

USENIX Association

**Proceedings of the
Eighteenth Symposium on
Usable Privacy and Security (SOUPS 2022)**

**August 7–9, 2022
Boston, MA, USA**

© 2022 by The USENIX Association

All Rights Reserved

This volume is published as a collective work. Rights to individual papers remain with the author or the author's employer. Permission is granted for the noncommercial reproduction of the complete work for educational or research purposes. Permission is granted to print, primarily for one person's exclusive use, a single copy of these Proceedings. USENIX acknowledges all trademarks herein.

ISBN 978-1-939133-30-4

Symposium Organizers

General Chair

Sonia Chiasson, *Carleton University*
Apu Kapadia, *Indiana University Bloomington*

Technical Papers Co-Chairs

Manya Sleeper, *Google*
Rick Wash, *Michigan State University*

Technical Papers Committee

Olabode Anise, *Figma*
Nalin Asanka Gamagedara Arachchilage,
University of Auckland
Hala Assal, *Carleton University*
Rebecca Balebako, *Google*
Alexandru Bardas, *University of Kansas*
Lujó Bauer, *Carnegie Mellon University*
Jasmine Bowers, *MITRE*
Cristian Bravo-Lillo, *Ciberseguridad Humana*
Lynne Coventry, *Northumbria University*
Lorrie Cranor, *Carnegie Mellon University*
Sauvik Das, *Georgia Institute of Technology*
Mary Ellen Zurko, *MIT Lincoln Laboratory*
Jinjuan Feng, *Townson University*
Carrie Gates, *Bank of America*
Julie Haney, *National Institute of Standards and Technology*
(*NIST*)
Jun Ho Huh, *Samsung Research*
Hyoungshick Kim, *Sungkyunkwan University*
Bart Knijnenburg, *Clemson University*
Janne Lindqvist, *Aalto University*
Jennifer Marlow, *Google*
Abigail Marsh, *Macalester College*
Peter Mayer, *Karlsruhe Institute of Technology*
Susan E. McGregor, *Data Science Institute,*
Columbia University
Mainack Mondal, *Indian Institute of Technology Kharagpur*
Alena Naiakshina, *Ruhr-University Bochum*
James Nicholson, *Northumbria University*
Simon Parkin, *Delft University of Technology*
Emilee Rader, *Michigan State University*
Robert Reeder, *Google*
Karen Renaud, *University of Glasgow*
Scott Ruoti, *The University of Tennessee*
Florian Schaub, *University of Michigan*
Kent Seamons, *Brigham Young University*
Jose Miguel Such, *King's College London*
Blase Ur, *University of Chicago*

Kami Vaniea, *University of Edinburgh*
Emanuel von Zezschwitz, *Google*
Daniel Votipka, *Tufts University*
Tara Whalen, *Cloudflare*
Josephine Wolff, *Tufts University*
Heng Xu, *American University*
Yaxing Yao, *University of Maryland, Baltimore County*
Leah Zhang-Kennedy, *University of Waterloo*

Lightning Talks and Demos Co-Chairs

Sanchari Das, *University of Denver*
Kopo Marvin Ramokapane, *University of Bristol*

Lightning Talks and Demos Junior Co-Chair

Taslima Akter, *Indiana University Bloomington*

Karat Award Chair

Blase Ur, *University of Chicago*

Posters Co-Chairs

Hala Assal, *Carleton University*
Camille Cobb, *University of Illinois at Urbana–Champaign*

Posters Junior Co-Chair

Kelsey Fulton, *University of Maryland*

Tutorials and Workshops Co-Chairs

Yaxing Yao, *University of Maryland, Baltimore County*
Leah Zhang-Kennedy, *University of Waterloo*

Tutorials and Workshops Junior Co-Chair

Harjot Kaur, *Leibniz University Hannover*

Mentoring Co-Chairs

Mohamed Khamis, *University of Glasgow*
Scott Ruoti, *The University of Tennessee*

Mentoring Junior Co-Chairs

Eva Gerlitz, *University of Bonn*
Ananta Soneji, *Arizona State University*

Publicity Co-Chairs

Adam Aviv, *The George Washington University*
Martin Degeling, *Ruhr-University Bochum*

Publicity Junior Co-Chair

Yixin Zou, *University of Michigan*

Email List Chair

Lorrie Cranor, *Carnegie Mellon University*

Accessibility Chair

Liz Markel, *USENIX Association*

USENIX Liaison

Casey Henderson, *USENIX Association*

External Reviewers

Jody Jacobs
James Mattei
Carson Powers
Sarah Radway

Ronald Thompson
Samantha Katcher
Rhea Votipka
Benjamin Berens

Kailani R. Jones
Dalton Brucker-Hahn
Yousif Dafalla
Shawn Robertson

William Seymour
Natalie Janosik
Adam Jenkins
Brian Singer

Message from the SOUPS 2022 Program Co-Chairs

Welcome to SOUPS 2022!

With the conference in its 18th year, our SOUPS community has collectively ensured an excellent and exciting conference program despite the challenges and obstacles caused by the global pandemic. With 37 papers accepted out of 133 submissions (28% acceptance rate), the technical program covers a wide range of topics within usable privacy and security. The conference also includes workshops, posters, lightning talks, mentorship activities, and a keynote.

In 2016, SOUPS became an independent conference body. For the last six years, we have partnered with USENIX for hosting and administrative support, a move that has enabled continued growth for the conference. We thank all the members of the USENIX staff for their work in organizing SOUPS and supporting our community. We particularly appreciate their support and flexibility this year, including managing the hybrid event. Their team has been fantastic at making the process seamless.

In 2018, we co-located with the USENIX Security Symposium for the first time, and we have continued that co-location in a hybrid format for 2022. Co-locating the two conferences allows for interactions and shared ideas between SOUPS and USENIX Security attendees. We have found this beneficial for both conferences and look forward to the opportunity again this year. We hope that this year's hybrid format will allow us to return to in-person interactions, facilitate participation for those joining remotely, and encourage interactions between both groups. Whether you join us in person or remotely this year, we hope you will find SOUPS 2022 engaging and meaningful.

SOUPS relies on a range of volunteers for all of its activities. Steering Committee members provide oversight and guidance and are elected for three-year terms. Organizing Committee members help determine the conference content for a particular year, often serving two-year terms to facilitate the transition of knowledge. Technical Papers Committee members are chosen by the Technical Papers Co-Chairs each year. SOUPS is a product of the hard work by many people, starting with researchers who decide to submit their work to SOUPS, and including all of the SOUPS Organizers, the SOUPS Steering Committee, the technical paper reviewers, the workshop organizers, the poster jury, and the USENIX staff. We are grateful and thank each and every one of you for your contributions to SOUPS 2022.

Sonia has served as General Chair of SOUPS and Chair of the Steering Committee for 2021 and 2022. Apu was appointed as Vice Chair in 2022 and will take on the role of General Chair for the following two years. If you are interested in helping with SOUPS 2023 in any way, please contact Apu.

SOUPS would not be possible without the generous support of our sponsors – thank you. Please visit our website to view the recipients of the SOUPS 2022 awards. Congratulations to all recipients for their outstanding work.

Sonia Chiasson, *Carleton University, General Chair*

Apu Kapadia, *Indiana University, Vice Chair*

Manya Sleeper, *Google, Technical Papers Co-Chair*

Rick Wash, *Michigan State University, Technical Papers Co-Chair*

Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)

August 7–9, 2022

Boston, MA, USA

Monday, August 8

Expertise and Learning

Replication: Stories as Informal Lessons about Security 1
Katharina Pfeffer and Alexandra Mai, *SBA Research*; Edgar Weippl, *University of Vienna*; Emilee Rader, *Michigan State University*; Katharina Krombholz, *CISPA Helmholtz Center for Information Security*

DualCheck: Exploiting Human Verification Tasks for Opportunistic Online Safety Microlearning 19
Ryo Yoshikawa, Hideya Ochiai, and Koji Yatani, *The University of Tokyo*

Understanding Non-Experts' Security- and Privacy-Related Questions on a Q&A Site 39
Ayako A. Hasegawa, *NICT*; Naomi Yamashita, *NTT / Kyoto University*; Tatsuya Mori, *Waseda University / NICT / RIKEN AIP*; Daisuke Inoue, *NICT*; Mitsuaki Akiyama, *NTT*

The Nerd Factor: The Potential of S&P Adepts to Serve as a Social Resource in the User's Quest for More Secure and Privacy-Preserving Behavior 57
Nina Gerber, *Technical University of Darmstadt*; Karola Marky, *Leibniz University Hannover and University of Glasgow*

"I don't know why I check this..." - Investigating Expert Users' Strategies to Detect Email Signature Spoofing Attacks 77
Peter Mayer, *SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology*; Damian Poddebniak, *Münster University of Applied Sciences*; Konstantin Fischer and Marcus Brinkmann, *Ruhr University Bochum*; Juraj Somorovsky, *Paderborn University*; Angela Sasse, *Ruhr University Bochum*; Sebastian Schinzel, *Münster University of Applied Sciences*; Melanie Volkamer, *SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology*

User Understanding of Security and Privacy Concepts

Industrial practitioners' mental models of adversarial machine learning 97
Lukas Bieringer, *QuantPi*; Kathrin Grosse, *University of Cagliari*; Michael Backes, *CISPA Helmholtz Center for Information Security*; Battista Biggio, *University of Cagliari, Pluribus One*; Katharina Krombholz, *CISPA Helmholtz Center for Information Security*

Replication: The Effect of Differential Privacy Communication on German Users' Comprehension and Data Sharing Attitudes117
Patrick Kührtreiber, Viktoriya Pak, and Delphine Reinhardt, *University of Göttingen*

Comparing User Perceptions of Anti-Stalkerware Apps with the Technical Reality 135
Matthias Fassel and Simon Anell, *CISPA Helmholtz Center for Information Security*; Sabine Houy, *Umeå University*; Martina Lindorfer, *TU Wien*; Katharina Krombholz, *CISPA Helmholtz Center for Information Security*

Users' Perceptions of Chrome Compromised Credential Notification 155
Yue Huang, Borke Obada-Obieh, and Konstantin Beznosov, *University of British Columbia*

Exploring User-Suitable Metaphors for Differentially Private Data Analyses175
Farzaneh Karegar and Ala Sarah Alaqra, *Karlstad University*; Simone Fischer-Hübner, *Karlstad University and Chalmers University of Technology*

An Empirical Study of a Decentralized Identity Wallet: Usability, Security, and Perspectives on User Control 195
Maina Korir, *University of Bedfordshire*; Simon Parkin, *TU Delft*; Paul Dunphy, *OneSpan*

Privacy and Security Tools

- Usability and Security of Trusted Platform Module (TPM) Library APIs** 213
Siddharth Prakash Rao and Gabriela Limonta, *Nokia Bell Labs*; Janne Lindqvist, *Aalto University*
- Increasing security without decreasing usability: A comparison of various verifiable voting systems** 233
Melanie Volkamer, *Karlsruhe Institute of Technology*; Oksana Kulyk, *IT University Copenhagen*; Jonas Ludwig and Niklas Fuhrberg, *Karlsruhe Institute of Technology*
- Presenting Suspicious Details in User-Facing E-mail Headers Does Not Improve Phishing Detection** 253
Sarah Zheng and Ingolf Becker, *UCL*
- Evaluating the Usability of Privacy Choice Mechanisms** 273
Hana Habib and Lorrie Faith Cranor, *Carnegie Mellon University*
- Detecting iPhone Security Compromise in Simulated Stalking Scenarios: Strategies and Obstacles** 291
Andrea Gallardo, Hanseul Kim, Tianying Li, Lujo Bauer, and Lorrie Cranor, *Carnegie Mellon University*

Tuesday, August 9

Methods

- If You Can't Get Them to the Lab: Evaluating a Virtual Study Environment with Security Information Workers**... 313
Nicolas Huaman, Alexander Krause, and Dominik Wermke, *CISPA Helmholtz Center for Information Security*;
Jan H. Klemmer and Christian Stransky, *Leibniz University Hannover*; Yasemin Acar, *George Washington University*;
Sascha Fahl, *CISPA Helmholtz Center for Information Security*
- Is it a concern or a preference? An investigation into the ability of privacy scales to capture and distinguish granular privacy constructs** 331
Jessica Colnago, *Google*; Lorrie Faith Cranor and Alessandro Acquisti, *Carnegie Mellon University*; Kate Hazel Stanton, *University of Pittsburgh*
- On recruiting and retaining users for security-sensitive longitudinal measurement panels** 347
Akira Yamada, *KDDI Research, Inc. and National Institute of Information and Communications Technology*;
Kyle Crichton, *Carnegie Mellon University*; Yukiko Sawaya, *KDDI Research, Inc.*; Jin-Dong Dong and Sarah Pearman, *Carnegie Mellon University*; Ayumu Kubota, *KDDI Research, Inc.*; Nicolas Christin, *Carnegie Mellon University*
- Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys** 367
Jenny Tang, *Wellesley College*; Eleanor Birrell, *Pomona College*; Ada Lerner, *Northeastern University*

Understanding Specific User Populations and Behaviors

- Aunties, Strangers, and the FBI: Online Privacy Concerns and Experiences of Muslim-American Women** 387
Tanisha Afnan and Yixin Zou, *University of Michigan School of Information*; Maryam Mustafa, *Lahore University of Management Sciences*; Mustafa Naseem and Florian Schaub, *University of Michigan School of Information*
- An open door may tempt a saint: Examining situational and individual determinants of privacy-invading behavior** .. 407
Markus Langer, *Saarland University*; Rudolf Siegel and Michael Schilling, *CISPA Helmholtz Center for Information Security*; Tim Hunsicker and Cornelius J. König, *Saarland University*
- Investigating How University Students in the United States Encounter and Deal With Misinformation in Private WhatsApp Chats During COVID-19** 427
K. J. Kevin Feng, *Princeton University*; Kevin Song, Kejing Li, Oishee Chakrabarti, and Marshini Chetty, *University of Chicago*
- Anti-Privacy and Anti-Security Advice on TikTok: Case Studies of Technology-Enabled Surveillance and Control in Intimate Partner and Parent-Child Relationships** 447
Miranda Wei, Eric Zeng, Tadayoshi Kohno, and Franziska Roesner, *Paul G. Allen School of Computer Science & Engineering, University of Washington*
- "Fast, Easy, Convenient." Studying Adoption and Perception of Digital Covid Certificates** 463
Franziska Herbert, Marvin Kowalewski, Theodor Schnitzler, and Leona Lassak, *Ruhr University Bochum*; Markus Dürmuth, *Leibniz University Hannover*

“As soon as it’s a risk, I want to require MFA”: How Administrators Configure Risk-based Authentication 483
Philipp Markert and Theodor Schnitzler, *Ruhr University Bochum*; Maximilian Golla, *Max Planck Institute for Security and Privacy*; Markus Dürmuth, *Leibniz University Hannover*

Passwords and Authentication

Let’s Hash: Helping Developers with Password Security 503
Lisa Geierhaas and Anna-Marie Ortloff, *University of Bonn*; Matthew Smith, *University of Bonn, FKIE Fraunhofer*; Alena Naiakshina, *Ruhr University Bochum*

Exploring User Authentication with Windows Hello in a Small Business Environment. 523
Florian M. Farke, Leona Lassak, and Jannis Pinter, *Ruhr University Bochum*; Markus Dürmuth, *Leibniz University Hannover*

Improving Password Generation Through the Design of a Password Composition Policy Description Language . . 541
Anuj Gautam, Shan Lalani, and Scott Ruoti, *The University of Tennessee*

Password policies of most top websites fail to follow best practices. 561
Kevin Lee, Sten Sjöberg, and Arvind Narayanan, *Department of Computer Science and Center for Information Technology Policy, Princeton University*

Do Password Managers Nudge Secure (Random) Passwords? 581
Samira Zibaei, Dinah Rinoa Malapaya, Benjamin Mercier, Amirali Salehi-Abari, and Julie Thorpe, *Ontario Tech University*

Let The Right One In: Attestation as a Usable CAPTCHA Alternative. 599
Tara Whalen, Thibault Meunier, and Mrudula Kodali, *Cloudflare Inc.*; Alex Davidson, *Brave*; Marwan Fayed and Armando Faz-Hernández, *Cloudflare Inc.*; Watson Ladd, *Sealance Corp.*; Deepak Maram, *Cornell Tech*; Nick Sullivan, Benedikt Christoph Wolters, Maxime Guerreiro, and Andrew Galloni, *Cloudflare Inc.*

IoT and Ubiquitous Computing

Being Hacked: Understanding Victims’ Experiences of IoT Hacking. 613
Asreen Rostami, *RISE Research Institutes of Sweden & Stockholm University*; Minna Vigen, *Stockholm University*; Shahid Raza, *RISE Research Institutes of Sweden*; Barry Brown, *Stockholm University & Department of Computer Science, University of Copenhagen*

Runtime Permissions for Privacy in Proactive Intelligent Assistants 633
Nathan Malkin and David Wagner, *University of California, Berkeley*; Serge Egelman, *University of California, Berkeley & International Computer Science Institute*

Normative and Non-Social Beliefs about Sensor Data: Implications for Collective Privacy Management 653
Emilee Rader, *Michigan State University*

Sharing without Scaring: Enabling Smartphones to Become Aware of Temporary Sharing. 671
Jiayi Chen and Urs Hengartner, *University of Waterloo*; Hassan Khan, *University of Guelph*

Balancing Power Dynamics in Smart Homes: Nannies’ Perspectives on How Cameras Reflect and Affect Relationships 687
Julia Bernd, *International Computer Science Institute*; Ruba Abu-Salma, *King’s College London*; Junghyun Choy and Alisa Frik, *International Computer Science Institute*

Replication: Stories as Informal Lessons about Security

Katharina Pfeffer
SBA Research

Alexandra Mai
SBA Research

Edgar Weippl
University of Vienna

Emilee Rader
Michigan State University

Katharina Krombholz
CISPA Helmholtz Center for Information Security

Abstract

Anecdotal stories about security threats told to non-experts by friends, peers, or the media have been shown to be important in forming mental models and secure behaviors. In 2012, Rader et al. conducted a survey ($n=301$) of security stories with a student sample to determine factors that influence security perceptions and behavior. We replicated this survey with a more diverse sample ($n=299$), including different age groups and educational backgrounds. We were able to confirm many of the original findings, providing further evidence that certain characteristics of stories increase the likelihood of learning and retelling. Moreover, we contribute new insights into how people learn from stories, such as that younger and higher educated people are less likely to change their thinking or be emotionally influenced by stories. We (re)discovered all of the threat themes found by Rader et al., suggesting that these threats have not been eliminated in the last decade, and found new ones such as ransomware and data breaches. Our findings help to improve the design of security advice and education for non-experts.

1 Introduction

Today, computers, mobile devices, and IoT devices permeate almost every aspect of our daily lives, forcing all users (including those with little to no security background) to make critical decisions about their IT security and privacy. These range from whether to click on a link or update software, to which password, antivirus software, or messaging service to choose. Although the usability of the devices has improved and security measures have been automated to a certain ex-

tent, the complexity of the decisions people have to make has continued to grow.

Several studies [16, 30, 31] have shown that people often make decisions based on incorrect or inaccurate mental models and misperceptions of security threats that expose them and others to security risks. In general, it is difficult for people to develop accurate mental models of cyber security threats since they typically cannot experience them themselves (i.e., we often do not directly experience security threats, nor can we observe others doing so since they are usually subtle or invisible). Redmiles et al. [26] found that people often reject security advice because they have not yet had a related negative experience themselves. They also found that people are generally overwhelmed with security advice from many different sources, such as newspapers, social media, movies, IT professionals, friends and family. In addition, their results suggested that people find it difficult to trust advice that comes from institutions that are obviously guided by marketing ideas.

As a possible solution to the lack of direct personal experience with security threats, it was found that in addition to security advice from IT professionals or security training, which is often ignored, negative experience stories from friends, family, or the media have a major impact on security decision making. We define negative experience stories as statements people have heard or read that relate to cyber security threats that happened to someone else. Rader et al. [25] were the first to examine how stories influence people's thinking and behavior. They conducted a survey in 2012 (hereafter referred to as the *Rader study*) in which they asked 301 undergraduate students open- and closed-ended questions about security advice they had heard from others. Using qualitative and quantitative methods, they determined the characteristics of these stories that lead to changes in thinking and behavior. The Rader study focused on undergraduate students and hence allows to only draw conclusions for this specific population. Also, their study was conducted a decade ago, and since then technology usage and the nature of security threats has fundamentally changed. Later, Fennell et al. [13] conducted another

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

user study examining how security stories may affect people's willingness to adopt two-factor authentication. Although they were able to demonstrate that stories increase adoption, they were unable to determine exactly what aspects of the stories might have convinced people to do so.

We seek to understand if the results from the Rader study are replicable ten years after the original study was conducted. We furthermore examine the generalizability to a broader population. We anticipated differences within our diverse sample, since prior work found evidence that demographics influence mental models, security behavior, and the processing of security advice [5, 26, 31]. A more nuanced understanding of which stories are remembered and which lead to changes in thinking and behavior is an important step towards making security advice better and more personalized. We thus replicated the Rader study with the following modifications:

- We recruited a more diverse sample with different age groups, educational backgrounds, and employment statuses.
- We applied a different recruiting strategy using quotas for age and gender to obtain a sample representative of the U.S. population.
- We examined the changed threat landscape reported in our stories and the changed media usage during the last decade.
- We applied inductive (instead of deductive) coding for the full stories, resulting in more in-depth themes grounded in our data.

Our found threat themes are similar to those of the Rader study, but ransomware and data breaches emerged as two new themes. We were able to confirm many of the original findings, such as that stories with a lesson affect our participant's behavior, while stories with serious threats affect thinking and the likelihood of retelling. Our results also confirm that stories that elicit fear or anger affect both thinking and behavior. In contrast to the Rader study, we found that stories told in a work environment are more likely to lead to behavior change than those told in casual contexts such as at home or in a coffee shop. We also report additional findings, which have not been examined in the Rader study, e.g. that younger and higher educated participants are less likely to report a change in their thinking.

2 Related work

Security advice and stories: Redmiles et al. [27] conducted semi-structured interviews to investigate from where people get security advice and found that a primary source is negative events they have experienced themselves or that have been passed on by peers, family, or the media. They also conducted a quantitative survey [26] to examine how people's security beliefs, knowledge, and demographics correlate with their

choice of security advice sources and their security behavior. Their findings suggest possible differences based on people's age and social status. In both studies, the trustworthiness of the advice source and the content of the advice play an important role in whether advice is accepted or rejected. In contrast to Redmiles et al., we do not ask how people decide which security advice to follow, but rather what effect stories have on people's thinking and behavior.

Fagan et al. [12] found that people decide to (not) follow security advice by weighting the benefits of following and the risks and costs of not following (balancing security and convenience). With our study, we investigate how stories can impact people's risk perception and security decisions. Ion et al. [19] compared the security practices of non-experts and security experts and found differences in the tools they use and their security behaviors. In this paper, we have a closer look at how stories impact the security tool usage and behavior of non-experts.

Rader et al. [25] were the first to study how security stories told by non-experts influence thinking and behavior. We replicate their study in this paper with a more diverse sample and some additional and modified survey questions. Rader et al. [24] conducted another study comparing three sources of security advice: news articles, web pages with security advice, and stories from friends or family (using the sample from the study described above). They found that personal stories often focused on who was carrying out the attacks, rather than how they were carried out or what the consequences were. Fennell et al. [13] showed that stories do indeed increase the people's willingness to adopt two-factor authentication. They hypothesized that focusing on negative consequences might work better than focusing on benefits. We investigate their hypothesis for our participants' security stories.

Mental models and risk perception: Mental models of the internet and security risks influence people's security behavior and decision making. Wash [30] identified eight non-expert mental models of security threats. Wash and Rader [31] quantified these mental models in a large-scale survey and found correlations between weak or incorrect mental models and insecure behavior. Asgharpour et al. [4] showed that risk communication often fails since it does not take the mental models of non-experts into account.

Kang et al. [20] examined experts' and non-experts' mental models of the internet and discovered that they often affect privacy and security decisions. Specifically, they found that a better understanding of risks can lead to a more secure behavior. Fulton et al. [16] showed that entertainment media, such as movies or series, can affect people's mental models by allowing them to learn from the actors' experiences (which, however, do not always correspond to reality). In this paper, we assume that security stories have an influence on people's mental models that must be considered alongside the influence of entertainment media, observation, and personal experience.

Nurse et al. [23] showed that in cyber security risk communication, characteristics of the message source (e.g., intent, reputation), the message (e.g., specificity, credibility), and the message recipient (e.g., beliefs, expertise) affect the effectiveness of the communication. In this paper, we investigate how characteristics of the storyteller, the story, and the recipient affect the likelihood of thinking and behavior change.

Psychology of Behavior Change: One theory commonly used to explain the adoption of secure behavior is the *Motivation-Ability-Trigger model* [14]. This says that a behavior only gets adopted if a person has the motivation, the ability, and is triggered to do so. We think that stories can affect all of these three properties, as people can share ideas to motivate and make each other aware, pass on strategies how to change a behavior, and trigger the behavior change by (re-)telling negative experiences to be avoided. Das et al. [9] showed that social processes often act as trigger to adapt secure behavior and were effective at raising security sensitivity.

The *Extended Parallel Process Model (EPPM)* [32] explains the role of fear-inducing communication in triggering behavior change. It states that although fear determines the intensity of the response, it is only effective if the person is also provided with a viable solution to the threat.

Another frequently cited theory is the *theory of planned behavior* [3], which was extended by Ng et al. [22]. Ng et al. stated that behavior change is affected by (i) perceived behavior control (over the ability to practice computer security), (ii) subjective norm (social pressure to perform an action), and (iii) attitude (influenced by the perceived usefulness). For (ii), the influence of peers, family, the mass media, as well as the work environment is important. In this study, we examine the influence of stories on the subjective norm, i.e. the social pressure to adopt a behavior. Along these lines, Ruoti et al. [29] developed a four-stage process for the adoption of security measures: learning, evaluation of risks, estimation of impact, and weighing trade-offs to different coping strategies.

Generally, it has been shown that *social proof and personal examples* have an impact on secure behavior and decision making. Das et al. [11] demonstrated that when showing people the usage of security features of friends, they were more likely to be adopted. Das et al. also showed in a retrospective interview study [10] that social influence, especially the observability of security feature usage, can affect people's thinking and behavior. Similar, Harbach et al. [17] revealed that risk communication with personalized examples (i.e. which information is at risk when installing an app) can foster secure behavior. With our work we further examine how security stories, much like personal examples, can affect people's thinking and behavior, serving as substitute for the lack of observability of security feature usage.

3 Methodology

We replicated the study design from the Rader study with a few changes. The authors of the original study shared their anonymized data with us for statistical comparison. In this section, we explain the modifications we did to the original questionnaire, our prestudy, how we recruited a diverse sample, and how we analyzed our survey data.

3.1 Questionnaire

Rader's study questionnaire started with an introduction text explaining the goal of the survey. Afterwards, participants answered four open questions where they had to name cyber security threats, protection measures, and stories they had heard related to security threats. These questions were used to help participants remember any stories they might have heard or read. Finally, they had to choose one story they could most easily recall details about and answer the subsequent questions in regards to that story. Most of the following questions were multiple choice. In the last part of the survey, after the participants had thought about the story for a while, they were asked to write down the story as if they would tell it to a friend using as many details as possible. In total, the survey consisted of 7 open-ended questions and 38 closed (multiple choice or checkbox) questions.

For the replication study, we changed the wording of the original questionnaire in the introduction text and in several questions to explicitly include mobile threats (see Appendix 9.1). Moreover, we changed and shortened the original web use skill measure where respondents had to rate their knowledge of technical terms such as phishing, meme, or cache on a Likert scale. We updated the terms to be up to date and included more highly understood terms since our audience mainly consisted of non-experts. According to Hargittai et al. [18], adjusting web-skill measures based on the characteristics of the targeted population helps to reduce non-responses (i.e., non-experts might quit the survey when faced with numerous lesser known items due to frustration).

We also included a bogus term (filtibly), which served as a attention check question. We discarded all responses that rated their knowledge of this term as "good" or "full". We considered the rating "little" still acceptable, since we assumed that the participants might remember having seen a word kind of like "Filtibly" before. We shortened the questions where participants had to rate emotions the story made them feel on a Likert scale, to shorten the time and concentration needed to complete the survey. We added an additional question asking whether the participants received formal training in IT security since we expected differences based on this characteristic. However, we did not find any significant correlations in participants with or without formal training in any of our regression models. Since we wanted to compare the participants' own negative experiences with their reported stories, we included a

Table 1: Demographics

		Sample	Quotas [6]
Gender	Female	53%	51%
	Male	45%	49%
	Prefer not to say	2%	-
Age	18-34	28%	27%
	35-44	18%	17%
	45-59	26%	26%
	>59	28%	30%
Education	High school	8%	
	Technical, vocational school	5%	
	Some college	25%	
	Bachelor degree	36%	
	Master degree	19%	
	Doctoral degree	4%	
	Other	2%	
	None	<1%	
Employment	Employed full time	54%	
	Employed part time	12%	
	Retired	16%	
	Unemployed	9%	
	Student and empl. part-time	1%	
	Student	3%	
	Disabled	2%	
	Other	3%	

question asking about cyber security threats they experienced themselves. We placed this question after they told the story they have heard or read, in order to not confuse them or prime them towards thinking about their own experiences instead of stories they have heard.

We conducted a prestudy ($n=16$) to test the comprehensibility of our questionnaire. At the end of the prestudy, we included a question to ask participants about survey parts that were unclear to them and to make improvement suggestions. We found that responses to the question about the moral of the story and learnings from the story were redundant, thus we merged this questions. Otherwise, no problems arose.

3.2 Recruitment and Participants

We hosted our survey on SurveyMonkey [1] and used their participant pool for recruitment, which allowed us to put quotas on age and gender to ensure that our sample largely matched these quotas of the U.S. population as published by the United States Census Bureau [6] (see Table 1). However, our sample is not representative of culturally different regions. Completing the survey took an average 13 minutes. We compensated each participant with US\$5 for their time and effort.

We started our study paying for 350 participants, assuming that we will have to exclude about 15% invalid responses, similar to the Rader study. However, it turned out that about half of our responses did not meet our criteria (see below). After consulting with SurveyMonkey, they relaunched our survey free of charge until we had collected enough responses that matched our original quotas and criteria. For both launches, we received in total 622 responses, from which we excluded:

- 239 since they were unusable (participants did not re-

member a story, wrote a story not related to cyber security, or answered inconsistently),

- 19 since they failed the attention check question, i.e. rated their understanding of "Filtibly" as "good" or "full" (we accepted 28 ratings as "little"),
- 52 since they wrote a story about themselves,
- 13 since they gave an advice instead of writing a story.

This left us with 299 usable responses.

3.3 Analysis

We used a combination of (i) qualitative coding to account for the subtleties of the stories told, and (ii) quantitative statistical analyses to calculate differences between demographic groups and compare our results with those of the Rader study.

Qualitative Coding: To code the full stories and the responses to the open-ended questions, we used inductive thematic coding [7]. Our goal was to find repeating patterns in the data and use them to build theories. The Rader study used (i) deductive thematic coding with a pre-defined codebook consisting of story themes to code the full stories and (ii) inductive thematic coding to code the open-ended questions, creating the codebook by grouping recurring themes into higher-level themes and sub-themes.

We applied the second approach, i.e. inductive thematic coding, to both the full stories and the open-ended questions, aiming at grounding our analysis as close as possible in the meaning of the data. This allowed us to gain more in-depth results including meta reflections from the full stories. We created a codebook (see Appendix 9.2 for the final version) based on recurring patterns in the full stories and the open-ended questions. First, two independent researchers open-coded all full stories and open-ended questions to create an initial codebook of themes and sub-themes. One of the researchers had not previously read Rader's study, and the second researcher also attempted to look at the emerging categories in an unbiased manner.

Second, both researchers discussed the emerging themes and jointly created a joint version of the codebook. Although reading the Rader study may have influenced one researcher's coding process, we are confident that we also included an unbiased view by jointly creating the codebook. Besides, since our goal was to compare our findings with those of the original study, we do not see it as problematic that our codebook may have been influenced by the codes of the Rader study.

Third, both coded all responses independently. Fourth, they discussed the differences and decided to merge certain sub-themes of the codebook that were too similar and therefore resulted in different code assignments. For example, the second codebook had a "ransomware" topic with "enterprise" and "public entity" sub-themes, which were merged. For the remaining differences, we calculated the inter-coder-agreement

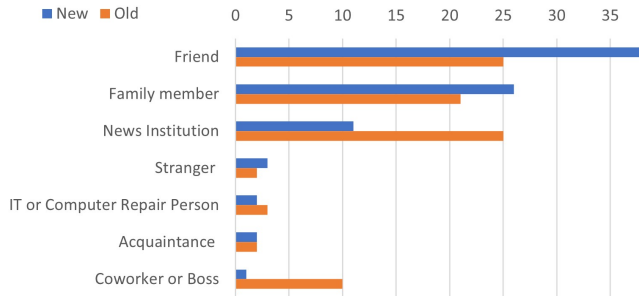


Figure 1: Source of hearing/reading the story (percentages).

with a Cohen’s Kappa κ [8] of 0.89, which shows a good level of agreement. Finally, both researchers met and reached agreement on all code assignments.

Quantitative Analysis/ Statistical Tests: We used logistic regression for binary dependent variables and OLS regression for interval scaled dependent variables. We calculated models for the same factors as the Rader study to make the results comparable. Additionally, we created models with different demographics (e.g., age, education level) as factors.

4 Results: Descriptive Statistics

Most of our participants chose a story told by friends or family members (see Figure 1), which is similar to the findings of the Rader study. However, we discovered that fewer stories in our sample came from news institutions. We also found that a lower percentage of stories (34% in our sample versus 55% in the Rader study) were told face-to-face, as more people used social networks, instant messaging, or the phone. We attribute this to either changes in technology use over the past decade or to the global pandemic. The majority of our participants (64%) heard the story within the last year.

In line with the Rader study, we found that 96% of our participants believed that the story was true (see Table 2). Almost half of the stories (48%) were retold, most (57%) within a day and almost all within a week (90%). 59% of the stories were autobiographical, meaning that the protagonist was the person telling the story. Our results and the Rader study show that most of the stories contain a lesson about something you should always do or never do, or both.

Our participants had to rate the seriousness of the threat described in their chosen story on a Likert scale of 1-5. We

Table 2: Facts about stories

	New	Old
Believed story to be true	96%	95%
Retold Story	48%	45%
Autobiographical	59%	51%
Contains lesson	71%	72%
Change of behavior	52%	52%
Change of thinking (mean, 1-5 scale)	3.1	2.9
Seriousness of threat (mean, 1-5 scale)	4.1	3.7

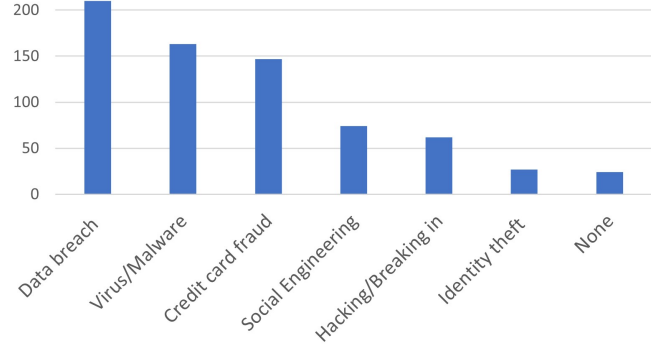


Figure 2: Threats that happened to participants (counts).

report a mean score of 4.05, indicating very high severity, which is higher than the more moderate mean score of 3.5 in the Rader study. Possible reasons for this are discussed in Section 6. The average story had a moderate effect ($M=3.1$) on participants’ thinking and influenced their behavior in half of the cases in both studies.

Stories told by our participants affected single or multiple individuals, companies, governmental or educational institutions, or society as a whole. The reported threats resulted in the loss of money, time, data, reputation, or the availability of critical infrastructure such as a gas pipeline, electricity, or the healthcare system. We asked our participants which threats happened to them personally (see Figure 2). We found that more than two thirds had already fallen victim to a data breach. Many also experienced credit card fraud or having a virus or malware. Fewer participants reported that hacking or social engineering such as phishing had happened to them.

5 Qualitative Results

In this section, we report and discuss our qualitative findings in comparison to the Rader study. Note that although we report numbers for both studies, they are not directly comparable since many themes overlap (e.g., "Hacking/Breaking In", "Virus/Malware", and "Social Engineering"). For all themes, multiple assignments are possible for one story.

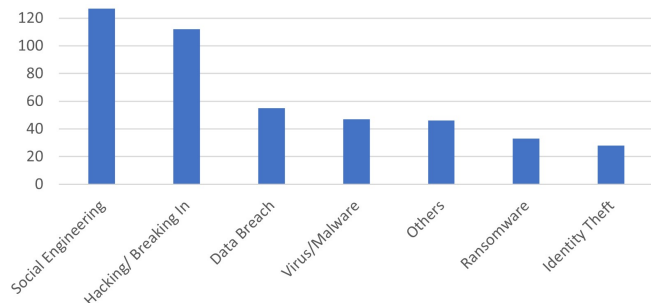


Figure 3: Threat categories of reported stories (counts).

5.1 Full Stories

For the full stories, we constructed sub-themes to each threat category, in contrast to the Rader study, where only the top-level themes were coded. Figure 3 shows the distribution of emerged threats of the reported stories.

Since we asked participants to report on only *one* story they remember most vividly, these numbers do not claim to be representative of security story themes in general. However, we use them to show trends in such themes.

Comparing the threats found in the stories with the threats experienced personally by the participants (see Figure 2), it seems that incidents related to hacking or social engineering are mentioned much more often in the stories than they were experienced personally. Note that we did not introduce a specific category for ransomware in our questionnaire, thus these threats may have been reported as hacking or viruses/malware in Figure 2.

Social Engineering The largest fraction of stories were related to social engineering threats (127), which were often coded together with the categories "Virus/Malware" (47) and "Hacking/Breaking In" (112). Social engineering threats include phishing messages (71), where people are tricked into clicking on fraudulent links or attachments, and fraudulent messages or calls (25), where attackers take false identities and tell fraudulent stories to steal valuable information or money. These threats occurred via social media, email, or via the phone. Moreover, participants were told about fraudulent websites (15), apps (4), or USB sticks (2) tricking people into giving away sensitive information. We identified several pretexts under which these social engineering attacks are usually carried out. Many participants shared stories of using fake friend requests or messages on social media platforms to gain trust and subsequently steal information. Some stories were about sophisticated threats where the attacker went through multiple stages:

"Someone pretend to be his high school classmate [and] requested friend connection. [They] chatted and exchanged email addresses. Tom shared [his] work email. Tom received [an] email from Facebook (fake, phishing). Tom clicked [on the] email content and [his] computer got infected by a virus. The virus infected Tom's company network and the hacker stole company data."

In many cases, the attacker posed as a friend or relative in need of money, a representative of the bank or tax office, or a co-worker or boss. Some stories also claimed that the victim had won money or an item.

"The incident had occurred after my friend had sent personal information to another Instagram user who had claimed they would send them money via cash app."

Several stories were about a fraudulent pop-ups or IT help desk numbers (8).

"He came across an old man whose computer was infected and was asking to call a support number to fix it. The person called the support number which was actually a hacker."

Comparison Rader study: They also found many stories related to phishing (53) using similar tricks as reported in our study. In line with our findings, they reported that phishing messages ranged from emails pretending to be from a bank to more sophisticated attempts, where someone started a chat with the victim via Facebook or an online game. Unfortunately, this shows that phishing is a persistent problem that has not been solved in the last decade.

Hacking/Breaking In The majority of stories in this category were about hacked bank, email, or social media accounts (91), which usually included a hacked password. As a result of the hacked account, various threats occurred, such as attackers sending spam emails or messages, or making transactions. Some incidents were more serious than others, such as:

"Someone hacked the email of a vendor and provided fraudulent wire details to pay for an invoice."

Several stories dealt with the hacking of (public) WiFis (2), cameras (16), or celebrities (3).

"A stranger hacked into the camera and was spying on the child and started speaking to the child through the security camera."

Comparison Rader study: They also reported 59 stories in the category "Breaking In". In this category, our results are very similar to the Rader study, as many incidents of hacked computers, systems and accounts were also reported there. Their participants also often talked about negative affects of the "hacking", such as altered accounts or profile information, or sending fraudulent messages. This shows that such hacking incidents are a long-term challenge that has not yet been solved.

Data Breach We found many stories of data breaches affecting banks, credit institutes, shops, retailers, or institutions (49) and the governmental (2), educational (1), or healthcare (3) sector. These stories described personal data such as social security numbers or credit card information being stolen, and affected customers often being informed of the incident via email.

Comparison Rader study: They describe theft (75) only in the context of stealing personal information or money through unauthorized credit card use, fraudulent websites, or as part of a phishing scam. They did not report stories about data breaches, which our participants frequently described. For this reason, and in line with the cyber security report [2], we assume that the frequency of data breaches has mainly increased in recent years.

Virus/Malware We found many generic stories related to a virus or malware attacking the victims' computers or phones (22). From those that described the viruses in more detail, seven mentioned that screens behaved differently (e.g., turned blue or blinked), nine that devices slowed down or crashed, and five that people were logged off. Three stories involved link redirects, where the victim was always redirected to a site chosen by the attacker, independent from the URL that was entered into the browser.

"Everytime my friend opened up his web browser, it would go to a fake looking Google search engine site. A virus was planted from someone so that it forced him to use that site to search with."

One participant also mentioned that the virus or malware was stealing data, which was associated with phishing.

Comparison Rader study: They describe similar stories to the ones we found and grouped them under the category of PC effects (95). Their participants also frequently reported that their computers behaved differently than usual due to a virus or malware, such as freezing, being slow, or losing information. In comparison to the Rader study, we found a decrease in stories about viruses, possibly due to the increasing importance of newly emerging threats such as data breaches or ransomware. Still, 23% of our respondents said they had already been a victim of a virus or malware (see Figure 2) showing that these threats are still prevalent.

Ransomware A large percentage of the stories were about ransomware that affected both individuals (13) and companies/public institutions (20). For instance, stories were told about ransomware affecting critical public infrastructure such as oil pipeline companies or hospitals. These stories describe attacks that locked computers or encrypted data and asked the owner to pay a ransom in order to regain data access. Various reasons have been cited as the source of the ransomware, including viruses, clicking on fraudulent links, attachments or pop-ups, or connecting fraudulent external devices. The amount of requested ransom ranged from four hundred to several millions of dollars. Of these stories that mentioned whether the ransom was paid or not, 76% (13) did pay the ransom. In most stories, data access was returned after the ransom was paid. However, in some stories this was not the case.

"My friends computer was locked. She got a message [that] there was a virus and she had to call a number. She did and they needed \$500. She paid and it did not resolve the problem. She had to take it in and pay more."

"He paid the requested ransom but he still lost all his files."

Those who did not pay the ransom either found a way to remove the ransomware themselves, had backups of their data, or faced serious consequences.

"My brother-in-law's small real estate company received a ransom notice. They were told that unless they paid \$100,000 all their files would be destroyed. He thought it was a joke at first. It was not. Luckily they had an off-site backup that saved the day."

Comparison Rader study: Ransomware was not reported because this type of cyber threat, although it already existed, was very rare in 2012.

Identity Theft Although this theme often appears along with others, we decided to code it as a separate theme, as it was re-occurring. This category includes stories about people whose identities were stolen so that the attacker could open up credit card accounts, conduct fraudulent transactions, or purchase houses and other expensive items in the victim's name. As a result, the victims' credit scores or reputations were often ruined. In some cases, it took them a long time to resolve the problem. Some stories report the usage of fraudulent websites or hacking as the source of identity theft, while most claim not to know why this happened.

Comparison Rader study: They also report on identity theft as part of their category "Theft" (see above) and give the example of identity theft by a fraudulent website that their participant claimed had been "hacked."

Others This category includes various themes that appeared more frequently but could not be assigned to an overarching theme. Two stories were about whistle-blowing and six about that Facebook generally steals data and is not respecting people's privacy. Two other stories were about cyber bullying that led to serious psychological consequences for the victim. Three stories described a person catching a scammer to prevent the scam or to set an example. Two stories mentioned software vulnerabilities leading to security attacks.

5.2 Retelling Stories

When asking our participants with whom they shared the story and why they did so, three main themes emerged:

The majority of participants (64) explained that the story contains a general risk which has to be shared with everybody, while thirty-one participants reported that they shared the story only with impacted people. Impacted people ranged from those who potentially fell victim to a data breach or hack to those who might open spam messages from a specific person.

"My friend's Facebook account got hacked. Watch out for links from him."

Six participants explicitly mentioned that they shared the story with others who they assume to not be knowledgeable (e.g., elderly people) or who they assume to be highly knowledgeable and therefore, interested in their story.

Our participants mentioned a variety of emotions as reasons for retelling the story, which were scary/dangerous (10), unexpected/unbelievable/crazy (7), relevant/informative (10), interesting (8), funny/entertaining (2), and frustrating/sad (3).

The most common reason for retelling (97) was to create awareness and knowledge to protect others from falling for the same threat. Fourteen participants described that an action was required such as changing the provider, reacting to a shutdown due to ransomware, or reacting to a shutdown of computers in a work environment. Six participants answered that the story fitted the conversation, two aimed at getting other opinions, and two simply wanted to spread gossip.

Comparison Rader study: They did not report qualitative findings on why people retell stories.

5.3 Learnings and Behavior Changes

The themes for participants' learnings and behavior changes overlapped since learning and behavior is often intertwined, so we coded them together. Five main themes emerged:

Behavior Most of our participants (215) expressed that they learned some kind of security awareness or caution from the story. While many expressed these in general phrases, e.g. "To be very careful online" or "Security is important", others were specific about their behavior change. Fifty participants reported to have changed their password security practices and usage as a result of the story heard, such as "keep different passwords for different accounts" or "always update passwords". Another fifteen updated their software or changed to a more secure version. Twelve participants started to monitor their accounts or credit card charges more carefully. Another six participants did back-ups of their data, mostly as a response to stories about ransomware. Two participants stopped connecting to insecure WiFis and one changed their privacy settings. Five participants mentioned that they communicated with others about their security concerns. Four even took such radical actions as to quit using Facebook or using credit cards.

Comparison Rader study: In line with our findings, the Rader study found that many participants described their behavior change on a very generic level which means that they seemed to not have taken actionable advice from the stories, but rather vague learnings. As an exception their findings indicated that participants explicitly mentioned to have changed their password habits as well as using antivirus (see paragraph "Tools/Services"), which we can also confirm with our study. When it comes to specific fields in which participants learned something or reported behavior change, the Rader study reported similar findings regarding caution when clicking on links, downloads and shady websites, where participants learned actionable lessons. We also confirm findings from the Rader study of participants being more keen to update software and monitoring their accounts. One theme which we found in our data was not reported in the Rader

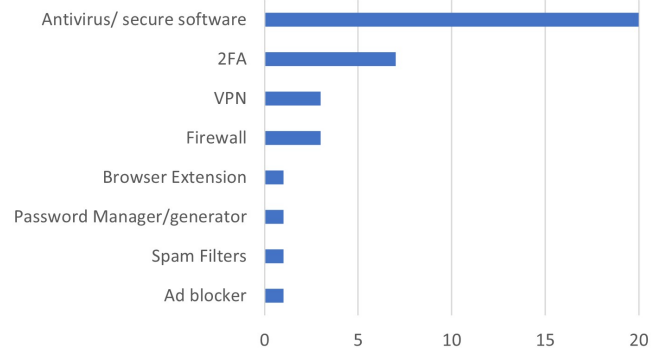


Figure 4: Reported tools/services participants started using after hearing the story (counts).

study: Backing-up data, which was often reported along with ransomware.

Distrust Thirty-one participants mentioned a general distrust in data protection online as well as in security applied by companies or institutions. For example,

"Even though you think your data is undoubtedly secure, there is always a chance it could be compromised."

Five mentioned a distrust in a specific technology such as email (4), credit cards (1), or apps. One participant wrote, "Just because Android apps are in the Google Play store does not necessarily mean that they do not contain malware."

Comparison Rader study: They also found a theme describing that the internet is generally a dangerous place and that their participants often distrusted companies as well strangers in the internet. This is an interesting finding, as it speaks for an experienced helplessness without the participant having learned anything that could improve their situation.

Tools/Services Many participants stated that they started to use a new tool or service after hearing the story (see Figure 4), where the most prominent tool was antivirus software, followed by 2FA, VPNs and firewalls.

Comparison Rader study: They also found that the most participants reported to start using antivirus and keeping it up-to-date. However, they did not report on participants mentioning 2FA or VPNs. This is likely because these tools have grown in popularity over the last decade.

Education Another theme that emerged was that many participants (18) said that they started to educate themselves more about possible threats online as well as prevention mechanisms, such as:

"I ended up reading more about scams as well as watching videos on the topic."

Some mentioned to also educate their employees or vulnerable people (i.e., elderly). This is in line with the theme we

created for why participants retold stories, where we found that sharing them with elderly people who might not be as tech-savvy was often mentioned.

Comparison Rader study: No such theme was reported.

Others 20 participants said they had learned that everyone can be impacted, which was a belief that was not as present for them before they heard the story. Examples are attacks on close individuals which made it clear to participants that such threats are not only discussed in the media but happen in reality, as well as data breaches or ransom attacks on big companies which were considered to have security in place.

"It shows how a big company can be hacked and required to pay despite having security software."

Three participants described that their views were reinforced.

6 Quantitative Results

In this section, we report our quantitative findings in comparison to those of the Rader study. Note that although the change in behavior and thinking is self-reported as a causal relationship by our participants, we cannot infer causality from our survey, only correlations. For all logistic regression models, calculated for binary dependent variables such as change in behavior (yes/no), retelling (yes/no), we report odds ratios. For the OLS regressions, calculated for interval scaled variables such as change in thinking (1-5), angry/anxious (1-5), seriousness of threat (1-5), we report estimates to interpret our results.

6.1 Stories' influence on thinking and behavior

In line with the Rader study, we found that specific properties of a story change the thinking and behavior of our participants. There were two types of properties, related to the content and the source. We built two respective regression models, which we can directly compare to the results of the Rader study.

Content influences: Table 3 shows that when a story *contains a lesson*, i.e., claims something which one should always or never do, then the odds that the participant reported they had changed their behavior are about twice as high as for stories without a lesson. This replicates the results of the Rader study. For the influence of stories with lessons on thinking, we

Table 3: Content influences on thinking and behavior

	Change in Behavior		Change in Thinking	
	New	Old	New	Old
(Intercept)	0.19	0.27	1.66	2.27
Contains a lesson	2.02 **	2.33 **	0.18	0.26 .
Seriousness of threat	1.31 *	1.14	0.27 ***	0.15 **
Autobiographical	1.42	1.79 *	0.43 **	0.15

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

found a positive but non-significant correlation, in contrast to the Rader study which found a significant correlation between stories containing a lesson and reported change in thinking. This shows that lessons directly affect behavior, but there was inconsistency in perceived change in thinking.

The *seriousness of the threat* described in the story significantly impacts the change in behavior and change in thinking. We found that the odds of a reported change in behavior are 31% higher when the seriousness of the threat increases by one, while the Rader study did not find statistically significant results. Moreover, we found a strong influence of serious threats on the change in thinking, which is in line with the Rader study. We hypothesize that the seriousness of threats is an influential property of the story, as people usually have a *negativity bias* [28]. This means that people tend to give more weight to negative events than to positive ones. Therefore, negative stories are more likely to be present in people's minds and to influence their thinking and behavior.

Whether a story is *autobiographical*, i.e. the protagonist is the same person as the one telling the story, seems to have some influence on thinking and behavior. However, it is unclear which of the two are influenced more, since we found a significant correlation between autobiographical stories and thinking, whereas the Rader study found a correlation with behavior change. This only shows that, since thinking and behavior are so deeply intertwined, a distinction might not always be possible. We cannot be sure why this correlation exists. It might be that autobiographical stories are perceived as more credible or easier for people to identify with.

Source influences: Where and from whom a story is heard also influences the thinking and behavior, as shown in Table 4. When the story is heard in a *casual context* such as at a friend's or relative's house, at a coffee shop, or at home, then our results show that the odds are 41% lower that participants reported they had changed their behavior. This is in contrast to the Rader study where the odds were 95% higher for that casual context changes the behavior. We compared casual context to a more formal context such as at work, in class, or in the library, which seems to have increased the odds for changing the thinking of our participants. Due to these conflicting results, we searched for other variables (e.g. demographic differences) in our data that could explain the difference. We found that participants of age > 60 are more likely to hear the story in a casual context (presumably since they are often retired), as well as participants from 18-29 years

Table 4: Source influences on thinking and behavior

	Change in Behavior		Change in Thinking	
	New	Old	New	Old
(Intercept)	1.11	0.21	3.30	2.50
Casual Context	0.59 .	1.95 .	0.27 .	0.28
Knowledgeable Source	1.12	1.40 **	0.32 ***	0.11

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Casual context: compared to "Home context" as baseline.

old. However, we did not find a significant influence of the age group on the change of thinking or behavior, which is why we excluded it from our source influence regression model. When it comes to the change in thinking, we found a positive correlation with the reported change in thinking. This finding is also statistically significant, in contrast to the Rader study. Basically, our results suggest that a formal context more likely influences perceived change in behavior, while the casual context more likely influences a change in thinking. We suspect that people might feel more pressured at work to behave in a certain expected way after hearing a story (e.g., from a co-worker or boss). However, since we found different results in comparison to the Rader study, this hypothesis should be taken with a grain of salt.

We found that stories from a *knowledgeable source* (expertise of the source rated on a 1-5 Likert scale) significantly increase the change of thinking. Although we also found a positive correlation for behavior change, this result is not statistically significant. However, since the Rader study found the same correlation with statistical significance, we hypothesize that an effect of source expertise on both change in thinking and behavior exists. Sources with greater knowledge may be perceived as more trustworthy.

Table 5: Influence of emotions on thinking and behavior

	<i>Change in Behavior</i>		<i>Change in Thinking</i>	
	New	Old	New	Old
(Intercept)	0.16	0.27	1.79	1.83
Happy	1.22	0.91	0.04	0.07
Sad	0.95	0.64	0.05	0.15
Anxious	1.46 *	1.88 *	0.33 ***	0.24 *
Anger	1.44*	1.84 **	0.14 .	0.19 *

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Influence of emotions: We asked our participants to what extent (on a 1-5 Likert scale) they experienced the emotions listed in Table 5 after hearing the story. In line with the Rader study, we found a significant impact of feeling *anxious* or *angry* about a story on both thinking and behavior, and negative odds (although not statistically significant) for the influence of feeling *sad* on the change in behavior. This could again be explained with the negativity bias [28], saying that negative events are more impactful than positive ones, and with the EPPM model [32], stating that fear can induce behavior change. Moreover, our results and the Rader study have shown that stories involving serious threats influence reported changes in thinking and behavior, and we hypothesize that such stories are more likely to make participants anxious and angry.

6.2 Story Retelling

While the Rader study found that whether a story *contains a lesson* does significantly increase the odds of retelling this

Table 6: Content influence on retelling

	<i>Retelling</i>	
	New	Old
(Intercept)	0.13	0.17
Contains a lesson	1.51	2.30 **
Seriousness of Threat	1.46 **	1.30 *
Autobiographical	1.28	1.07

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

story, this correlation was not statistically significant in our data, although we also found a positive correlation (see Table 6). However, in line with the Rader study, our regression model for content properties shows a 46% increased chance of the influence of the *seriousness of the threat* on the retelling. Hence, the seriousness of threat seems to be a pivotal property of a story, which significantly influences our participants' thinking, behavior, and retelling.

Table 7: Source influences on retelling

	<i>Retelling</i>	
	New	Old
(Intercept)	0.65	0.29
Casual Context	0.91	0.88
Knowledgeable Source	1.14	1.41 **

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Casual context: compared to "Home context" as baseline.

We did not find statistically significant effects for *autobiographical* stories, nor for any of the source properties (see Table 7) on whether a story is retold or not. This means, we could not replicate the correlation between a knowledgeable source and retelling a story in the Rader study.

Table 8: Influence of emotions on retelling

	<i>Retelling</i>	
	New	Old (Re-calculated)
(Intercept)	0.14	0.28
Happy	1.21	1.47.
Sad	0.95	0.89
Anxious	1.40 *	1.10
Anger	1.47 **	1.24

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

For the influence of emotions on retelling (see Table 8), we found a strong correlation between stories that made participants *anxious* (40% increase in the odds of retelling) or *angry* (40% increase in the odds of retelling). This correlation was statistically significant, in contrast to the Rader study. Similar to the influence of these feelings on thinking and behavior, we think that this correlation can be explained by that more exciting stories are more likely retold. This also matches with our qualitative results regarding participant's answers on why they retold the story (see Section 5.2).

6.3 Demographics' influence

We fitted various regression models to investigate the influence of demographics on variables of our interest such as

change in thinking and behavior, emotions, and factors that have been shown to influence perceptions and behavior in the Rader study (e.g., seriousness of threat or context). In this section, we only report those models where we found statistically significant correlations.

Table 9: Influence of demographics on seriousness of threat and change of thinking

	<i>Seriousness of Threat</i>	<i>Change in Thinking</i>
<i>(Intercept)</i>	4.25	3.61
Age		
18-29	-0.72 ***	-0.32
30-44	-0.20	0.07
45-60	0.05	0.09
Education		
Some college	0.10	-0.48
Techn., voc. school	-0.04	-0.53
Bachelor Degree	-0.11	-0.70 *
Master degree	-0.13	-0.66 *
Doctoral Degree	0.04	-0.67

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Age: ">60" as baseline; Education: "High school" as baseline.

Table 9 shows that younger participants between 18-29 perceive the seriousness of threat statistically significantly lower than other age groups and are slightly less likely to report a change of thinking after hearing a story, which is however not a significant results ($p>0.1$). For participants with higher education, we found a statistically significantly lower likeliness of changing their thinking.

Table 10: Influence of demographics on emotions

	<i>Angry</i>	<i>Anxious</i>
<i>(Intercept)</i>	3.30	2.57
Age		
18-29	-0.48 *	0.17
30-44	-0.13	0.04
45-60	0.27	0.17
Education		
Some college	-0.34	-0.33 .
Techn., voc. school	-0.70 *	-0.71 *
Bachelor Degree	-0.59 **	-0.48 *
Master degree	-0.55 *	-0.48 *
Doctoral Degree	-1.25 ***	-0.87 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Age: ">60" as baseline; Education: "High school" as baseline.

Table 10 shows, that participants age 18-29 and with higher education also reported feeling angry about a story less often. Hence, younger and higher educated participants are less likely emotionally affected by stories and assess their seriousness lower. This might explain why they are less likely to change their thinking, which we found to have a positive correlation with the perceived seriousness of the threat and feeling anxious or angry (see Table 3). We did not find differences in the influence of participant's demographics on the reported behavior change or chance of retelling a story.

7 Discussion

Comparison Rader study: The results of our replication study confirm many of the original findings. We found that neither the topics of the stories nor the ways in which participants learn best from stories have changed much across time and demographics. The threat landscape we discovered in our full stories is very similar to that of the Rader study with the exception of the newly found categories of "Ransomware" and "Data Breaches." This is in line with the Accenture Cyber Threat Intelligence Report of 2021 [2], stating that ransomware as well as infostealing were active problems in 2021. Other threats such as social engineering, hacking, viruses, and identity theft have been an unsolved problem for over a decade. Although we changed the introductory text and some questions to explicitly include mobile threats in addition to computer threats, we did not find neither more stories of mobile threats than the Rader study, nor did we find new story themes related to mobile threats. This shows that the results of the Rader study are still largely valid today, across age and educational differences.

Similar to the Rader study we found that behavioral changes based on stories can help both prevent security threats and respond to them after they occur. Stories often conveyed strategies for responding to threats, such as advice on whether or not to pay ransom, or awareness of data breaches. Our results also suggest that participants often learn distrust through stories, which in many cases was only described on a general level and was unhelpful. Only in some cases distrust led to secure behavior, such as frequent monitoring of accounts. Consistent with the EPPM model [32], we hypothesize that this may be the case since effective responses to fear are only possible if viable solutions to threats are offered. Another interesting finding was that stories can encourage participants to educate themselves about certain security-related topics. We can confirm that the stories were mainly related to what happened rather than why. We suspect that this is due to the fact that security threats often cannot be traced back to their source and are only noticed when they occur.

Stories and psychology of behavior change: Our findings suggest that the threats in the commonly told stories differ from those that participants had experienced themselves. This means that stories can broaden the range of threat awareness. In addition, we found evidence that stories can influence security risk perception, as participants often reported learning that anyone can be affected by security threats after hearing a story that happened to a close relative or a company they previously considered secure. Moreover, our results confirm that stories can influence participants' thinking, which in turn can change their motivation and capabilities and serve as a trigger for adopting secure behavior.

Our qualitative and quantitative survey results suggest that our participants' thinking and behavior are intertwined. For

the seriousness of the threat and the emotions of anger and anxiety, we found a correlation with thinking, behavior, and retelling. Although our results suggest that some factors only affect behavior (such as containing a lesson) and others only thinking (such as autobiographical stories and knowledgeable sources), we argue that the two are difficult to separate. If changes in behavior only, but not in thinking, are reported for one factor, this could also be because people's mental models are often tacit [21] and people may not be aware that they are changing. Behavioral changes tend to be more obvious and therefore easier for participants to recognize and report. Likewise, if participants reported a change of thinking without a change in behavior, altered mental models could affect their security decisions without them being aware.

Advice based on stories: Surprisingly, we found that it made no significant difference whether stories came from the media or from friends or family. It would be interesting to see whether this finding holds in the future as the media landscape continues to change. We argue that our results can be used to better design media articles on security threats, advice, and security training. We suggest that those should focus on stories containing lessons with concrete actions and serious threats to the individual. This is in line with Nurse et al. [23], suggesting that clear actions increase the effectiveness of risk communication, and with the EPPM model [32], stating that fear can encourage behavior change if a concrete solution is presented. One idea could be to create an online platform, e.g. on social media, where people can share security incidents, since we found that autobiographical stories positively effect learning. This platform could serve as a story pool for media articles or security training, which could pass on the most relevant or often occurring stories. When a knowledgeable person leads the training or writes the media article, this could further increase the impact of the stories told.

Participants often reported that they retold stories since they fitted the conversation, which shows that bringing IT security on people's agenda on its own already improves the likelihood of sharing stories and learning from them. This finding could be used, e.g. in companies, to encourage employees (e.g. in specifically therefore scheduled meetings) to share security incidents. We found a positive correlation of sharing security stories at work and reported behavior changes, so this could help people learning from their colleagues. We hypothesize that the reason stories heard in a work context have a greater influence on behavior than those heard in a casual context may be that they influence the subjective norm, which in the theory of planned behavior is the social pressure to perform an action [3, 22]. However, it is up to future work to investigate this further, as the Rader study found that stories in casual contexts have a greater impact on perceived behavior changes.

We derive from our results that younger and better educated participants are harder to reach with security stories, as they commonly perceive threats as less serious and are less likely

to report being emotionally affected by stories and change their thinking. We hypothesize that this is because people growing up with information technology have had more exposure to security reports and therefore, perceive security threats as less shocking. Education possibly increases the chances to have heard about similar security incidents before, thus being less anxious or angry about them, and less likely to change the thinking as a result. Future work to explain these correlations is required. However, we still found that this group is influenced by autobiographical stories from knowledgeable sources, which should be kept in mind when designing security advice. We found that elderly and less educated participants might learn easier from security stories. However, according to Frik et al. [15] they are also at higher security risk due to less knowledge and experience with technology, but do not always perceive threats as more severe. Hence, for elderly people or those retired (who do not have access to training at work-places), it would be especially useful to create a platform for sharing stories in their own words.

Methodological reflections: It was generally straightforward to replicate the original study since all needed material was available. We noticed that our participants' responses were similar to the Rader study in terms of complexity and technical details. In line with the Rader study, we also found that some participants gave superficial and general answers when asked how the story changed their thinking or behavior. It could be that these participants were unable to draw specific conclusions from the stories or that the setting of an online survey did not encourage them to explain details. It would be interesting to explore the impact of different stories on people's thinking and behavior in a qualitative interview study in the future. Because the data in our study and the original study consist only of self-reported stories, future work could examine in a prospective study how different stories affect people's security behaviors in the wild. Although we only asked participants to tell us one story that they remembered most clearly, and thus may have missed others, we believe that the reported stories are the ones that are retold most often and thus have the greatest effect.

8 Conclusion

With our replication study, we confirm most of the Rader study's findings regarding which characteristics of stories lead to changes in thinking, behavior, and retelling. In addition, our diverse sample allowed us to examine differences among participants of various age groups and educational backgrounds. Based on our findings, we provide guidance on how security training or media content on IT security can be better designed. We strongly suggest that security stories should be considered alongside professional training and personal experience as important sources of security advice.

9 Acknowledgments

We thank the anonymous reviewers for their valuable feedback on our work. SBA Research (SBA-K1) is a COMET Center within the framework of COMET – Competence Centers for Excellent Technologies Program and funded by BMK, BMDW, and the province of Vienna. The COMET program is managed by FFG.

References

- [1] SurveyMonkey. <https://www.surveymonkey.com/>, accessed: 2022-06-01.
- [2] Accenture. Cyber threat intelligence report - vol 2 21. https://www.accenture.com/_acnmedia/PDF-158/Accenture-2021-Cyber-Threat-Intelligence-Report.pdf, accessed: 2022-06-01.
- [3] Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.
- [4] Farzaneh Asgharpour, Debin Liu, and L Jean Camp. Mental models of security risks. In *International conference on financial cryptography and data security*, pages 367–377. Springer, 2007.
- [5] Vanessa Boothroyd. *Older Adults’ Perceptions of Online Risk*. PhD thesis, Carleton University, 2014.
- [6] United States Census Bureau. ACS demographic and housing estimates. <https://data.census.gov/cedsci/table?q=DP05&tid=ACSDP5Y2020.DP05>, accessed: 2022-06-01.
- [7] Victoria Clarke, Virginia Braun, and Nikki Hayfield. Thematic analysis. *Qualitative psychology: A practical guide to research methods*, 222:248, 2015.
- [8] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- [9] Sauvik Das, Laura A Dabbish, and Jason I Hong. A typology of perceived triggers for end-user security and privacy behaviors. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 97–115, 2019.
- [10] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A Dabbish, and Jason I Hong. The effect of social influence on security sensitivity. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 143–157, 2014.
- [11] Sauvik Das, Adam DI Kramer, Laura A Dabbish, and Jason I Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 739–749, 2014.
- [12] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 59–75, 2016.
- [13] Chris Fennell and Rick Wash. Do stories help people adopt two-factor authentication? *Studies*, 1(2):3, 2019.
- [14] Brian J Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, pages 1–7, 2009.
- [15] Alisa Frik, Leysan Nurgalieva, Julia Bernd, Joyce Lee, Florian Schaub, and Serge Egelman. Privacy and security threat models and mitigation strategies of older adults. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.
- [16] Kelsey R Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L Mazurek. The effect of entertainment media on mental models of computer security. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 79–95, 2019.
- [17] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2647–2656, 2014.
- [18] Eszter Hargittai and Yuli Patrick Hsieh. Succinct survey measures of web-use skills. *Social Science Computer Review*, 30(1):95–107, 2012.
- [19] Iulia Ion, Rob Reeder, and Sunny Consolvo. “... no one can hack my mind”: Comparing expert and non-expert security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, 2015.
- [20] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. “my data just goes everywhere:” user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 39–52, 2015.
- [21] Anne R Kearney and Stephen Kaplan. Toward a methodology for the measurement of knowledge structures of ordinary people: the conceptual content cognitive map (3cm). *Environment and behavior*, 29(5):579–617, 1997.

- [22] Boon-Yuen Ng and Mohammad Rahim. A socio-behavioral study of home computer users' intention to practice security. *Pacific Asia Conference on Information Systems (PACIS)*, 2005.
- [23] Jason RC Nurse, Sadie Creese, Michael Goldsmith, and Koen Lamberts. Trustworthy and effective communication of cybersecurity risks: A review. In *2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST)*. IEEE, 2011.
- [24] Emilee Rader and Rick Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, 2015.
- [25] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS 2012)*, pages 1–17, 2012.
- [26] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How i learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677, 2016.
- [27] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288. IEEE, 2016.
- [28] Paul Rozin and Edward B Royzman. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320, 2001.
- [29] Scott Ruoti, Tyler Monson, Justin Wu, Daniel Zappala, and Kent Seamons. Weighing context and trade-offs: How suburban adults selected their online security posture. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 211–228, 2017.
- [30] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS 2010)*, pages 1–16, 2010.
- [31] Rick Wash and Emilee Rader. Too much knowledge? security beliefs and protective behaviors among united states internet users. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325, 2015.
- [32] Kim Witte. Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs*, 59(4), 1992.

Appendix

9.1 Questionnaire

In this survey, we are interested in things you have heard about or learned through stories from others related to protecting your computer or mobile device and yourself from cyber security threats. We are NOT interested in something that happened to you personally, only in stories about other people you've heard, e.g. from a friend, coworker or acquaintance, social media sites, blogs and newspapers, or any other source you can think of.

Cyber security threats might include things like hackers, viruses, malicious apps, identity theft, shady URLs in spam emails, etc. It can be very hard sometimes to tell when someone is facing a cyber security threat- symptoms might include when someone's computer or mobile device is slow or freezes unexpectedly, when programs won't close, or lock up, unwanted popup windows, spam email, posts appearing in someone's Instagram or Facebook account without their permission or knowledge, or other undesirable computer or mobile device issues. Sometimes people cope with these threats by using tools such as anti-virus or firewall software, or by making sure to back up their data, or not clicking links or installing apps from people they don't know or trust.

We will start with 4 longer open questions to help you start to remember stories you have heard or read about cyber security. Afterwards, we will continue with shorter questions, which are mainly multiple choice.

1. First, please make a list of as many different kinds of computer or mobile security problems, or threats that you can think of, using only a couple of words to describe each of them.

Open-ended answer

2. Next, think of all the different ways you can protect yourself and your computer or mobile device from cyber security problems or threats, and make a list of these below.

Open-ended answer

3. Take a moment to think back to times in the past when you remember being told or reading about a story related to computer or mobile security. Please make a list of as many of these stories as you can remember, using only a couple of words to describe each story (you may want to read over your answers to the previous questions to jog your memory).

Open-ended answer

4. Finally, please choose one story for which you can most easily recall details about where you were and what happened when you heard or read the story (You can go back to review your list). In a sentence or two, briefly summarize what happened. You will be answering further questions about this story in the rest of the survey.

Open-ended answer

5. How long ago did you hear or read the story?

Answers: Within the last day/ Within the last week/ Within the last month/ Within the last year/ Longer than one year ago / Don't remember

6. Where were you when you heard or read the story?

Answers: At a coffee shop/ At a friend or relative's house/ At home/ At work/ In a computer lab In class/ Other (please specify)

7. Via what medium did you hear or read the story?

Answers: In person (face-to-face)/ Phone/ Text message/ Chat (instant messaging)/ Video chat/ Email/ Blog post/ Social network site (Instagram, Facebook, Twitter, etc.)/ Print news media (physical newspaper, magazine, etc.)/ Broadcast news media (TV, Radio, etc.)/ Online news media (CNN.com, Yahoo News, etc.)/ Don't remember/ Other (please specify)

8. From what source did you hear or read the story?

Answers: Family member/ Friend/ Acquaintance/ Coworker or Boss/ IT or Computer Repair Person/ Stranger/ News Institution/ Don't Remember/ Other (please specify)

9. How knowledgeable do you think the source you selected above is about cyber security? Please rate the source's knowledge from 1 (Not Knowledgeable) to 5 (Very Knowledgeable).

10. Did you tell, send, post, or otherwise share this story with anybody else?

Answers: Yes/ No/ Don't remember

11. Approximately how many times did you share the story?

Answers: 1/ 2/ 3/ More than 3/ Don't remember

12. With whom did you share the story (select all that apply)?

Answers: Family member/ Friend Acquaintance/ Coworker or Boss/ IT or Computer Repair person/ Stranger/ News Institution/ Follower/ Don't Remember/ Other (please specify)

13. How long after you first heard or read the story did you first share it with others?

Answers: Within one day/ Within one week/ Within one month/ Within one year/ Longer than one year/ Don't Remember/ Other (please specify)

14. Please briefly describe why you shared this story with others.

Open-ended answer

15. Was this story about the same person who told the story to you or who wrote it?

Answers: Yes/ No/ Don't Remember/ Other (please specify)

16. How serious was the threat or problem? Please rate the severity from 1 (Not Serious At All) to 5 (Very Serious).

17. Did the story end well or badly for the main character? Please rate the outcome from 1 (Very Well) to 5 (Very Badly).

18. In general, was the story about something you should ALWAYS do (e.g., wash your hands after using the bathroom), or something you should NEVER do (e.g.,

stick your tongue to a frozen flagpole)?

Answers: Always do/ Never do/ Both/ Neither/ Other (please specify)

19. What did you learn from this story?

Open-ended answer

20. This story made me feel... Sad/ Happy/ Helpless/ Curious/ Angry/ Anxious (Not at all - Somewhat - Mostly - Extremely)

21. Did you start doing anything differently to try to protect yourself from IT security threats or problems after hearing this story?

Answers: Yes/ No/ Other (please specify)

22. Please describe one thing you started doing differently after hearing this story.

Open-ended answer

23. Do you believe this story actually happened?

Answers: Yes/ No/ Don't know

24. How much do you think hearing this story has affected the way you think about cyber security threats?

Please rate it from 1 (A Lot) to 5 (Not At All)

You're almost done!

25. You have now answered a number of questions about a story, you remembered being told or reading about, related to a computer or mobile security threat or problem. Below, please write the story as if you were telling it to a friend. Use as much detail as you can, including any thoughts or recollections you might have had about what happened as you were filling out the survey. Use about 4-5 sentences to describe the story.

Open-ended answer

26. Have you ever had one of the following experiences? Select all that apply:

Answers: Fell victim to a phishing email message or other scam email/ Received a notification from a company that your information was involved in a data breach/ Had a virus on your computer or mobile device/ Someone broke in or hacked your computer, mobile device, or account/ Stranger used your credit card number without your knowledge or permission/ Identity theft more extensive than use of your credit card number without permission/ None of the above

27. What is your age in years?

Open-ended answer

28. What gender do you identify as?

Answers: Female/ Male/ Prefer not to say/ Other

29. What is your highest completed level of education?

Answers: None/ High school/ Technical, vocational school AFTER high school/ Some college/ Bachelor degree/ Master degree/ Doctoral degree/ Other (please specify)

30. What is your current employment status? *Answers:* Employed full time/ Employed part time/ Unemployed looking for work / Unemployed not looking for work/ Retired/ Student /Student and employed part-time/ Disabled/ Other (please specify)

31. Please rate your understanding of each term below

from None (no understanding) to Full (full understanding). Wiki/ Meme/ Phishing/ Bookmark/ Cache/ TLS/ AJAX/ RSS/ Filitbly

32. Have you ever received formal training in computer science, software engineering, IT, computer networks, or a related technical field?

Answers: Yes/ No/ I'm not sure

9.2 Codebook

Table 11: Codes and counts for full stories

A Ransomware	33	B Data Breach	55	C Social Engineering	127
A.1 Company/Public Institution	20	B.1 Shop/Company/Bank	49	C.1 Phishing/Scam messages	71
A.2 Individual	13	B.2 Governmental	2	C.2 Scam Call	25
A.3 Ransom payed: yes	13	B.3 Educational	1	C.3 Fraudulent website	15
A.4 Ransom payed: no	4	B.4 Healthcare	3	C.4 Fraudulent pop-Up	8
				C.5 Fraudulent app	4
				C.6 Revenge	2
				C.7 Fraudulent device	2
D Virus/Malware	47	E Hacking	112	F Others	74
D.1 General	22	E.1 Account/password/data	91	F.1 Others	31
D.2 Screen different	7	E.2 Device	16	F.2 Whistleblower	2
D.3 Computer slow	2	E.3 WiFi	2	F.3 Cyber bullying	2
D.4 Computer crash	7	E.4 Celebrity	3	F.4 Facebook privacy	6
D.5 Logged out	5			F.5 Catching scammer	3
D.6 Link redirection	3			F.6 Security vulnerabilities	2
D.7 Stealing data	1			F.7 Identity theft/Credit card fraud	28

Table 12: Codes and counts for reported learnings and behavior changes

O Behavior		P Distrust		Q Tools/Services	
O.1 Security awareness/caution	215	P.1 General/ Company/ Institution	31	Q.1 Firewall	3
O.2 Change settings	1	P.2 Credit cards	1	Q.2 Ad blockers	1
O.3 Credit/ account monitoring/ protection	12	P.3 Email	4	Q.3 VPN	3
O.4 Back ups	6	P.4 Technology/Devices	4	Q.4 2FA	7
O.5 Connect to trusted WiFis	2			Q.5 Spam Filters	1
O.6 Updating/ securing software	15			Q.6 Antivirus/ secure software	20
O.7 Password hygiene/usage	50			Q.7 Password manager/ generator	1
O.8 Exchange with others about security (concerns)	5			Q.8 Browser Extension	1
O.9 Stop using tool/service					
S Education		T Ransom should be		V Other	
S.1 Employees	4	T.1 paid	2	V.1 Everyone can be impacted	20
S.2 Elderly	3	T.2 not paid	1	V.2 View reinforcement	3
S.3 General/ Self	18			V.3 Stop using credit card	3
				V.4 Stop using Facebook	1
				V.5 Other	27

Table 13: Codes and counts for why stories were retold

K Shared with		L Incident was		M Reason	
K.1 impacted persons	31	L.1 scary/dangerous	10	M.1 Action required	14
K.2 all/ general risk	64	L.2 unexpected/unbelievable/ crazy	7	M.2 Knowledge/Awareness/ Warning/ Protection	97
K.3 other (not/knowledgeable)	6	L.3 relevant/informative	10	M.3 Fitted conversation	6
		L.4 interesting	8	M.4 Get other opinions	2
		L.5 funny/entertaining	2	M.5 Gossip	2
		L.6 frustrating/ sad	3		

DualCheck: Exploiting Human Verification Tasks for Opportunistic Online Safety Microlearning

Ryo Yoshikawa
The University of Tokyo
ryo@iis-lab.org

Hideya Ochiai
The University of Tokyo
ochiai@elab.ic.i.u-tokyo.ac.jp

Koji Yatani
The University of Tokyo
koji@iis-lab.org

Abstract

Learning online safety and ethics is becoming more critical for the general user population. However, they do not receive such learning opportunities regularly, and are often left behind. We were therefore motivated to design an interactive system to provide more frequent learning opportunities to the general user population. This paper presents our explorations on the integration of opportunistic microlearning about online safety and ethics into human verification. Our instantiation of this concept, called DualCheck, asks users to respond to questions related to online safety and ethics while human verification would be executed in a similar manner to reCAPTCHA v2. In this manner, DualCheck offers users microlearning opportunities when they use online services. Our 15-day user study confirmed the positive learning effect of DualCheck. The quantitative and qualitative results revealed participants' positive experience with attitude toward DualCheck, and also found its significantly higher perceived usability than text-based CAPTCHA and picture-based reCAPTCHA.

1 Introduction

As many general users enjoy online services and communication regularly, understanding online safety and ethics is becoming an essential and critical literacy. However, they do not necessarily have sufficient opportunities to learn online safety and ethics. According to recent surveys conducted by Information-technology Promotion Agency (IPA) in Japan [6], only 17.9% of smart device users claimed that they had taken explicit training on online ethics.

Furthermore, such training typically occurs at school or workplace, and the frequency is also limited. This suggests that general users may not have constant opportunities for learning online safety and ethics.

We were therefore motivated to design an interactive system to provide more frequent learning opportunities to users. More specifically, we were interested in how we can exploit existing interactions which users are already familiar with that purpose. In this work, we exploit human verification tasks which are commonly seen in online forms. For instance, CAPTCHA [17] and its variants are widely used and well recognized. As human verification is common in many online services, an integration of learning opportunities would increase the frequency of such training in an opportunistic manner. Our research questions in this work are, therefore, 1) how the integration of opportunistic learning on online safety and ethics into human verification can support people's learning; and 2) how the user experience of such a system would be different from existing human verification tasks.

This paper presents our investigations on integrating opportunistic microlearning of online safety and ethics into human verification tasks to answer these two research questions. We develop DualCheck as a proof of our concept (Figure 1). Users see DualCheck as a human verification task at the end of online forms. They then read the question and answer by clicking one of the five choices. The system presents users the correct answer and explanation for their learning. It then enables a button to move to the next page 5 second after users' responses to the question. The system does not consider any information about whether their responses are correct or not for human verification. Instead, human verification is expected to be performed in a similar manner to the checkbox-based reCAPTCHA v2. In this manner, DualCheck can achieve reliable human verification while offering microlearning of online safety and ethics in an opportunistic manner.

Our evaluation through a 15-day deployment study confirmed significant improvements on the accuracies (correct answer rates) for 9 of the 10 questions used

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

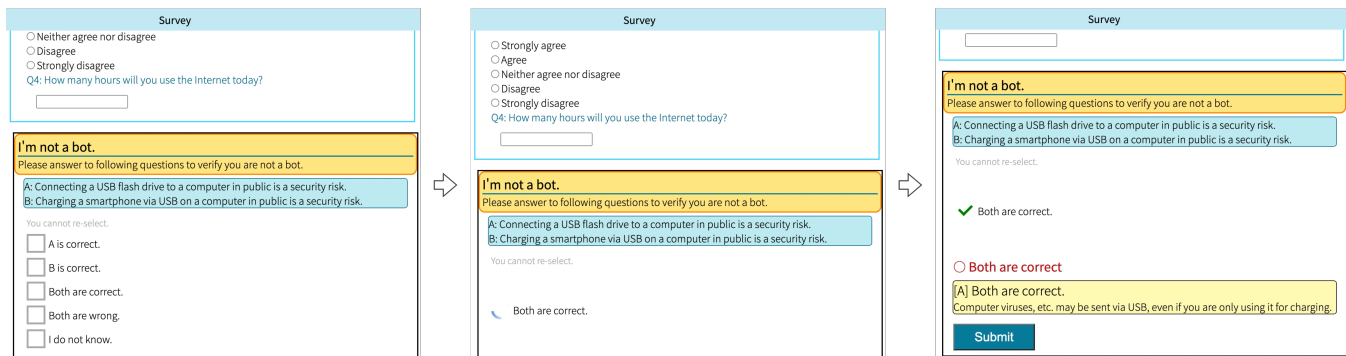


Figure 1: The DualCheck interface. Left: DualCheck can be integrated into online forms as a human verification mechanism. We note that our current prototype does not implement the human verification mechanism because our primary objective of this work is to validate the effect of opportunistic microlearning through DualCheck instead of evaluating the robustness of human verification. Middle: Users choose one of the choices after reading the question presented by DualCheck. The current implementation simply pretends to be performing human verification like reCAPTCHA v2. Right: DualCheck presents the correct answer and explanation about the given question. The system enables the submit button five seconds after it shows the correct answer and explanation. In this manner, users have an opportunity to read them. The system does not consider whether users have chosen the correct answer or not for human verification. Instead, it is expected to perform human verification through a mechanism like reCAPTCHA v2.

throughout the deployment study. In addition, our participants exhibited significantly higher accuracies on 5 of another 10 questions about online safety and ethics than general Internet users who do not use DualCheck. The perceived usability of DualCheck was significantly higher than text-based CAPTCHA and picture-based reCAPTCHA. Our qualitative results support participants' positive attitudes toward DualCheck.

The primary contributions of this work are two-folded:

- Development of DualCheck, our proof of concept of the integration of opportunistic microlearning about online safety and ethics into human verification tasks; and
- Evaluation of DualCheck through a 15-day deployment study, confirming its positive learning effect and user experience.

2 Related Work

2.1 Online Safety and Ethics Learning

Learning online safety and ethics is critical for general users as they now have multiple computer devices and access various media and online social platforms. However, people lack learning opportunities of such knowledge, often being left at risk. School curricula in different countries now include learning about online safety and ethics, but they are not necessarily effective. According to surveys conducted by Information Technology Promotion Agency (IPA) in Japan in 2019 [6, 7], only 38.0% of teenagers explicitly responded that they had online ethics training. The percentage of such

people becomes even lower in older generations; for example, the number becomes only 9.6% of the survey respondents in their 70s. Furthermore, such training occurred at school or workplace for 76.8% of the respondents who claimed they had such training. Furthermore, their survey [7] also found that most of the respondents did not possess even basic knowledge concerning information security. For example, only 28.5% of the respondents were aware of the concept of “malware,” and only 13.6% answered all three questions about malware correctly. A study by Grimes et al. [5] in the United States showed similar results; they found that older adults have lower awareness of online safety. Their study showed that older adults possess considerably less knowledge and awareness of Internet security hazards than university students.

These survey results indicate that learning opportunities are limited outside schools and workplaces, and thus people's knowledge about online safety and ethics is also constrained. In particular, people do not have learning opportunities regularly. Reinheimer et al. investigated the effectiveness of an awareness and education program on phishing [12]. They found that participants' phishing discrimination capabilities were maintained up to four months after the education program, but degradation occurred after that. Their result thus confirms that regular training is critical.

Existing work attempted to utilize games to motivate people's learning about online safety, Sheng et al. [14] integrated anti-phishing knowledge into a video game. They [13] further confirmed the insufficiency of people's cyber hygiene behaviors and knowledge through conducting roleplay-based phishing attacks. Although such approaches can be beneficial, further explorations on online safety and

ethics learning approaches are necessary as Drury et al. [9] suggested that attacks and threats are evolving and becoming more complex and sophisticated.

Our work exploits a human verification task seen in various online forms for online safety and ethics learning. As people often encounter such tasks during their Internet use, our DualCheck can offer more frequent learning opportunities than existing school curriculum or training at a workplace. The main objective of this work is to validate the effect of the integration of microlearning into human verification.

2.2 Opportunistic microlearning

Microlearning is a learning style where learners undergo small learning units repeatedly. Tasks in microlearning are deliberately designed to be small so that learners can complete them within a short amount of time. Another merit of microlearning is that it can be integrated into users' interactions or tasks to provide learning moments in an opportunistic manner. Prior work in the field of Human-Computer Interaction has examined the learning effect of opportunistic microlearning systems.

Many projects targeted vocabulary development through their opportunistic microlearning systems. Trusty and Truong [16] developed a browser extension that automatically translated words on a Web page in English to a foreign language users were learning. The foreign translations were thus integrated into the existing context in English, offering opportunistic vocabulary learning when users were reading Web pages. Their user study revealed that participants were able to acquire 50 new foreign words per month on average. Cai et al. [2] created a vocabulary-learning system that exploits users' waiting time during a text chat. The study showed that participants learned 57 words in two weeks on average, indicating that the system was effective for vocabulary learning. Dingler et al. [3] implemented QuickLearn to exploit mobile notifications for microlearning. It presents users vocabulary questions via mobile notifications. In this manner, QuickLearn offers lightweight access to vocabulary learning materials even if users are on the go or only have a limited amount of attention. In their experiment, participants learned 18 words per week on average.

While vocabulary development is a common opportunistic microlearning application, this work extends its scope to online safety and ethics learning. Mohammed et al. [10] conducted a study incorporating microlearning into ICT education for elementary school students. They found that microlearning with flashcards and videos increased learning ability by up to 18% compared to textbook-based education and also resulted in better retention of long-term memory. Our investigation of this work demonstrates how effective opportunistic microlearning of online safety and ethics would be in a case of the integration into human verification.

2.3 GUI-based human verification

Human verification systems distinguish users from bots to prevent malicious automated access. CAPTCHA [17], developed by Ahn et al., is one of the most widely-used human verification systems. An early version of CAPTCHA required users to correctly type a visually-skewed string. Ahn et al. also developed reCAPTCHA [18]. It provides the same function as the original CAPTCHA but can also improve OCR software. However, problems in functionality and usability were also recognized. Yan and El Ahmad [20] discussed the robustness and usability of text-based CAPTCHAs. They pointed out usability concerns owing to the degree of distortion of the text and the presence of confusing characters.

To address such usability issues, research has examined alternative forms of human verification. Yamamoto et al. [19] designed a task of reordering four-frame cartoons. Fanelle et al. designed new audio CAPTCHAs that are primarily used by users with visual impairments [4]. Their designs were superior to those of existing audio CAPTCHA in terms of accuracy and speed. Recent developments on CAPTCHA have led to more lightweight interaction for human verification. reCAPTCHA v2 only requires the user to click a checkbox. reCAPTCHA v3 does not even require any explicit interaction from users.

The objective of this work is not to propose a novel human verification task nor evaluate its robustness and usability. Our primary advantage is the integration of opportunistic microlearning into a human verification task. Tanthavech et al. [15] showed Math CAPTCHA, which asks users to solve simple calculation problems, received the highest user experience rating among the five human verification task designs. One possible reason for this is that such tasks might have served as quick brain exercises. We hypothesize that a human verification task would become more acceptable if users could perceive benefits directly from it. This work examines this hypothesis in the context of online safety and ethics learning.

3 DualCheck

3.1 System Implementation

Our system, DualCheck, provides opportunistic microlearning of online safety and ethics through the human verification task of ticking a checkbox, similar to reCAPTCHA v2. More specifically, our interface presents users with a multiple-choice question about online safety and ethics. In this manner, users can learn online safety and ethics while performing human verification tasks.

Figure 1 shows the interface implemented in a Web environment. Our interface can be easily integrated into online forms for human verification. DualCheck shows a multiple-choice question about online safety and ethics comprising two statements, and users are asked to determine

whether each statement is correct. The following five choices are provided as responses: “Only Statement A is correct,” “Only Statement B is correct,” “Both statements are correct,” “Both statements are wrong,” and “I cannot tell.” While deeply investigating the learning effect of question and response formats is out of scope of this work, we decided to employ a multi-choice question for DualCheck because it is a common question style and reCAPTCHA v2 would fit this style well. After users tick one of the checkboxes, the system presents the correct answer and a short explanation to encourage them to acquire the appropriate knowledge. The system offers users 5 seconds before enabling the button to move to the next page. In this manner, it encourages users to read the correct answer and explanation.

The human verification in our system would not be based on whether users answer questions correctly. Instead, human verification is expected to be performed in a manner similar to that in reCAPTCHA v2, i.e., analyzing the cursor behavior when clicking a checkbox. We also note that reCAPTCHA v2 or equivalent human verification mechanisms are not integrated into our current prototype due to the unavailability of these codes. Moreover, our primary purpose of this work is to investigate the effect of opportunistic microlearning instead of human verification performance.

The question curation and user studies for DualCheck were executed in the local language of the authors though other languages can be accommodated. We translated the questions and answers into English for the report in this paper.

3.2 Question Curation

We created a set of questions for our demonstration and deployment study of DualCheck. We set the following two criteria to create questions: 1) questions should cover common issues and practices related to online safety and ethics; 2) questions should neither be too difficult nor too well-known. Using these criteria, we conducted a literature survey of existing online safety and ethics guidelines and learning materials designed for high school or older users. These references included materials for teaching high school students [11] and materials to educate the public on the latest knowledge of Internet hazards [8].

We initially prepared 29 questions from these resources, which covered various common online threats. One of the authors, an expert on network security systems, reviewed them to filter out questions considered too difficult or obvious. We also revised the phrasing of the questions based on their feedback. Finally, we had 25 questions.

We then conducted a crowdsourcing-based study to validate these 25 questions. The objective of this part of the study was two-fold: 1) confirming whether the questions were comprehensible and 2) observing how many participants would respond to these questions correctly. We used a crowdsourcing service available in the country of the authors.

Each crowdsourcing participant was asked to answer a subset of the 25 questions in a multiple-choice format. In addition, the task included a quality control question where its answer was obvious even for the general user populations (e.g., “I posted the password to my account on an SNS.”). This question curation process was approved by our institutional review board.

331 crowdsourcing participants volunteered for this study in total, and we collected 100–115 responses for each question (110 on average). 20 participants failed the quality control question, and their responses were discarded. Table A.1 in the Appendix A includes the entire set of questions and their percentage of the correct answers.

We found that 15 of the 25 questions exhibited correct answer rates greater than 80%. These questions would not be appropriate for our deployment study because people are already aware of these online safety and ethics issues. Consequently, we chose the remaining ten questions where the correct answer rates were below 80% with small modifications on their phrasing. Q1–10 in Table 6 are the final set of the ten questions and corresponding answers that we used in our deployment study.

4 Deployment Study

A 15-day user study was conducted to evaluate DualCheck. The primary objective of our study is to examine the effect of microlearning supported by DualCheck rather than its human verification performance. As explained in Subsection 3.1, our current prototype does not integrate a human verification mechanism. The following user study protocol was approved by our institutional review board.

4.1 Task Design

We designed a deception study to avoid potential bias in the evaluation of DualCheck. In contrast to other microlearning systems, DualCheck offers implicit, opportunistic learning. Thus, we designed a study similar to an experience sampling method, probing participants’ Internet usage through a short questionnaire (e.g., how much time they spent on social networking sites on that day). We created four different sets of such questionnaires, and used randomly during the deployment study. We then included DualCheck at the bottom as a human verification task. The participants were then asked to respond to the questionnaire multiple times a day throughout the experiment. DualCheck showed one of the ten questions shown in Table 6. Each question was exposed to the participants four times throughout the experiment. The order of these questions was randomly shuffled, and they were presented to all participants in the same order.

4.2 Procedure

4.2.1 Day #1

We asked participants to review their consent forms and sign them. We then asked them to fill a pre-experimental questionnaire that included demographic questions. We also asked the 10 questions about online safety and ethics that were also used during the deployment study. However, we did not provide the answers to these questions. The performance on these 10 questions served as the baseline for the later analysis.

We then explained the tasks and questionnaire form used in our deployment study. The details of the DualCheck implementation was not explained to the participants, in particular DualCheck did not include an actual human verification mechanism.

4.2.2 Day #2–#14

We sent emails that included the link to our short online questionnaire three times a day between Days #2 and #13, and four times on Day #14. Our participants then filled the questionnaire and responded to the questions in DualCheck. We set two modes in DualCheck for a comparison of the learning performance: *OneTime* and *Repeat*. The *OneTime* mode indicates that DualCheck shows participants the correct answer and explanation immediately after they tick one of the choices. In the *Repeat* mode, DualCheck forced participants to respond to the question until they ticked the correct answer. When they initially ticked a wrong answer, the system showed the correct answer and explanation, and asked participants to update their responses. After they chose the correct answer, the system enabled the button to submit a form. This mode was derived from the behavior of existing CAPTCHA systems where users would need to succeed the given verification tasks to pass. We constantly monitored the participants' responses and reminded them if they had not responded an hour before the submission deadline of each questionnaire.

4.2.3 Day #15

At the end of the study, we first debriefed all participants and revealed that this was a deception study and informed them of the true objective of the study, examining how DualCheck would influence on the learning of the content of the questions. However, we did not reveal how the human verification was performed in DualCheck nor that DualCheck did not implement an actual human verification mechanism. They were then offered an explicit opportunity to withdraw themselves from this study, but none of them withdrew.

Subsequently, we asked them to complete the post-experimental questionnaire. This questionnaire comprised 2 sections. The first section contained 30 questions that gauge respondents' knowledge on online safety and

ethics. 10 of the questions were the same as those used in the deployment study. Another 10 questions were similar to the first ten, and were simply paraphrased to appear different (Q1a–10a in Table 7). The remaining 10 questions were new questions that participants had never given and were used as distractors. The presentation order of the questions was randomized for each participant. The second section was designed to probe the participants' experience and perceived usability of DualCheck. For usability assessment, we used the System Usability Scale (SUS) [1]. In addition, we included free-form questions to collect opinions on DualCheck.

4.3 Participants

We recruited 34 participants (25 females and 9 males; 6, 9, 13, 4, and 2 in their 20s, 30s, 40s, 50s, and 60s, respectively) using the same crowdsourcing service used in our question curation. None of the participants participated in the study related to the question curation. We randomly split the participants into two groups to compare the effects of the presentation modes of DualCheck: 16 for *OneTime* and 18 for *Repeat*. The participants were offered approximately 22 USD in local currency for the completion of the 14-day short online questionnaires. They were additionally offered 2.6 USD in local currency for the completion of the post-experimental questionnaire. All the participants completed the entire experiment including the post-experimental questionnaire.

4.4 Hypotheses

We summarize our hypotheses to test through our deployment study below:

- H1. The accuracies of the 10 questions used throughout the deployment study (Q1–10) would be higher at the post-experiment phase than the pre-experiment phase. This is because we expected participants to learn online safety and ethics through DualCheck.
- H2. The accuracies of the 10 questions that are similar to Q1–10 but only shown at the post-experimental questionnaire (Q1a–10a) would be higher compared to those by general Internet users who do not use DualCheck. This is because participants would develop relevant knowledge to answer these questions correctly through DualCheck.
- H3. The usability of DualCheck would be higher than that of Text-based CAPTCHA and picture-based reCAPTCHA. This is because the human verification is as simple as the checkbox-based reCAPTCHA.
- H4. The accuracies of Q1–10 in the *Repeat* mode would be higher than those in the *OneTime* mode. This is because participants would learn more by responding to questions until reaching the correct answers.

H5. The usability of the *Repeat* mode would be lower than that of the *OneTime* mode. This is because participants would be forced to choose the correct answer.

5 Results

5.1 Learning Performance

The primary objective of our deployment study is to examine the effect of DualCheck on microlearning. Thus, we first investigated the improvements in the accuracy (percentage of correct answers) for the ten questions used throughout the deployment study (Q1–10 in Table 6).

The mean accuracies of the ten questions (Q1–10) in the pre-experimental and post-experimental questionnaires were 0.68 ($SD=0.11$) and 0.94 ($SD=0.04$), respectively. We then conducted a two-way ANOVA test with the factors of the experiment phase (pre-experiment and post-experiment) and system mode (*OneTime* and *Repeat*). It revealed a significant result on the experiment phase ($F(1, 32)=89.38$, $p<.001$, generalized $\eta^2=.50$), but not the system mode ($F(1, 32)=1.12$, $p=.30$, generalized $\eta^2=.02$). This suggests significant improvements in accuracy for the ten questions our participants had seen during the deployment study.

We then further looked into the accuracy differences of these ten questions between the pre-experimental and post-experimental phases. Table 1 shows the accuracy breakdowns for the 10 questions (Q1–10). We conducted a binomial test for each question to better understand these differences. The binomial test is a statistical test that uses the binomial distribution to determine whether the proportion of data in two categories is significantly deviated from the theoretically-expected distribution. The accuracy in the post-experiment phase would be the same as in the pre-experiment phase if DualCheck did not contribute to participants’ learning effectively. Table 1 includes the 95% confidence intervals and p values derived from our binomial tests. All questions except Q2 revealed significant positive results. This result confirms strong positive learning effect of DualCheck.

We next looked into the performance of the ten questions similar to Q1–10, which were only exposed to our participants at the time of the post-experimental questionnaire (denoted as Q1a–Q10a). As these questions were not answered at the pre-experiment phase, we separately collected the reference accuracy for them through another crowdsourcing task. By taking a similar data collection method to our question curation, we recruited 50 new crowdsourcing participants (17 females and 33 males; 6, 16, 17, 7, and 4 in their 20s, 30s, 40s, 50s, and 60s, respectively) who had not participated in any study related to this project.

	Pre-test accuracy	Post-test accuracy	95% CI	p
Q1	0.79	0.97	[0.85, 1.00]	<.01 **
Q2	0.94	0.94	[0.80, 0.99]	1.00
Q3	0.68	0.91	[0.76, 0.98]	<.01 **
Q4	0.68	0.97	[0.85, 1.00]	<.001 ***
Q5	0.65	1.00	[0.90, 1.00]	<.001 ***
Q6	0.56	0.94	[0.80, 0.99]	<.001 ***
Q7	0.65	0.94	[0.80, 0.99]	<.001 ***
Q8	0.56	0.97	[0.85, 1.00]	<.001 ***
Q9	0.62	0.85	[0.69, 0.95]	<.01 **
Q10	0.68	0.94	[0.80, 0.99]	<.01 **

Table 1: The accuracies of Q1–10 observed in the pre-experimental and post-experimental questionnaire in the deployment study. In this and later tables, we also include the binomial test result for each question.

	Reference accuracy	Pre-test accuracy	95% CI	p
Q1	0.86	0.79	[0.62, 0.91]	.32
Q2	0.86	0.94	[0.80, 0.99]	.22
Q3	0.72	0.68	[0.50, 0.83]	.57
Q4	0.66	0.68	[0.50, 0.83]	1.00
Q5	0.68	0.65	[0.47, 0.80]	.71
Q6	0.48	0.56	[0.38, 0.73]	.39
Q7	0.80	0.65	[0.47, 0.80]	<.05 *
Q8	0.76	0.56	[0.38, 0.73]	<.05 *
Q9	0.62	0.62	[0.44, 0.78]	1.00
Q10	0.82	0.68	[0.50, 0.83]	<.05 *

Table 2: The accuracies of Q1–10 observed in a separate data collection study (denoted as “reference accuracy”) and the pre-experimental questionnaire in the deployment study.

	Reference accuracy	Post-test accuracy	95% CI	p
Q1a	0.74	0.85	[0.69, 0.95]	.17
Q2a	0.98	0.88	[0.73, 0.97]	<.01 **
Q3a	0.52	0.74	[0.56, 0.87]	<.05 *
Q4a	0.64	0.88	[0.73, 0.97]	<.01 **
Q5a	0.90	1.00	[0.90, 1.00]	<.05 **
Q6a	0.20	0.74	[0.56, 0.87]	<.001 ***
Q7a	0.80	0.91	[0.76, 0.98]	.13
Q8a	0.70	0.97	[0.85, 1.00]	<.001 ***
Q9a	0.72	0.71	[0.53, 0.85]	.85
Q10a	0.98	0.97	[0.85, 1.00]	.50

Table 3: The accuracies of Q1a–10a observed in a separate data collection study (denoted as “reference accuracy”) and the post-experimental questionnaire in the deployment study.

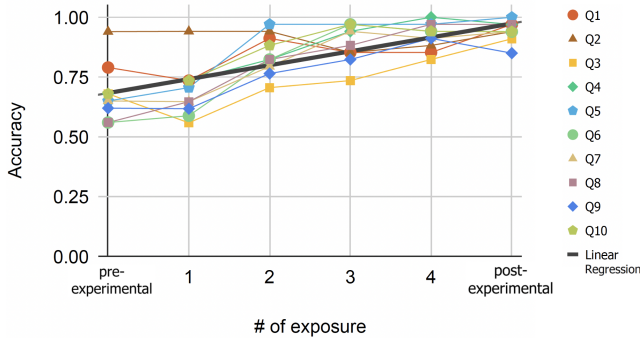


Figure 2: The accuracy transition across the number of exposure to the 10 questions. The plot includes participants’ overall performance in each of the four exposures as well that in the pre-experimental and post-experimental questionnaires, which results in the six measurement points. The regression result was $y = 0.06x + 0.68$ (adjusted $R^2=0.57$).

They were asked to respond to the 30 questions of Q1–10, Q1a–10a, and 10 distractor questions. We then derived the accuracy rates of Q1a–10a, which we regard as the reference accuracy, ultimately summarized in Table 3. Each crowdsourcing participant was compensated approximately 1.7 USD in their local currency at the completion of the task.

The average accuracy of Q1–10 in a separate data collection study explained in the previous paragraph was 0.73 ($SD=0.16$) while it was 0.68 ($SD=0.11$) in the pre-experimental questionnaire. Our t test did not find a significant result between these two groups ($t(49,33)=1.26$, $p=.21$, Cohen’s $d=0.13$). Table 2 shows the accuracy difference between the pre-experiment phase in our deployment study and a separate data collection study (denoted as “reference accuracy”). Our binomial tests confirmed significant differences in Q7, 8, and 10, where the accuracy in the pre-experimental questionnaire was lower. While we observed some accuracy differences, we did not find a significant difference in the average accuracies. We thus concluded that the performance comparison on Q1a–10a between these two groups would not be strongly biased in favor of either way.

The average accuracies of Q1a–10a were 0.72 ($SD=0.14$) and 0.94 ($SD=0.04$) in a separate data collection study and the post-experimental questionnaire, respectively. Our t test revealed a significant result between these two groups ($t(49,33)=8.49$, $p<.001$, Cohen’s $d=0.70$). Table 3 presents the accuracy difference on Q1a–10a between the post-experiment phase in our deployment study and a separate data collection study. Our binomial tests confirmed significant differences in 6 of the 10 questions (Q2a, Q3a, Q4a, Q5a, Q6a, and Q8a). All these significant results except Q2a were associated with higher accuracies in the post-experiment phase in our deployment study.

We further examined how the accuracies were improved during the experiment. Figure 2 presents the accuracies across

Verification system	Mean SUS (SD)
Text-based CAPTCHA	54.45 (15.46)
Picture-based reCAPTCHA	53.10 (18.82)
Checkbox-based reCAPTCHA	80.60 (13.28)
DualCheck <i>OneTime</i>	69.38 (13.02)
DualCheck <i>Repeat</i>	78.47 (13.96)
DualCheck average of both modes	74.19 (14.10)

Table 4: The mean SUS scores and their standard deviations of DualCheck and existing human verification systems.

questions and the number of exposures. As explained above, participants saw each of the ten questions four times. Our linear regression analysis shows a significant effect of the number of exposure (estimated coefficient: 0.07, $p<.001$). The goodness of fit was .52 (adjusted R^2). Due to large variances in the accuracies we observed in the deployment study, the fitting was not very strong. However, our analysis results confirm an increasing trend of accuracies, suggesting a positive learning effect caused by DualCheck.

5.2 Usability Comparison

We next examined the usability of DualCheck through the System Usability Scale (SUS) [1]. To better understand the SUS results, we conducted another data collection on the SUS scores of the existing CAPTCHA systems. They included text-based CAPTCHA, picture-based reCAPTCHA, and the reCAPTCHA Checkbox. We designed another task to collect these SUS scores in the same crowdsourcing service. All participants were offered an opportunity to participate in this data collection and a compensation of approximately 1 USD in the local currency. Consequently, 50 new participants who did not participate in our question curation or deployment study participated in this scoring task.

Table 4 presents the average SUS scores and the standard deviations of DualCheck and the three human verification systems mentioned above. A one-way ANOVA revealed significant differences in the factors of the human verification interfaces ($F(3,180)=37.51$, $p<.001$, generalized $\eta^2=.63$). Our Scheffe’s test further showed that the SUS score of DualCheck was significantly higher than those of text-based CAPTCHA ($p<.001$) and picture-based CAPTCHA ($p<.001$). Our t test did not find a significant difference between the *OneTime* and *Repeat* modes in DualCheck ($t(15,17)=-1.96$, $p=0.06$, Cohen’s $d=0.67$). These statistical results confirm that the perceived usability of DualCheck was significantly higher than that of text-based CAPTCHA and picture-based reCAPTCHA.



Figure 3: The distribution of the responses about question difficulty.



Figure 4: The distribution of the responses about whether participants felt that they were able to acquire new knowledge about online safety and ethics through DualCheck.

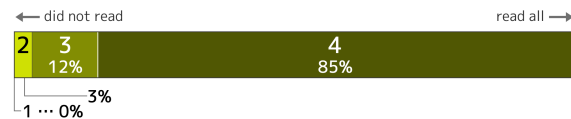


Figure 5: The distribution of the responses about how carefully participants read the questions, correct answers, and explanations.

5.3 Impressions on Questions in DualCheck

We further analyzed the participants’ responses to our questions about the questions presented in DualCheck. Figure 3 shows the distribution of the participants’ responses to the question about the overall difficulty of the questions they saw in DualCheck (1: The questions were too easy–5: The questions were too difficult). 20 participants (59%) considered that the questions were at the appropriate level, confirming that our question curation was properly executed. Figure 4 summarizes how strongly participants agreed that they were able to acquire new knowledge about online safety and ethics through questions provided by DualCheck. All but one participant agreed that they were able to learn through the questions. Figure 5 shows the participants’ responses to the question about whether they thought they read the correct answers and explanations on a 4-point Likert scale. 29 participants (85%) responded that they read questions, correct answers, and explanations. All of these results suggest participants’ positive experience with DualCheck.

5.4 Qualitative Results

We further examined the comments we received through open-ended questions to deepen our understanding of participants’ experiences with DualCheck. Two of the authors jointly conducted thematic analysis and developed six themes that categorize the quotes of comments for overall deployment study. We discarded the quotes that these two authors disagreed in categorization. As a result, all categories had the perfect agreement between the two authors. Table 5 shows

Theme and Subtheme	# quotes
Questions	
Question difficulty	15
Issues on question presentation	9
Issues on answer explanations	5
Advantages of DualCheck	
Perceived advantages	15
Usability of DualCheck	
Positive opinions on usability	11
Issues on usability	10
Suggestions	
Possible improvements	8

Table 5: The categorization of participants’ comments collected in the deployment study. We note that we only considered the comments that two of the authors agreed in their categorization and used for our analysis. Thus, all the categories above exhibited the perfect agreement.

our categorization and quote occurrence for each category. We note that the quotes presented below were originally written in our local language, and we translated them into English as faithfully as possible for the report in this paper.

5.4.1 Perceived Benefits of DualCheck

We observed explicit comments where participants appreciated DualCheck for offering unique microlearning opportunities. For instance, P28 and P33 offered their appreciation on DualCheck over existing human verification systems by highlighting its direct benefits to users.

I’m worried about phishing scams and other sophisticated scams these days, so I thought it would be good to have a lot of such problems. This is much better for learning than doing puzzles that are not easy to use, so I would like to see this implemented in general Websites. [P33]

I thought it would be more interesting than a bot detection system that requires input of known illegible strings, and it would kill two birds with one stone because it would be simple and learnable. [P28]

We further examined the participants’ responses to an open-ended question about which questions were the most memorable. Fourteen participants explicitly mentioned that the question about cookies (Q6) was the most memorable. P28 and P29 shared the following comments about Q6.

I have gained more knowledge about information literacy in general, which I had been unclear about. In particular, I have gained accurate and clear knowledge about cookies. I also learned that I should be careful about key-marked sites, which I had blindly trusted in the past. [P28]

I learned a lot because I knew the name of cookies, but not the details. [P29]

Q1, Q3, and Q9 were mentioned by 5, 6, and 4 participants, respectively. P6 commented on how memorable Q3 was and how it promoted awareness of the SSL presentation and the URL in a browser.

There was a lot of information that I didn't know, but the question on URLs starting with https:// left a particular impression on me. I don't usually check URLs, so I thought I'd pay attention to it from now on. [P6]

Both the quantitative and qualitative results strongly confirm the benefits of DualCheck, particularly its capability to offer microlearning opportunities.

5.4.2 Possible Improvements

Our participants suggested several improvements to DualCheck. Five participants explicitly commented that they would like to see more variation in the questions. Undo and redo features were common requests; they were suggested by two participants who were grouped into the *OneTime* condition. Our SUS comparison did not indicate statistically significant differences between the two presentation modes. Thus, the *Repeat* mode may solve these issues. Future studies should examine how to fine-tune the interface settings of DualCheck to improve the learning experience while reducing users' cognitive load. In general, DualCheck successfully encouraged participants to read questions carefully.

I thought it was very good that I could study every time. The fact that the questions are repeated every day, and that I can't re-select the options, allows me to concentrate on reading the questions and learn about things that I've only vaguely been familiar with. [P32]

The same participant also commented that the question content would substantially impact on the user experience of DualCheck. This may suggest a future research direction of personalization on topics.

I felt that it was very annoying for those who were not interested in the content, because it took a lot of brainpower to prove that I was not a bot. I was also interested in the content of this problem, so I felt I learned a lot, but if it had been a fashion problem, for example, I would have hated it. [P32]

Participants were motivated to receive more detailed explanations. The general opinion was that these improvements would not only make explanations more accessible to the general user populations but also help users learn online safety and ethics by themselves.

I remember that I always answered the same question wrong. As for the safety of the Internet, even though I understood what I should not do (such as not clicking on links unnecessarily), I did not understand the technical terms (such as domain names) properly, so I think I answered some of the questions on a hunch. It would have been nice to have a simple explanation of these IT terms. [P1]

For example, when the question is about "writing with storage services", I thought it would be good to have one or two examples of service names to show what kind of storage services are available. I was a little confused at first if it was the one I was thinking of or not. I also thought that it might be difficult to understand for people who have never used that service before. [P8]

Participants also suggested dynamic adjustment of difficulty depending on people's correct responses, more complex response styles (e.g., the "Choose all that apply" response style), and integration of gamification (e.g., awarding points for correct responses).

6 Discussion

As shown through our quantitative results, we observed positive learning effects of DualCheck. The subjective ratings and open-ended comments we obtained in the post-experimental questionnaire also support participants' positive experience in learning online safety and ethics. We conclude that our results support H1.

The accuracy of the ten questions used throughout the deployment study (Q1–10) had significant improvements except for Q2. The accuracy of Q2 was 0.94 even at the time of the pre-experiment phase, and it remained the same after the experiment. We do not have clear reasons why only Q2 exhibited such high accuracy. The accuracy of the remaining questions in the pre-experiment phase was below 80%, which were in line with our results during the question curation. We thus concluded that our question choice was appropriate in general.

5 of the 10 questions similar to Q1–Q10 and asked only in the post-experimental questionnaire (Q3a, Q4a, Q5a, Q6a, and Q8a) showed significant accuracy improvements compared to the reference accuracy. This is a promising result as participants were able to extend their knowledge to answer unseen questions correctly to some extent.

The accuracy of Q2a in the deployment study was significantly lower than the reference accuracy. This result might be related to the fact that participants did not have improvements in the accuracy of Q2. Our deployment participants were able to answer correctly from the beginning and thus might not have paid careful attention to the explanation offered by the system. This result suggests

that a future system should provide variations of the same questions (e.g., paraphrasing or converting expressions from the affirmative to the negative form) or different questions about the same topic to reinforce users' learning. In conclusion, H2 is not fully supported in this study.

Although DualCheck increases the overall performance time for human verification tasks, the usability assessment we obtained showed a higher rating for DualCheck than text-based CAPTCHA and picture-based reCAPTCHA. Our qualitative evidence also suggests that participants were able to explicitly observe the learning benefits of DualCheck, which could contribute to its higher perceived usability. Our SUS results showed that we did not have a significant result between DualCheck and checkbox-based reCAPTCHA are interpretable because interaction requested by both systems was equivalent. We thus conclude that our results support H3. This result also suggest that users could be more willing to engage in microlearning during human verification tasks because they can perceive more direct benefits to them.

Our results did not reveal strong evidence about the two presentation modes of DualCheck in terms of learning effect and perceived usability. Thus, H4 and H5 are not supported. However, other factors, such as question content, the frequency of presenting the same question, and users' personal preferences, might have influenced this result, and future work should further examine these effects.

7 Limitations and Future Work

There are several limitations to be discussed to clarify the scope and contributions of this work. We recruited our study participants through a crowdsourcing service available in the country of authors. This implies that our participants might have been more accustomed to using online services and human verification systems than the general user populations. As they could be considered active Internet users, they might be more attentive to online safety and ethics, which might have led to a positive bias toward DualCheck. Future work should conduct a wider scale of user studies to validate the effect of DualCheck.

We took the design of an experience sampling method for our deployment study to offer repeated exposure to DualCheck. In a more realistic setting, users would not see our system as frequently as our deployment study. Thus, understanding the learning effect of DualCheck in a more realistic setting requires additional studies.

While our current investigation focused on online safety and ethics questions, future work may expand the scope to other kinds of privacy and safety threats and practices (e.g., fraud in the physical world and fake news). The results of our study anticipate positive learning effects on these topics, and further examinations are encouraged.

Another important future research direction is to investigate the effect of question and response formats. Different question

formats (e.g., dichotomous or free-form questions) might have different learning effect. Similarly, response methods can also influence on learning behavior. Even using the same question, users might exhibit different accuracy rates depending on the question and response formats. Our current implementation utilizes an interaction modality derived from reCAPTCHA v2 (ticking a checkbox), but advanced CAPTCHA systems does not even require explicit interaction like reCAPTCHA-v3. With such technology, a future system can completely decouple human verification and interaction for microlearning, which would allow researchers to explore different forms of microlearning. Our work serves as a foundation of such future work to integrate human verification and microlearning.

The administration of questions is necessary in a practical setting. Officers in charge of information management for organizations may take this responsibility to employ DualCheck for their members. In particular, we envision that DualCheck can complement existing learning activities at educational institutions. Future work should examine the longer-term effect of DualCheck as well as its deployment in a more practical setting.

8 Conclusion

Learning online safety and ethics is becoming more critical. However, they lack such learning opportunities and are often left behind. We introduce DualCheck, a microlearning system that is integrated into human verification tasks. Users are asked to respond to questions related to online safety and ethics while human verification would be executed in a similar manner to reCAPTCHA v2. In this manner, DualCheck offers users microlearning opportunities when they use online services. Our 15-day user study confirmed the positive learning effect of DualCheck. The quantitative and qualitative results also supported participants' positive attitudes toward DualCheck. The usability of DualCheck was rated significantly higher than those of text-based CAPTCHA and picture-based reCAPTCHA. We plan to further investigate the effect of DualCheck by expanding our studies to a wider user population and incorporating more learning topics.

	Statement and answer
Q1	A: Connecting a USB flash drive to a computer in public is a security risk. B: Charging a smartphone via USB on a computer in public is a security risk. Correct Answer: Both statements are correct.
Q2	A: On social networking sites, there is no privacy problem in sharing selfies and other information if you give limited access. B: On social networking sites, if you don't post any personal information, your identity will not be identified. Correct Answer: Both statements are wrong.
Q3	A: This is the first time I visited this Website, but I thought it was safe because it had a key symbol on my browser, so I entered my personal information. B: I entered my personal information on a Website beginning with http:// . It is risky to enter personal information on such a Website. Correct Answer: Only statement B is correct.
Q4	A: Passwords should be a combination of letters, numbers, and symbols that are difficult to remember. B: Passwords are safer if they are based on personal information, such as your hobbies, and avoid famous words that are easily guessed. Correct Answer: Only statement A is correct.
Q5	A: When the earthquake struck, local people posts the situation in the area. Even if you don't know whether it is true information, it is better to share the information quickly. B: When spreading information when an earthquake or other event occurs, it is better to only spread posts by the government or news organizations. Correct Answer: Only statement B is correct.
Q6	A: A cookie is a piece of information that sends a user's name and other personal information to a site administrator. B: Cookies are used for retargeting advertisements and other purposes. Correct Answer: Only statement B is correct.
Q7	A: Documents created with online storage services and document creation tools are not disclosed to the public. B: Documents created with online services can be seen by others through searches. Correct Answer: Only statement B is correct.
Q8	A: The procedure for requesting information about an offensive social networking account has been made easier due to a change in the law. B: Even if there is an offensive SNS account, it is difficult to identify their source address. Correct Answer: Only statement A is correct.
Q9	A: To verify that the email you received was sent from a real bank or other sources, you check the back of the @ in the source address. B: Checking the domain is one of the most important things to ensure that the URL sent to you is authentic. Correct Answer: Only statement B is correct.
Q10	A: Photos taken with a smartphone may contain location information. B: If you post a photo without the location information to a social networking site, your location will not be identified. Correct Answer: Only statement A is correct.

Table 6: The questions used in this work. Q1–10 are derived from our question curation process. They were originally written in the local language of the authors, and are translated into English as faithfully as possible.

Q1a	<p>A: If you use a computer’s USB port only to charge your smartphone, no viruses or other devices will be transferred.</p> <p>B: If you connect a USB flash drive to a shared computer, viruses and other malicious programs may be copied.</p> <p>Correct Answer: Only statement B is correct.</p>
Q2a	<p>A: On social networking sites, if you limit the number of people you can follow, there is no problem if you tweet personal information.</p> <p>B: Your identity can be identified based on your following relationship on social networking sites.</p> <p>Correct Answer: Only statement B is correct.</p>
Q3a	<p>A: Websites that start with http://... do not support encrypted communication.</p> <p>B: If the Website is capable of encrypted communication, it is safe to send personal information.</p> <p>Correct Answer: Only statement A is correct.</p>
Q4a	<p>A: Passwords should be a meaningless string of characters with symbols.</p> <p>B: It is preferable to create a password based on a hobby or something that you keep secret from others.</p> <p>Correct Answer: Only statement A is correct.</p>
Q5a	<p>A: An earthquake occurred, but there was no information from the news media or government, so I spread a post made by a person claiming to be a local.</p> <p>B: When the earthquake occurred, a person claiming to be a scholar on Twitter explained the situation. It is considered as credible information.</p> <p>Correct Answer: Both statements are wrong.</p>
Q6a	<p>A: The use of cookies can customize ads.</p> <p>B: Allowing the use of cookies is likely to leak personal information.</p> <p>Only statement A is correct.</p>
Q7a	<p>A: Documents created with online document creation tools are not likely to show up in a Web search.</p> <p>B: It is important to check the publication settings of documents created with online tools.</p> <p>Correct Answer: Only statement B is correct.</p>
Q8a	<p>A: It is difficult to identify the source address of an anonymous social networking account.</p> <p>B: You can file a request for disclosure of sender information against an offensive social networking account.</p> <p>Correct Answer: Only statement B is correct.</p>
Q9a	<p>A: Checking the domain of the URL is important to confirm whether it is genuine or not.</p> <p>B: I received an email claiming to be from my bank. It was the same domain as the bank’s email, so I figured it was the right email.</p> <p>Correct Answer: Only statement A is correct.</p>
Q10a	<p>A: The scenery and objects in the photo could lead to the identification of personal information.</p> <p>B: Location information may be stored in the photo.</p> <p>Correct Answer: Both statements are correct.</p>

Table 7: The 20 questions used in this work. We created another 10 questions (Q1a–10a) that are similar to Q1–10 to measure the deployment study participants’ learning. They were originally written in the local language of the authors, and are translated into English as faithfully as possible.

Acknowledgements

We appreciate our lab members for giving us very helpful feedback and advice. Especially, Anran Xu offered great help on related work and thoughtful insights on this project. We also thank all participants for their help. This research is part of the results of Value Exchange Engineering, a joint research project between Mercari, Inc. and the RIISE.

References

- [1] John Brooke. SUS—a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.
- [2] Carrie J. Cai, Philip J. Guo, James R. Glass, and Robert C. Miller. Wait-learning: Leveraging wait time for second language education. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3701–3710. Association for Computing Machinery, 2015.
- [3] Tilman Dingler, Dominik Weber, Martin Pielot, Jennifer Cooper, Chung-Cheng Chang, and Niels Henze. Language learning on-the-go: Opportune moments and design of mobile microlearning sessions. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17. Association for Computing Machinery, 2017.
- [4] Valerie Fanelle, Sepideh Karimi, Aditi Shah, Bharath Subramanian, and Sauvik Das. Blind and human: Exploring more usable audio CAPTCHA designs. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 111–125. USENIX Association, August 2020.
- [5] Galen A. Grimes, Michelle G. Hough, Elizabeth Mazur, and Margaret L. Signorella. Older adults' knowledge of internet hazards. *Educational Gerontology*, 36:173 – 192, 2010.
- [6] Japan Information-technology Promotion Agency. Awareness survey on information security ethics in fy2019 report. <https://www.ipa.go.jp/files/000080783.pdf>, 2019. (Written in Japanese).
- [7] Japan Information-technology Promotion Agency. Awareness survey on information security threats in FY2019 report. <https://www.ipa.go.jp/files/000080784.pdf>, 2019. (Written in Japanese).
- [8] Japan Information-technology Promotion Agency. 10 major threats to information security 2021. <https://www.ipa.go.jp/files/000088835.pdf>, 2021. (Written in Japanese).
- [9] Ulrike Meyer and Vincent Drury. Certified phishing: taking a look at public key certificates of phishing websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 211–223. USENIX Association, 2019.
- [10] Gona Sirwan Mohammed, Karzan Wakil, and Sarkhell Sirwan Nawroly. The effectiveness of microlearning to improve students' learning ability. *International Journal of Educational Research Review*, 3(3):32–38, 2018.
- [11] Ministry of Education, Culture, Sports, Science, and Technology. “Information I”, Chapter 1 of the teaching materials of information science for high school teachers. https://www.mext.go.jp/content/20200722-mxt_jogai02-100013300_003.pdf, 2019. (Written in Japanese).
- [12] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 259–284. USENIX Association, August 2020.
- [13] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI conference on human factors in computing systems*, CHI '10, pages 373–382. Association for Computing Machinery, 2010.
- [14] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security (SOUPS 2007)*, pages 88–99. Association for Computing Machinery, 2007.
- [15] Nitirat Tanthavech and Apichaya Nimkoompai. Captcha: Impact of website security on user experience. In *Proceedings of the 2019 4th International Conference on Intelligent Information Technology*, ICIIT '19, page 37–41. Association for Computing Machinery, 2019.
- [16] Andrew Trusty and Khai N. Truong. Augmenting the web for second language vocabulary learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 3179–3188. Association for Computing Machinery, 2011.

- [17] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. Captcha: Using hard ai problems for security. In Eli Biham, editor, *Advances in Cryptology — EUROCRYPT 2003*, pages 294–311, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- [18] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.
- [19] Takumi Yamamoto, Tokuichiro Suzuki, and Masakatsu Nishigaki. A proposal of four-panel cartoon captcha. In *2011 IEEE International Conference on Advanced Information Networking and Applications*, pages 159–166, 2011.
- [20] Jeff Yan and Ahmad Salah El Ahmad. Usability of captchas or usability issues in captcha design. In *Proceedings of the 4th Symposium on Usable Privacy and Security (SOUPS 2008)*, page 44–52. Association for Computing Machinery, 2008.

A Questions used during our Question Curation

Table [A.1](#) had 25 questions (Qc1 – Qc25) through the question curation phase. We then collected the percentage of the correct answers. shows the questions and their accuracies.

B Distractor questions used in the pre-expermental and post-expermental questionnaire

Table [B.1](#) shows the 10 distractor questions used in the pre-expermental and post-expermental questionnaire.

ID	accuracy	Statements
Qc1	0.87	A: Even if you post anonymously, there is a chance that you will be identified. B: I want to say something bad about my friend, but if I do so directly, it will damage our relationship, so I post it on an anonymous forum. Correct Answer: Only statement A is correct.
Qc2	0.84	A: I received an email from a web service I use that ask me to change my password. The URL contained the company's name, so I assumed it was a real site and logged in. B: You need to be careful when click websites' links because scam sites can appear higher position in web searches. Correct Answer: Only statement B is correct.
Qc3	0.98	A: A friend sent me a link to a website he recommended. It is safe because it came from a trusted friend. B: Even if the link was sent by a friend, you need to check the URL carefully. Correct Answer: Only statement B is correct.
Qc4	0.89	A: In order to get more people to watch my favorite drama, I posted a scene from that drama on social networking sites to spread the word. B: Pictures and other materials posted by individuals are not registered with the Patent Office and are not copyrighted, so they may be freely reproduced. Correct Answer: Both statements are wrong.
Qc5	0.73	A: On social networking sites, there is no privacy problem in publishing selfies and other photos as long as the account is limited public. B: On social networking sites, as long as you don't post any personal information, your identity will not be identified. Correct Answer: Both statements are wrong.
Qc6	0.94	A: The advantage of anonymous message boards is that people can post easily, and there is no problem if they post wrong things. B: Anonymous forums can be dangerous as inaccurate content may be posted. Correct Answer: Only statement B is correct.
Qc7	0.85	A: If you connect to a wireless LAN from a trusted provider, you do not have to worry about others seeing your communications. B: Before connecting to a free wireless LAN, you should thoroughly check the terms and rules of use of the wireless LAN. Correct Answer: Only statement B is correct.
Qc8	0.69	A: Connecting a USB flash drive to a computer in an Internet cafe, etc. is a security risk. B: Charging smartphones via USB on computers in Internet cafes, etc., is a security risk. Correct Answer: Both statements are correct.
Qc9	0.82	Which of the following is the correct address for Google? A: https://google-co.jp B: https://google.co.jp C: https://google.co.jp D: https://goog1e.co.jp Correct Answer: B is the correct URL.
Qc10	0.83	A: Photos taken with a smartphone may contain location information. B: If you post a photo to a social networking site, the location information are removed automatically, so your location will not be identified. Correct Answer: Only statement A is correct.
Qc11	0.95	A: I received an email I don't recognize. There was a link to unsubscribe, so I clicked on it and took the necessary steps to unsubscribe. B: While browsing a website, the message "This smartphone has been compromised" was displayed, so I followed the instructions on the screen. Correct Answer: Both statements are wrong.
Qc12	0.54	A: Passwords should be a combination of letters, numbers, and symbols that are difficult to remember. B: Passwords are safer if you avoid famous words that can be easily guessed, and create passwords based on personal things like your hobbies. Correct Answer: Only statement A is correct.
Qc13	0.61	A: When an earthquake occurred, people were sending out information about the area. Even if the authenticity of the information is unknown, it is better to spread the information quickly. B: When spreading information after an earthquake or other event, it is better to only spread posts from the government or news organizations. Correct Answer: Only statement B is correct.

ID	accuracy	Statements
Qc14	0.62	A: This is the first time I visited a website, but my browser had a key symbol on it, so I thought it was safe and entered my personal information. B: It is risky to enter personal information on a website that begins with http:// . Correct Answer: Only statement B is correct.
Qc15	0.91	A: I saw information about COVID-19 on a social networking site. Since the profile said the author was a doctor, I thought it was correct and spread it. B: Several people mentioned the information about COVID-19, so I thought it was correct and spread it. Correct Answer: Both statements are wrong.
Qc16	0.88	A: Fingerprint and face recognition are not vulnerable to being breached because only you can unlock. B: Fingerprint and face recognition enhance security when they are combined with password locks. Correct Answer: Only statement B is correct.
Qc17	0.91	A: I got a warning message while browsing a website. It instructed me to install an application in the Google Play/App Store, so I downloaded it, thinking it was safe. B: An advertisement recommended an application. It was highly rated in the app store, so I thought it was safe and downloaded it. Correct Answer: Both statements are wrong.
Qc18	0.98	A: I posted a picture I liked that I found on a social networking site, claiming it to be my own work. B: A music program I forgot to record was reprinted on a social networking site, so I downloaded it to watch it later. Correct Answer: Both statements are wrong.
Qc19	0.69	A: Documents created with online storage services and document creation tools are never made available to the outside world. B: Documents created with online services may be seen by others through searches. Correct Answer: Only statement B is correct.
Qc20	0.88	A: I posted a photo of myself with a friend under a limited public access permission on an SNS at my own discretion. B: A post such as “The train I’m on is delayed” could identify where I live, etc. Correct Answer: Only statement B is correct.
Qc21	0.47	A: A cookie is a piece of information that sends a user’s name and other personal information to a site administrator. B: Cookies are used for targeted advertisement and other purposes. Correct Answer: Only statement B is correct.
Qc22	0.61	A: Legal changes have made it easier to request information about offensive social networking accounts. B: Even if there is an offensive social network account, it is difficult to identify the source of the slander. Correct Answer: Only statement A is correct.
Qc23	0.61	A: To minimize the damage caused by ransomware, backups need to be taken regularly. B: If you are a victim of ransomware, you will only lose the use of your data, which is not a problem if you have proper backups. Correct Answer: Only statement A is correct.
Qc24	0.85	A: Two-factor authentication can be set up to reduce the risk of unauthorized login. B: Two-factor authentication may include biometrics and one-time passwords. Correct Answer: Both statements are correct.
Qc25	0.41	A: To verify that an email you receive is from a real bank or other organization, just look at the back of the @ in the source address. B: One of the most important things to make sure that the URL sent to you is authentic is to check the domain. Correct Answer: Only statement B is correct.

Table A.1: 25 questions used during our Question Curation. They were originally written in a local language where the authors curated the questions, and are translated into English as faithfully as possible.

ID	Statements
D1	A: To minimize the damage caused by ransomware, backups need to be taken regularly. B: If you are a victim of ransomware, you will only lose the use of your data, which is not a problem if you have proper backups. Correct Answer: Only statement A is correct.
D2	A: Even if you post anonymously, there is a chance that you will be identified. B: I want to say something bad about my friend, but if I do so directly, it will damage our relationship, so I post it on an anonymous forum. Correct Answer: Only statement A is correct.
D3	A: A friend sent me a link to a website he recommended. It is safe because it came from a trusted friend. B: Even if the link was sent by a friend, you need to check the URL carefully. Correct Answer: Only statement B is correct.
D4	A: In order to get more people to watch my favorite drama, I posted a scene from that drama on social networking sites to spread the word. B: Pictures and other materials posted by individuals are not registered with the Patent Office and are not copyrighted, so they may be freely reproduced. Correct Answer: Both statements are wrong.
D5	A: The advantage of anonymous message boards is that people can post easily, and there is no problem if they post wrong things. B: Anonymous forums can be dangerous as inaccurate content may be posted. Correct Answer: Only statement B is correct.
D6	A: I received an email I don't recognize. There was a link to unsubscribe, so I clicked on it and took the necessary steps to unsubscribe. B: While browsing a website, the message "This smartphone has been compromised" was displayed, so I followed the instructions on the screen. Correct Answer: Both statements are wrong.
D7	A: I saw information about COVID-19 on a social networking site. Since the profile said the author was a doctor, I thought it was correct and spread it. B: Several people mentioned the information about COVID-19, so I thought it was correct and spread it. Correct Answer: Both statements are wrong.
D8	A: Fingerprint and face recognition are not vulnerable to being breached because only you can unlock. B: Fingerprint and face recognition enhance security when they are combined with password locks. Correct Answer: Only statement B is correct.
D9	A: I got a warning message while browsing a website. It instructed me to install an application in the Google Play/App Store, so I downloaded it, thinking it was safe. B: An advertisement recommended an application. It was highly rated in the app store, so I thought it was safe and downloaded it. Correct Answer: Both statements are wrong.
D10	A: I posted a picture I liked that I found on a social networking site, claiming it to be my own work. B: A music program I forgot to record was reprinted on a social networking site, so I downloaded it to watch it later. Correct Answer: Both statements are wrong.
D11	A: I received an email from a web service I use that ask me to change my password. The URL contained the company's name, so I assumed it was a real site and logged in. B: You need to be careful when click websites' links because scam sites can appear higher position in web searches. Correct Answer: Only statement B is correct.
D12	A: If you connect to a wireless LAN from a trusted provider, you do not have to worry about others seeing your communications. B: Before connecting to a free wireless LAN, you should thoroughly check the terms and rules of use of the wireless LAN. Correct Answer: Only statement B is correct.

Table B.1: Distractor questions used in the pre-experimental and post-experimental questionnaire. We chose 10 questions from this set for each questionnaire. They were originally written in a local language where the authors conducted the user study, and are translated into English as faithfully as possible.

C The Experience Sampling Method Interface Used in Our Study

Figure C.1 shows the screenshot of the questionnaire we used in our survey. It consists of questionnaire for ESM and DualCheck. The ESM part asked participants to answer their recent Internet usage (hours they had spent in SNS, shopping sites, and news sites) in the example below.

*以下の設問で、所要時間等はすべて、半角で分数をご記入ください。(例：SNSの利用時間:120)
記入は最大可能な時間で差し支えありません。

[必須]Q0: Crowdworksの表示名をご記入ください。
表示名は、数字のIDではなく、ワーカーさん自身がお決めた名前です。

[必須]Q1: 今日、これまでにSNSを利用した時間を教えてください。概算で構いません。

[必須]Q2: 今日、これまでにネットショッピングサイトなどを利用した時間を教えてください。概算で構いません。

[必須]Q3: 今日、これまでにインターネットニュースなどを閲覧した時間を教えてください。概算で構いません。

[任意]Q4: その他に1時間以上利用したインターネットサービスがあれば、概要を教えてください。(「インターネットゲーム」「Webメール」など)

EDU CAPTCHA 私はロボットではありません
ボットによる投稿でないことを確認するため、以下の質問にお答えください。
次の文章のうち、正しいものを選んでください。
A: 受信したメールが本物の銀行などから送られたものか確かめるためには、送信元アドレスの@の後ろを見ればよい。
B: 送られてきたURLが本物が確認するために重要なことの一つに、ドメインを確認することがある。
一度選ぶと選びなおせません。ご注意ください。
 Aのみ正しい
 Bのみ正しい
 両方正しい
 両方誤り
 わからない

Figure C.1: The screenshot of the questionnaire we used during the deployment study.

D Post-experimental Questionnaire

We asked participants to complete post-experimental questions at the end of the study. The questionnaires consisted of two parts; the first part included 30 questions to gauge knowledge of online safety and ethics, and the second part was to probe the participants' experience and perceived usability of DualCheck. This section includes the questions we used in the second part. They were originally written in the local language of the authors, and are translated into English as faithfully as possible.

We referred DualCheck as “CAPTCHA Quiz” in this questionnaire.

- Please fill your ID of crowdsourcing service account.
- Please answer the following questions about your comfort with the CAPTCHA quiz. (We used SUS for this part.)
 - I think that I would like to use this system frequently.
 - I found the system unnecessarily complex.
 - I thought the system was easy to use.
 - I think that I would need the support of a technical person to be able to use this system.
 - I found the various functions in this system were well integrated.
 - I thought there was too much inconsistency in this system.
 - I would imagine that most people would learn to use this system very quickly.
 - I found the system very cumbersome to use.
 - I felt very confident using the system.
 - I needed to learn a lot of things before I could get going with this system.
- Please let us know if you have any feedback on the usability of the CAPTCHA quiz. Please tell us about any difficulties you had in operating the system or any points that made it easier to use. You can answer in free-form.
- Please answer the following items.
 - Overall, how difficult did you find the quiz? (1: Too easy – 5: Too difficult)
 - Do you think you gained new knowledge through this quiz? (1: Not at all – 5: Very much)
- How much did you read about the question and explanations of the CAPTCHA quiz? You can choose from the statements below.

- I answered randomly and did not read the questions, correct answers, or explanations.
 - I read the questions, but not the correct answers and explanations
 - I read the questions and checked the correct answers, but did not read the explanations.
 - I read all the questions, correct answers, and explanations
- Please tell us about any particularly memorable content or new knowledge you learned in the CAPTCHA quiz. You can answer in free form.
- Please let us know any comments you have about the questions in the CAPTCHA quiz (e.g., They were too easy, too difficult, or any doubts about the answers). You can answer in free form.
- Were you aware that the original purpose of the survey was the experiment for CAPTCHA quiz? You can choose from the statement below;
 - I was aware that the purpose of the survey was to investigate CAPTCHA quiz.
 - I felt that there might be another purpose of the study
 - I was not aware of it.
- Please let us know if you have any comments or advice regarding the mechanism or content of the CAPTCHA quiz. We would be happy to hear any suggestions you may have, such as how we could improve the functionality or content of the quiz.

Understanding Non-Experts' Security- and Privacy-Related Questions on a Q&A Site

Ayako A. Hasegawa
NICT

Naomi Yamashita
NTT / Kyoto University

Tatsuya Mori
Waseda University / NICT / RIKEN AIP

Daisuke Inoue
NICT

Mitsuaki Akiyama
NTT

Abstract

Non-expert users are often forced to make decisions about security and privacy in their daily lives. Prior research has shown that non-expert users ask strangers for advice about digital media use online. In this study, to clarify the security and privacy concerns of non-expert users in their daily lives, we investigated security- and privacy-related question posts on a Question-and-Answer (Q&A) site for non-expert users. We conducted a thematic analysis of 445 question posts. We identified seven themes among the questions and found that users asked about cyberattacks the most, followed by authentication and security software. We also found that there was a strong demand for answers, especially for questions related to privacy abuse and account/device management. Our findings provide key insights into what non-experts are struggling with when it comes to privacy and security and will help service providers and researchers make improvements to address these concerns.

1 Introduction

Security and privacy technologies are generally difficult for non-experts to understand and use because of the complexity of these concepts [78]. Indeed, researchers have demonstrated that misconceptions regarding security and privacy technologies are ingrained and pervasive in non-expert users [86, 89]. Today, security and privacy technologies are incorporated into every device and service. Non-expert users are often forced to make decisions about security and privacy in their daily lives [20, 65], such as whether

to permit apps to access their personal data [7] or whether to proceed against browser warnings [72]. They are therefore likely to have a variety of security and privacy concerns.

According to a study that investigated the advice sources of non-expert users pertaining to digital media use, 43% of young adults ask strangers online as well as family and friends for advice [54]. Hence, we can expect that Question-and-Answer (Q&A) sites for non-expert users contain many security- and privacy-related questions that non-expert users have in their daily lives. In the security and privacy research community, researchers have successfully identified the security and privacy concerns of developers during their development work by analyzing questions posted on Stack Overflow, a Q&A site for developers and programmers [47, 63, 88, 97]. However, little is known about security- and privacy-related questions posted on Q&A sites for non-expert users. By analyzing such questions, we can identify the issues these users face in their daily lives and provide insights to help stakeholders (e.g., service providers and security researchers) address these problems.

In this study, to clarify the security and privacy concerns of non-expert users in their daily lives, we investigated questions posted on Yahoo! Chiebukuro (Yahoo! 知恵袋) [36], the largest Q&A site for non-experts in Japan. We chose a Japanese Q&A site because a previous survey revealed that among Arabic, French, Japanese, Chinese, Korean, and Russian participants, the Japanese non-expert users had the lowest security behavior scores [80]. A lower score indicates less secure behavior; hence, we speculate that Japanese non-expert users are likely to have a greater variety of security- and privacy-related concerns in their daily lives. To support such users effectively, it is essential to identify frequent, serious, and sensitive question topics. Given these observations, we address the following research questions in this work.

RQ1 What types of security and privacy topics do non-expert users post questions about on the Q&A site?

RQ2 Among these topics, which do they perceive as more serious or sensitive?

We analyzed 445 questions that were posted in security

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

categories or that contained security- and privacy-related words in the question texts. For RQ1, we qualitatively coded topics for each question post and identified seven themes. We found that many non-expert users posted questions to determine whether they had been victimized/abused, to learn about response strategies for errors/damages, and to understand the necessity of security and privacy technologies. We also found that some users faced privacy abuse. For RQ2, for evaluating question seriousness, we measured the averages of coder-rated seriousness and the percentage of questions with rewards. We also measured the percentage of anonymous posts for evaluating question sensitivity. We found that the average of the coder-rated seriousness of questions in “privacy abuse” and “account/device management” was significantly higher than that of other themes. We also found that those who seek answers are likely to use a strategy of either appealing linguistically or offering rewards. On the other hand, we found no statistically significant difference in question sensitivity among the question themes.

This study makes the following contributions.

- To the best of our knowledge, this is the first qualitative security and privacy study of a Q&A site for non-expert users to demonstrate that a Q&A-site analysis can provide insights into what non-expert users are struggling with when it comes to security and privacy in their daily lives. We identified frequently asked question themes (“cyberattack,” “authentication,” and “security software”) and question themes that askers perceived as more serious (“privacy abuse” and “account/device management”). We also demonstrated that some of the concerns of non-expert users have not been sufficiently investigated in previous studies.
- We assessed the effectiveness of potential indicators of question seriousness and sensitivity to help researchers better understand and prioritize the concerns of non-expert users. The results suggest that researchers should complementarily incorporate multiple indicators.
- We provide design implications for Q&A sites to help non-expert users judge what and how much information they should reveal in their questions.

2 Related Work

In this section, we present a review of the literature closely related to this study. We first discuss studies that investigated non-expert users’ advice sources for security and privacy issues, and the contents and quality of the advice. Next, we go over previous studies on HCI and security/privacy that explored the posts and users (i.e., askers and responders) of Q&A sites. Finally, we identify the gaps in the previous studies and clarify how our study addresses these gaps.

2.1 Security and Privacy Advice

Many researchers have assessed the contents and quality of security and privacy advice given by experts to non-expert users or advice available on the web [10, 35, 58, 60, 71, 73]. Redmiles et al. showed that the majority of advice on the web was at least somewhat actionable and somewhat comprehensible [71]. Mossano et al. identified various issues such as contradictory or abstract advice [58]. Redmiles et al. also investigated non-expert users’ reactions to security advice and found that they determined whether to accept digital security advice based on the trustworthiness of the advice source [70]. Fagan et al. surveyed users who followed security advice and found that they rated the benefits of following, the risks of not following, and the costs of not following higher than those who did not follow the advice [16].

Other researchers have focused on advice sources [54, 65–70] and found that these include both informal (e.g., family and friends) and formal (e.g., technical support) sources, as well as both offline and online sources. Micheli et al. [54] investigated the advice sources of young adults for digital media use in 2016 and found that 43% of participants asked questions to strangers online. They also reported that males with higher Internet skills were significantly more likely to ask questions to strangers online.

2.2 Asking Questions on Q&A Sites

Q&A sites such as Yahoo! Answers offer people the opportunity to obtain desired information rapidly and efficiently online. Thus, Q&A sites have become an interesting and promising subject of research in computer science [6, 85].

User motivations. Askers post questions for various reasons, such as to obtain specific information, to obtain non-popular information, to gather diverse opinions and experiences, and to satisfy curiosity [39, 40]. Previous studies examining the motivation of responders commonly concluded that the primary motivation was altruism (e.g., to feel like they were helping someone) [59, 81].

Question topics and types. Researchers have examined Q&A sites to clarify people’s concerns (i.e., question topics) about specific issues, such as eating disorders [8] or cancer [62]. Other researchers have classified the types of questions posted on Q&A sites [2, 13, 27, 29, 32, 81]. For example, Choi et al. categorized question types as information-, advice-, opinion-, and non-information-seeking questions and found that advice- and opinion-seeking questions were the most popular on Yahoo! Answers [13, 81]. A key finding of these studies is that the frequency of question types differs among categories and Q&A sites.

Anonymity and sensitivity of posts. One of the most unique features of Q&A sites is anonymous posts. When users create accounts, some sites (e.g., Yahoo! Answers) al-

low pseudonyms, whereas with others (e.g., Quora¹ [33]), real names are mandatory. When users post questions, both types of sites typically offer anonymity. Researchers consider anonymity to be related to the sensitivity of a post [23, 64]. Naturally, the questions that are rated highly sensitive by coders are more likely to be asked anonymously [23]. Peddinti et al. [64] identified some of the question categories for which users are more likely to answer anonymously as religion, drugs, and sexual orientation.

Askers' strategies and question answerability. Although posting questions on Q&A sites has many benefits, these sites do not always work as expected because not all questions receive answers, and the quality of received answers is not always high. Therefore, askers utilize strategies such as specifying, clarifying, and signaling to ensure a higher chance of a response [39, 40]. Many studies have examined the answerability of questions on Q&A sites [6, 14, 23, 28, 46, 82, 99]. For example, Harper et al. [28] explored the variables that affect answer outcomes (such as number, length, effort, and quality of answers) and found that question topics, question types, levels of reward, and the site itself significantly affected one or more of these outcomes. Another study showed that the topics, uniqueness, and urgency of questions significantly affected the possibility of receiving answers [46]. As for allowing anonymity, it had no significant effect on the answer quality [23].

2.3 Security and Privacy Posts by Developers

Stack Overflow [34] is unique in that its target users are developers and programmers, and it has become the most popular information source for developers [1]. Many researchers have studied question topics on Stack Overflow to clarify developers' concerns and challenges related to security and privacy [47, 63, 88, 97]. For example, Tahaei et al. performed qualitative analysis to determine what developers ask about privacy-related issues on Stack Overflow and found that they often asked questions about privacy policies, privacy concerns, access control, and version changes [88]. Patnaik et al. identified the usability issues of cryptography libraries by qualitatively reviewing the questions on Stack Overflow [63]. Yang et al. conducted a large-scale study of questions with tags related to security on Stack Overflow and found that they covered a wide range of topics mainly belonging to five categories: web security, mobile security, cryptography, software security, and system security [97]. They also revealed that questions about passwords and signatures were posted frequently, but were less likely to be answered.

¹Quora initially required users to register their real names, but it has allowed users to use pseudonyms since 2021.

2.4 Research Gaps in Previous Studies

As mentioned in Section 2.1, nearly half of young adult users ask questions regarding digital media use to strangers online. Hence, in this study, we analyzed security- and privacy-related questions posted on a Q&A site. Although many researchers in the security and privacy community have investigated questions posted on Q&A sites for developers (as mentioned in Section 2.3), little is known about the questions posted by non-expert users. To clarify the security- and privacy-related questions posted by non-experts, we generally adopted the same analysis approaches and findings as previous Q&A site studies (see Sections 2.2 and 2.3), which we explain in detail in Section 3.3.

3 Methodology

We collected and analyzed security- and privacy-related questions posted on Yahoo! Chiebukuro (Yahoo! 知恵袋) [36], a site that was chosen because of its popularity and the wealth of features available to users (e.g., rewards for best answers, anonymous posts). In this section, we first present the mechanism of posting questions and receiving answers on Yahoo! Chiebukuro and then explain our data collection and analysis method.

3.1 Descriptions of the Target Q&A Site

Yahoo! Chiebukuro (Yahoo! 知恵袋) [36], where users share their knowledge and wisdom by answering questions, is the most popular Q&A site in Japan². The meaning of the Japanese word "Chiebukuro" is "bag of knowledge." It is provided only in Japanese and is available on the web and as an app (iOS and Android). Yahoo! Chiebukuro is generic, which means the site is not dedicated to a specific demographic of people (e.g., people with specific professions), and open, which means it is not invitation-only but is available to everyone. Anyone with a Yahoo! ID can post a question and answer for free. Yahoo! does not recommend that users include their real names in their Yahoo! IDs, and users can set random or favorite strings. Thus, we consider Yahoo IDs to be pseudonyms. Yahoo! Chiebukuro has various question categories spanning entertainment, romance, health, politics, technology, and more. It received approximately 4.5 million posts per month as of March 2021 [38].

Figure 1 shows a screenshot of the interface of a question on the Yahoo! Chiebukuro website. Herein, we present the mechanism of Yahoo! Chiebukuro in accordance with the four steps of a Q&A lifecycle: 1) an asker posts a question, 2) potential responders view the question, 3) responders post

²Yahoo! Answers, which is the global version of Yahoo! Chiebukuro, was closed in May 2021. The closure did not affect Yahoo! Chiebukuro because it is run by a different operating company.



Figure 1: Interface of a question in Yahoo! Chiebukuro.

answers, and 4) the question is closed either manually by the asker or automatically by the system.

1) Posting a question. An asker inputs the question text and, if necessary, attaches an image file (e.g., screenshot). The asker then selects one or two categories either manually or from a list of automatically recommended categories based on the question text. The categories are structured in a three-tier hierarchy (e.g., *Computer technology* > *Security* > *Network security*), and the Yahoo! ID of the asker is not anonymous by default. When posting a question using the Yahoo! Chiebukuro app, askers can opt to make their Yahoo! ID anonymous for free. When posting a question via the website, they can make their Yahoo! ID anonymous by paying with ChieCoins, which are used only on Yahoo! Chiebukuro and have no real-world value. Users can receive ChieCoins from the service by performing various actions such as registering, logging in, posting a question, posting an answer, and selecting the best answer; in addition, they receive ChieCoins if their answer is selected as the best answer. An asker can offer rewards for the best answer (25, 50, 100, 250, or 500 ChieCoins) to increase the probability of receiving answers. Each question has only the question text without any title or tag.

2) Viewing a question. A potential responder finds questions by selecting a category of interest or searching for a specific word. On an index page of each category/word, a potential responder can explore the questions by status (i.e., open or resolved) and sort them by newness, number of answers received at that time, or reward amount. On the in-

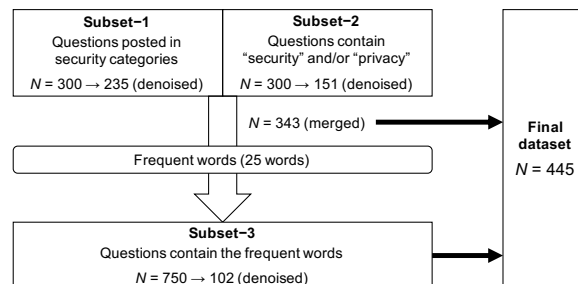


Figure 2: Data collection flow in this study.

dex page, a potential responder can see the beginning of the question text (about 40 Japanese characters), the main question category, the number of answers received at that time, an attached image (optional), any additional rewards (optional), and an anonymous-posts flag (optional) for each question.

3) Posting an answer. A responder inputs the answer text and if necessary, attaches an image file.

4) Closing a question. Each question is open for responders to answer for seven days by default. If a question does not receive any answers within this period, it is automatically deleted. If a question receives one or more answers, the asker can select the “best answer” from among them. When the asker selects the best answer, the question is marked as “resolved”, and no further answers will be accepted. A question that has received one or more answers and has been live for more than seven days is marked as “closed and waiting for the asker’s vote” until the asker selects the best answer.

3.2 Data Collection

As shown in Fig. 2, we created a dataset consisting of three subsets of questions collected in different ways: questions from security-related categories (Subset-1), questions containing the words “security” and/or “privacy” (Subset-2), and questions containing some words related to security and privacy (Subset-3). This merged dataset was created to cover a wide variety of security- and privacy-related questions.

Subset-1: Collected in security-related categories. Yahoo! Chiebukuro has three categories that are directly related to security: *Computer technology* > *Security* > *Network security*, “*Computer technology* > *Security* > *Cryptography and authentication*,” and “*Internet* > *Internet services* > *Computer virus measures and security practices*.” There are no categories that are directly related to privacy. We collected all question posts (comprising the question text, attached image, and some metadata) from these three categories. Note that we collected all posted questions regardless of whether they had received answers, even though a question with no answers is removed from the service later. We started collecting question posts in December 2021 and

continued for seven days until we had obtained 300 without random sampling. Then, two authors (security and privacy researchers) independently reviewed all the question posts to exclude any that satisfied any of the following conditions: (1) questions that were not related to computer security or privacy, (2) questions that were too vague, and (3) questions that the askers seemed to be using for an exam or homework. The discrepancies between the two coders were resolved by discussion and we finally obtained 235 question posts.

Subset-2: Collected with “security” and/or “privacy”. Although Yahoo! Chiebukuro has three categories directly related to security, askers may post security-related questions in categories besides these. For example, when an asker who wants to know about the security and privacy of smartphones posts a question, the automatic category recommendation system might recommend a category related to smartphones. Therefore, we collected question posts that contained the word “security” and/or “privacy” in the question text from all categories. We collected 300 question posts in the same way and period as Subset-1, and after performing the same exclusion, we obtained 151 question posts.

Subset-3: Collected with related words. Users might post security- and privacy-related questions that do not actually include the words “security” or “privacy,” e.g., “What does this warning mean?” with an attached image file. Therefore, we collected questions that contain specific words that appear frequently in security- and privacy-related topics in the question text from all categories. After merging Subsets-1 and -2 without overlapping ($N = 343$), we extracted frequent nouns in the question texts using MeCab [43] and mecab-ipadic-NEologd [79], which are Japanese morphological analyzers. The top 25 most frequent nouns were as follows: site, account, virus, setting, information, password, app, login, PC, software, screen, email, code, smartphone, authentication, (tele)phone, fraud, iPhone, Google, infection, connection, file, Internet, registration, and deletion. We believe these nouns are a representative, though not comprehensive, set of frequently used keywords related to the research theme of usable security and privacy [20]. We started collecting posts in all categories that included the above 25 nouns in the question text in January 2022. It took only one day to collect 30 question posts for each word (a total of 750 posts) without random sampling. After performing the same exclusion as Subsets-1 and -2, we obtained 102 question posts.

Final dataset. After merging Subsets-1, -2, and -3 without overlapping, we obtained a final dataset consisting of 445 question posts. Our sample size ($N = 445$) was sufficiently larger than that of a recent representative study ($N = 315$) in which privacy-related posts on a developer Q&A site were qualitatively reviewed [88]. In our dataset, the average text length was 168.6 Japanese characters (Med. 132), which is regarded as equivalent to 86.5 English words (Med. 67.7) [95]. Of the 445 question posts, 73 (16.4%)

had an attached image. After the period for receiving answers, 353 (79.3%) posts had received one or more answers (“resolved”: 43.1% and “closed and waiting for the asker’s vote”: 36.2%), and the remaining 92 (20.6%) posts received no answers (“deleted”).

3.3 Data Analysis

Analysis approach. To determine the question topics of non-expert users, we adopted a qualitative analysis approach (i.e., manual coding) rather than quantitative. A previous study that analyzed question posts on a Q&A site [88] demonstrated that the topic modeling yielded high-level results similar to the results of manual coding. We did not utilize topic modeling in this study because our preliminary investigation revealed that Yahoo! Chiebukuro users often post questions by attaching images instead of explaining their situation in detail using only words. In contrast to topic modeling, which lacks syntax and semantics, manual qualitative coding can provide deeper insights: for example, we can identify whether an asker was trying to preserve their privacy or abuse someone else’s privacy.

Coding procedure. Two authors (security and privacy researchers) reviewed the question texts and attached images using inductive thematic analysis [9]. For each question post, we coded the question topics (RQ1) and the askers’ perceived seriousness (RQ2). The two coders independently coded 100 randomly selected question posts and developed a codebook over the course of many discussions, which was then used to independently code all the collected question posts.

Question topics (RQ1). We represented question topics using themes and sub-themes. Following a previous study that analyzed question topics posted by developers on Stack Overflow [97], we categorized the themes in our study on the basis of security and privacy technologies and threats (e.g., theme: “authentication”). Sub-themes were categorized to describe the question topics in more detail and to cover the concepts of question types (i.e., whether the askers sought information or advice), question drivers (what prompted the askers to post questions), and phase of security and privacy practice (e.g., prevention or response). For each question post, we assigned one theme and one or two sub-themes, as askers sometimes asked two questions within the same post. For example, they might ask whether their devices have been infected, and if so, what they should do (e.g., theme: “cyber-attack,” and sub-themes: “have I been hacked?” and “how to handle this?”). Our final codebook consisted of seven themes and 19 sub-themes (excluding “other”). Of the 445 question posts, 416 were assigned one sub-theme, and the remaining 29 were assigned two sub-themes. We calculated the interrater reliability of the two coders’ theme assignment for all question posts and found that the Cohen’s Kappa coefficient

was 0.87, indicating high agreement.

Question seriousness (RQ2). According to Hsieh and Counts, a serious question can be defined as a one that you believe the question asker really wanted an answer for [32]. We adopt their definition in this study and utilize two evaluation measurements that may act complementarily.

The first measurement is the coder-rated seriousness of the question text. The coders manually reviewed the seriousness of each question text based on the above definition using a 5-point Likert scale, where 1 is not serious, 3 is moderately serious, and 5 is very serious [32]. The coders considered posts to have higher seriousness when the askers expressed certain signals such as expressions of urgency, anxiety, or a call for help. They judged based only on the question text, i.e., without looking at the metadata such as reward amount or anonymity. We provide some examples of question posts and the value of coder-rated seriousness in Table 2 of Appendix A. The correlation coefficient between the ratings of the two coders was $r = .773$, indicating adequate reliability. We calculated the average ratings of the two coders for each question topic.

The second measurement was the rewards (ChieCoins) for the best answer. Yahoo! Chiebukuro recommends that users who want to increase the probability of receiving answers should offer rewards for the best answer [37]. For each question topic, we calculated the percentage of question posts for which the askers offered rewards. Note that we did not report the average number of ChieCoins that askers offered. On Yahoo! Chiebukuro, askers must set rewards from either 25, 50, 100, 250, or 500 ChieCoins, so we cannot be certain that the level of seriousness perceived by askers exactly matches the reward amount.

The coder-rated seriousness is intended to capture the linguistic expressions of the askers, and the percentage of reward is indicative of the askers' behaviors when requesting answers. In this study, we judged a question as serious when either or both of these measurements were high.

Question sensitivity (RQ2). For measuring question sensitivity, we calculated the frequency of anonymous posts for each question topic. It is well known that anonymity can be used as a metric that captures the sensitivity of questions [64], i.e., askers tend to post sensitive questions anonymously [23].

3.4 Ethical Consideration

We followed the ethical principles laid out in the Menlo Report [5] and the ethical methods of studying online communities [77, 87, 91]. We also abided by Yahoo! Chiebukuro's Terms of Service. Our crawler sent requests with intervals of more than 15 seconds. We did not collect any personally identifiable information or the Yahoo! IDs of the askers. To investigate whether the posts were anonymous, we collected only the flag metadata that indicated whether the posts were

anonymous or non-anonymous. In this paper, we present only the aggregated data or the translated and abstracted contents of the original question posts (i.e., we avoid direct quotes) so that readers will not be able to identify the original question posts or askers. For the example shown in Fig. 1, we selected a post in which both the asker and responder were anonymous. Our study design was approved by our Institutional Review Board (IRB).

4 Results

4.1 RQ1: Question Topics

The final codebook for question topics consisted of seven themes and 19 sub-themes, excluding "Other." Table 1 presents an overview of the themes and their frequencies.

4.1.1 Cyberattack (40.7%)

The most frequent question theme was cyberattacks, which includes activities such as online fraud, phishing, malware, and account hijacking. Askers posted questions regarding the prevention of such cyberattacks, incident identification, and responses to these incidents. Note that we did not split the theme code into different attack types because in some cases, the description of the question was not clear, making it difficult to perform such classification.

Is this malicious? / Have I been hacked? (24.0%) Various triggers can make users anxious that they are facing a cyberattack. Examples of such triggers include suspicious messages (email, SMS, or popup), mistakenly accessing an unintended webpage, notifications from security software, suspicious activity logs that the user does not recognize, reduced operation speed of the device, and rapid draining of the device battery. Among cyberattacks, a frequently encountered event was one that we suspect to be a technical-support fraud: "*I received a warning that my computer has been infected with Trojan Horse and I need to call Microsoft Support Center. Is this a fraud or has my computer actually been infected?*" In some questions, the users copied and pasted the received messages into the questions and asked if these messages were fraudulent. Most of the messages received by the users were spoofed with URLs or sender email addresses using typical techniques such as typosquatting (e.g., AppleSupp0rt) or using an email address of a well-known free mail service (e.g., a message disguised as Google by using a Gmail address). As Reynolds et al. revealed, non-expert users are not even aware of the typical fraud techniques [75], so it is difficult for them to detect fraud on their own. We found that users who noticed that a site was a scam before they completely entered their personal information were worried about being victimized by attacks: "*[...] After entering my real name, I finally calmed down and closed the browser without entering my credit card*

Table 1: Results of question topics and the askers' perceived seriousness and sensitivity.

Theme	Sub-theme	Frequency*		Seriousness (RQ2)			Sensitivity (RQ2)		
				Ave. rating**	% Reward **	% Anonymous**			
Cyberattack (e.g., online fraud, phishing, malware, and account hijacking)	Is this malicious? / Have I been hacked?	40.7%	24.0%	3.6	3.8	29.3	29.9	50.5	
	How to handle this?		11.9%		4.1		41.5	50.9	
	Is there any possibility of being hacked?		4.7%		3.0		23.8	48.1	52.4
	How to prevent it?		3.4%		3.1		26.7	46.7	
	Other		2.5%		2.5		9.1	36.4	
Authentication	How to handle this?	16.2%	14.6%	3.6	3.7	29.2	26.2	36.9	
	Is it necessary/effective/trustworthy?		1.1%		–		–	36.1	–
	Other		0.4%		–		–	–	–
Security software	How to use it?	13.0%	6.7%	3.1	3.3	41.4	40.0	26.7	
	Which product do you recommend?		3.1%		2.9		35.7	28.6	
	Is it necessary/effective/trustworthy?		2.2%		3.1		40.0	31.0	40.0
	Other		0.9%		–		–	–	–
Privacy abuse (e.g., IPA, cyberstalking, parental control, and voyeurism)	How to escape from surveillance?	7.9%	2.9%	4.0	4.3	25.7	23.1	38.5	
	Am I under surveillance?		2.2%		4.0		30.0	50.0	
	Is this privacy abuse?		1.8%		–		–	42.9	–
	How to surveil a target?		0.7%		–		–	–	–
	Other		0.7%		–		–	–	
Account and device management	How to handle this?	7.0%	5.6%	3.9	4.0	45.2	44.0	48.0	
	What should I (not) do?		1.1%		–		–	48.4	–
	Other		0.2%		–		–	–	–
Secure connection (e.g., Wi-Fi and VPN)	How to use it?	6.5%	4.9%	3.2	3.4	27.6	31.8	36.4	
	Is it necessary/effective/trustworthy?		1.1%		–		–	48.3	–
	Other		0.4%		–		–	–	–
Privacy setting	How to set it?	5.6%	3.1%	3.7	3.8	36.0	35.7	42.9	
	Are my data disclosed?		2.2%		3.9		50.0	44.0	40.0
	Other		0.4%		–		–	–	–
Other			3.1%		2.9		14.3	35.7	

* For each question post, we assigned one theme and one or two sub-themes, as askers sometimes ask two questions within the same post.

** ‘–’ indicates that the sub-theme accounts for less than 2.0% of all question posts. These sub-themes are potentially influenced by an outlier.

information. Was my device already infected with a virus at the moment I accessed the URL?”

How to handle this? (11.9%) Many users seemed to have no idea what to do when they perceived that they had been subjected to a cyberattack: “When I was browsing web sites, a message saying ‘Your device is infected with 39 computer viruses’ was suddenly displayed. What should I do? I’ve never seen this message before, and I’m very worried. Please help me deal with this!” In cases where users have already undertaken the basic security measures, they may be looking for additional actions: “I accessed a phishing URL posing as Amazon and input my personal information, prepaid card number, and Amazon login information. Now I have changed my Amazon and prepaid card service passwords. Is there any other action I should take?” According to prior studies that analyzed the advice on anti-phishing and anti-account-compromise on the web, a minority of the websites provided complete advice for remediation [58, 60]. Hence, users may be unable to complete the necessary measures against online fraud.

Is there any possibility of being hacked? (4.7%) Users were worried about the types of situation in which they could be at risk of cyberattacks, as indicated by questions such as “Are smartwatches also at risk of being infected by viruses?” and “If a smartphone belonging to a member of my family

gets infected by a virus, is it possible that devices of other family members will get infected by the virus via Wi-Fi or other means?” A few users believed in unscientific conspiracy theories (e.g., the coronavirus containing malware code inside it) and were concerned about unrealistic cyberattacks (e.g., eavesdropping on thoughts).

How to prevent it? (3.4%) Some users were proactively contemplating prevention methods against cyberattacks, as indicated by the following questions: “How can I keep my computer and smartphone secure?” and “Is it better to log out every time after I use a Google account?” The prevention methods mentioned by users were not always effective or feasible: “I heard someone’s <service name> account had been hijacked on the news. To prevent account hijacking, what should I do? I have installed shopping apps on my smartphone. Is it effective to uninstall them after every time I use them?”

4.1.2 Authentication (16.2%)

Authentication is a security technique that most users encounter whenever they access services. Most of the questions in this question theme were posted when the users’ authentication had failed.

How to handle this? (14.6%) We found that many users

failed to receive security codes for multi-factor authentication because of discarded authenticator devices or fake email addresses registered for email verification: “I can’t log in to my <service’s name>’s account, where I set up a two-factor authentication with my phone number. Some days ago, I changed my phone number. How do I log in to the account again?” Some users had trouble using multi-factor authentication because of an implementation issue with the service or its app: “I confirmed the security code in the SMS app. However, when I go back to the original app, the screen for sending the security code is displayed, instead of the screen for inputting the security code. I’m stuck in this loop.” When authentication failed, some users tried to contact the service operator. However, they sometimes could not find the contact point: “I can’t log in to <service’s name>. I can’t find the contact form on the website, and the service doesn’t have a Twitter account, so I can’t contact them. [...] How do I get my account back?” Another user was irritated with a smartphone unlock issue that arose because of measures put in place during the COVID-19 pandemic: “I’ve been wearing a mask all the time due to COVID-19, and because of that, the Face ID doesn’t work. I end up having to input the passcode every time. That is inconvenient. [...] Is there any good way around that?”

Is it necessary/effective/trustworthy? (1.1%) Service providers and security researchers have stressed to users that two-factor authentication and two-step authentication are important technologies to improve the security of user accounts while maintaining their usability [15, 22, 74, 76]. Unfortunately, some users are skeptical about the necessity of these technologies: “I read reviews of a two-step authentication app and found many critical reviews. Do we really need two-step authentication?”

4.1.3 Security Software (13.0%)

Security software is often bundled with the OS or pre-installed in products, making it the most familiar security tool for most users. However, users often do not fully understand how to use it and how useful it is.

How to use it? (6.7%) Users struggle to set up security software and understand its features: “<Anti-virus software’s name>’s offline scan did not run. [...] What should I do?” and “The message says that silent mode is disabled, and the scheduled scan and detection notification are enabled. What does this mean?” We observed an unfortunate case in which the message displayed by the security software misled a user, though this issue may be peculiar to Japanese grammar. When the user saw the screen message saying that it was scanning for a Trojan Horse, they misunderstood that it had been detected on their device. Users also struggled to set up exception cases, i.e., legitimate access: “<Anti-virus software’s name> recently blocked my access to <service

name>, deeming it a suspicious site. How can I stop the blocking?”

Which product do you recommend? (3.1%) It was difficult for users to compare and choose between the technical advantages of various security products, so they sought opinions and recommendations from others: “What is the best anti-virus software? I currently use <software name>, but I frequently receive fraud emails. I plan to change to another software.” Users requested recommendations for software that has specific features and a good cost performance. Some users wondered which was better, using OS-bundled anti-virus software or purchasing their own anti-virus software.

Is it necessary/effective/trustworthy? (2.2%) Users, especially those who used their devices only for limited purposes, tended to be skeptical about the effectiveness and necessity of security software: “I use <anti-virus software’s name>, but I don’t see the benefits. When it runs in the background, my computer gets hot and the fan gets noisy. I want to uninstall it. I use this computer only for creating documents and surfing popular websites. Please tell me why I should use anti-virus software on my computer.”

4.1.4 Privacy Abuse (7.9%)

Privacy researchers have been worried about the prevalence of privacy abuse issues such as intimate partner abuse (IPA) [12, 17, 53, 92, 100], cyberstalking [41, 90], excessive parental control [21, 83, 96], voyeurism [51, 84], and bugging [51, 84]. In previous studies, privacy abuse has been researched in cooperation with professional organizations by means of closed questionnaires and interviews [17, 92, 100]. Surprisingly, we found a non-negligible number of questions on privacy abuse posted on the open Q&A site. We found questions from both the attackers’ and the victims’ points of view.

How to escape from surveillance? (2.9%) Users sought ways to escape surveillance by their partners (or ex-partners), friends, parents, acquaintances, schools, and companies. Users asked about various kinds of surveillance:

“When I was married to my ex-husband, I logged into my Instagram account from his Facebook account once. Since then, he seems to be logging into my Instagram account using his Facebook account. I find this very unpleasant, but I don’t know his Facebook password. Please tell me how to remove his surveillance.”

“My friend snooped on my smartphone and tried to use it. It has private chat logs and apps containing info on my sexual habits, so I don’t want it to be peeked at. [...]”

“I’m a student. My device is restricted by <security software’s name> that my parents set. Is there any way I can unlock it without using my parents’ devices? [...]”

Am I under surveillance? (2.2%) We found that some users, presumably children, wanted to know if they were being monitored by parental control features: *“I heard that parents can see children’s (browsing) histories with <security software’s name>. I remembered that the app had been pre-installed on my smartphone and I checked it. Then it asked me to agree to the privacy policy. It isn’t working, is it? My parents haven’t seen my history, have they?”* Another user was worried about voyeurism and bugging at the place they were staying: *“I hear something strange from the digital speakers on the ceiling of my hotel room. Is it possible that I’m being a target of voyeurism or being bugged?”*

Is this privacy abuse? (1.8%) Users asked for objective opinions on whether a certain action by themselves or another person constituted a privacy violation. *“My company asks me to submit a QR code for my private ID of <messaging service’s name>. This is a privacy violating action, isn’t it?”* and *“Please give me your opinion on children’s privacy and rights with conducting parental control. In the case of teenage children, to what extent do you think parents should intervene in their children’s smartphones? Specifically, please tell me about each of the following behaviors: keeping an eye on their location with a GPS, limiting the web sites they can visit, viewing their contact information, viewing their browsing histories, viewing incoming and outgoing call histories, and viewing their emails and chats.”*

How to surveil a target? (0.7%) Users were curious about the extent to which they could monitor a target using spyware apps: *“I want to know about the features of spyware apps, especially <spyware app’s name>. Is it possible to track targets even when they have turned off the GPS on their smartphone? How about when they have switched their smartphone to airplane mode?”* However, not all question posts were necessarily asked by malicious users. One user needed advice on monitoring their children to prevent them from being involved in a crime: *“[...] I found that my daughter created <SNS names> accounts. On her Twitter profile, she wrote messages asking to go on dates with adult males. I explained the various risks to her, and she agreed and deleted her accounts. However, today, I found that she received an email saying that her <SNS name> account had been restored. As a countermeasure, I set up her Gmail account so that I can view her emails. Should I take further countermeasures?”*

4.1.5 Account and Device Management (7.0%)

Questions in this theme deal with security- and privacy-related issues of account and device management, especially those related to setting up new accounts/devices and disposing of old ones.

How to handle this? (5.6%) Users asked for the appropriate account deletion procedure to protect their privacy: *“I want*

to delete my <service’s name> account. But I couldn’t find the delete option on my profile page. Can someone please tell me how to delete my account?” Previous studies on the presence of account deletion options on websites reported that not all websites provided such options [25, 31], which can cause confusion to the users.

What should I (not) do? (1.1%) A small number of users sought general advice on what to do with the apps and local data on their old devices when buying new ones. They also asked about the potential risks of simply discarding their old devices. As Ceci et al. reported, non-expert users are concerned about safe ways to dispose of their devices but seem to lack sufficient knowledge about how to do so [11].

4.1.6 Secure Connection (6.5%)

We found that a certain number of users tried to establish a secure connection encrypted by one or more security protocols. Most of the questions in this theme were about Wi-Fi and virtual private networks (VPNs).

How to use it? (4.9%) Users were confused by the many technical terms and names of security standards and encryption methods that appear on Wi-Fi connection setting screens. *“Which Wi-Fi security mode should I choose among WEP, WPA, WPA2, PSK, and 802.1X/EAP?”* Users also expressed confusion about frequently getting warning messages when they tried to connect to Wi-Fi networks: *“When I tried to connect to Wi-Fi using the IEEE802.11b standard, my iPhone screen showed that it was a legacy access point. Does this mean that there is a security problem?”* and *“When I use Wi-Fi on my iPhone, I get a ‘Privacy Warning’ message. Does this happen often? How do you deal with it?”*

Is it necessary/effective/trustworthy? (1.1%) Users seemed interested in the necessity, effectiveness, and trustworthiness of VPNs: *“I was recommended to use a VPN app as a trick to access a web site that my device can’t access. Are VPN apps secure?”* and *“Is VPN effective in making public Wi-Fi secure?”*

4.1.7 Privacy Setting (5.6%)

Application or website privacy settings can allow users to control their privacy. However, it is sometimes difficult for users to understand these settings and configure them appropriately.

How to set it? (3.1%) With regard to cookies, there have been numerous discussions about how service providers present users with cookie notifications (e.g., option, framing, and display position designs, as well as default) [50, 61, 93]. We observed that users suffered from different usability issues regarding cookies: *“When I visited the <service name>’s website, it asked me whether I would allow cookies. I mistakenly hit the allow button. Is it possible to change*

it to deny permission?” Another user had difficulty understanding the meaning and mechanism of personalization on the privacy setting page: “What does ‘Personalization based on your inferred identity’ on Twitter’s privacy setting page mean?”

Are my data disclosed? (2.2%) Users expressed concern about whether their data were disclosed or shared, especially because of unintended privacy settings: “*I browsed a certain company’s websites via Safari with my iPhone’s location information turned on. In this case, is my location information disclosed to the company? Is there a difference between using Wi-Fi at home and on a mobile line?*”

4.2 RQ2: Seriousness and Sensitivity

We examined relatively serious and sensitive question themes to better understand non-experts’ expectations and prioritize the themes accordingly. Note that every question theme is already regarded as at least some level of seriousness at the point of posting a question on a Q&A site.

4.2.1 Question Seriousness

For all the collected question posts, the average coder-rated seriousness was 3.5, and 31.5% (140/445) of the questions were posted with rewards for the best answer. The averages of the coder-rated seriousness and the percentage of question posts for which the askers offered rewards are listed in Table 1. We performed an unpaired *t*-test to compare the coder-rated seriousness score between the question posts of askers who offered rewards and those who did not. Although we found no significant difference ($p = .054$), those who offered rewards seem to express slightly more serious signals in their questions (avg. seriousness = 3.7) than those who did not offer rewards (avg. seriousness = 3.5). This indicates that askers who seek answers are likely to use a strategy of either appealing linguistically or offering rewards.

The average coder-rated seriousness was higher for questions under the themes of “privacy abuse” and “account/device management.” We observed that askers frequently expressed their anxiety in question posts under these themes. We performed a Kruskal-Wallis test to compare the coder-rated seriousness across the question themes and found that there was a statistically significant difference ($p < .001$). We then performed post hoc Wilcoxon rank-sum tests in which the *p*-values were adjusted using the Bonferroni method. We found that the average of coder-rated seriousness in “privacy abuse” and “account/device management” was significantly higher than in “security software” and “secure connection” at the 5% level. At the sub-theme level, the coder-rated seriousness of “*how to*”-type questions was higher than that of other types.

The percentage of question posts in which askers offered rewards was relatively higher under the themes of “ac-

count/device management” and “security software.” We performed a Fisher’s exact test to compare the percentage of question posts with rewards across the question themes and found that while the percentage varied moderately across themes, there was no significant difference ($p = .286$). The lower number of question posts in some question themes may have resulted in a lack of statistical power.

4.2.2 Question Sensitivity

The percentage of anonymous posts among all questions was 42.9% (191/445). While this percentage was relatively higher in “account and device management,” “secure connection,” and “cyberattack,” the Fisher’s exact test revealed no significant differences in themes ($p = .361$). As with the test for rewards, the lower number of question posts in some question themes may have resulted in a lack of statistical power. Researchers have treated “privacy abuse” as a highly sensitive topic, but we found that the percentage of anonymous posts in “privacy abuse” was not much higher than that in other themes. Users perceive the incident identification and responses to “cyberattack” as equally or more sensitive than “privacy abuse” because the incidents may expose their personal and sensitive information more broadly. It is also possible that Yahoo! Chiebukuro’s pseudonym-registration policy has an effect here, as users can keep their user IDs pseudonymized even if they do not use the anonymous post feature.

5 Discussion

In this section, we discuss the design implications for Q&A sites for non-expert users, how to leverage the Q&A-site analysis to facilitate usable security research, and the limitations.

5.1 Design Implications

We demonstrated that non-expert users post a variety of security- and privacy-related questions on Yahoo! Chiebukuro, which is a general purpose Q&A site. We believe that general Q&A sites should help non-expert users find a solution to their security- and privacy-related concerns by adopting an approach that both “pulls in” professionals and “hands off” to professionals. However, general Q&A sites may have little business motivation to provide such a support mechanism only for a specific category (including security and privacy) of questions. Having subsidies for such services provided by public agencies could be an effective solution. The call for such subsidies would not be limited to security- and privacy-related questions but would extend to various categories of serious questions that require immediate attention, such as urgent medical conditions, severe violence, and life-threatening disasters. Our specific sugges-

tions regarding cyberattacks and user privacy problems are detailed below.

Supporting users coping with cyberattacks. The most frequent theme of questions was “cyberattack.” Many non-expert users experienced issues related to incident identification and response. Non-expert users are vulnerable to attack techniques [3, 75]. Web-based knowledge related to basic attack tactics, symptoms, and advice can be utilized to create quick answers. Because of the low quality of anti-phishing advice on most websites (e.g., contradictory or abstract advice, and lack of suitable guidance) [58], the challenge is to create a usable knowledge base made up of consistent, specific, and actionable advice. We also need to understand that it is not always easy for users to find accurate information because many of the threats target users who are anxious and vulnerable. For example, technical-support scams use false alerts [55], and fake-removal-advertisement sites exploit malware-infected users’ solution search behavior [42]. Therefore, Q&A sites should collaborate with a knowledge base operated by a trusted organization to present users with appropriate information. Further, non-expert users often have difficulty explaining their issues. In our dataset, 16.4% of the question posts had an image file attached, and among them, screenshots were attached without detailed explanation. For Q&A sites to obtain appropriate information from the aforementioned type of knowledge base, first, it is necessary to obtain accurate information about the users’ issues. A possible application to support the use of information in the knowledge base is a security version of an “expert system,” which asks users for more information that is missing from their question posts and then presents a relevant solution from the knowledge base.

Helping users facing sensitive privacy problems. Users who asked questions as victims of privacy abuse require careful social support because their own privacy has been or could be severely compromised. Although anonymous online spaces provide a supportive environment for discussing potentially stigmatized sensitive topics [49], such spaces are usually created for communities facing similar issues [4]. Users may hesitate to ask questions about their privacy issues on an open and generic Q&A site, as they may become targets of slander. More than half of the questions about privacy abuse stemmed from the need to properly understand whether or not they were under surveillance or had been abused. To get answers to such questions, users have to reveal a certain amount of private information. However, non-expert users may find it difficult to judge what and how much information they should reveal.

Chatbots could be a useful tool for addressing the users’ risks of revealing private information on a public platform, as people tend to disclose their stigmatized experiences (e.g., experiences of failure or abuse, symptoms of depression) more actively to virtual agents than to humans [45, 48]. As

with security- and privacy-related questions, users may disclose sensitive content (e.g., privacy abuse) to a chatbot because they do not have to worry about slander or their private information spreading. Additionally, using chatbots allows users to exchange messages interactively and incrementally, which means users only need to disclose a sufficient and necessary amount of information for receiving their answers. In answering users’ questions, the chatbots themselves can respond in accordance with the aforementioned knowledge base. However, as pointed out by Zou et al. [100], security issues surrounding sensitive topics are complex, and there may be a variety of unsurfaced issues lurking. Therefore, it is also important to provide users with a feature that refers them to professionals for further advice [44, 100].

5.2 Exploring New Research Topics

As previously reported [47, 63, 88, 97], analyzing questions on a Q&A site for expert users (e.g., Stack Overflow) has helped researchers to better understand the security and privacy concerns of developers and programmers when developing systems. In this study, we confirmed that analyzing questions on a Q&A site for non-expert users can also allow researchers to understand the security and privacy concerns that such users are facing daily. Furthermore, our analysis of question seriousness suggests that askers who seek answers are likely to use a strategy of either appealing linguistically or offering rewards. This observation implies that researchers who analyze Q&A sites should complementarily incorporate multiple indicators to understand and prioritize the concerns of non-expert users.

Security and privacy concerns change over time as technology and lifestyles change. For example, in our dataset, the problems caused by lifestyle changes owing to COVID-19 include the inability to use Face ID, fear created by conspiracy theories, and issues with VPN settings stemming from the increase in remote work. As an efficient way to explore the usable security and privacy topics for non-expert users that have not yet been addressed, the research community should cultivate a research ecosystem that regularly extracts and clarifies the current user concerns from Q&A sites and works to resolve them. Among the questions obtained from our dataset, we highlight some usable security topics that need to be studied in more depth.

Support for authentication and account management. The second most common theme was “authentication.” We found that many users had difficulty receiving security codes for multi-factor authentication because of discarded authenticator devices or having registered with fake email addresses, not just users who failed to log in because of forgetting their credentials. In addition to an in-depth analysis of the reasons users forget to manage their credentials, researchers should further look for secure and usable ways of implementing account recovery. For example, researchers

should investigate whether services (especially non-Western services) have provided their contact points and appropriate support for users who encounter authentication errors. Some users were concerned about security and privacy in account management because they did not know how to properly create, delete, and/or link accounts. Future studies should thus cover a greater number of specific situations and diverse users.

Usability issues of security software. Usable security researchers have worked diligently on the various usability issues facing security technologies. However, we showed that many users still do not have a sufficient understanding of information about security technologies, how they work, and the merits of adopting them (see the “Security Software,” “Secure Connection,” and “Privacy Setting” parts in Section 4.1). While some security technologies (e.g., private browsing, Tor, ad-blockers, and firewalls) have been analyzed with respect to user perceptions [18, 24, 52, 86, 94], we believe that usability issues of security software such as anti-virus software and security terminology need to be studied more. In one unique approach, Zhang-Kennedy et al. succeeded in persuading users to update antivirus software by utilizing comic materials [98]. It will be necessary to investigate the usability of the features implemented in actual security software and that of the wording used in them. It will also be important to more extensively explore the user mental models about the effectiveness of the features.

5.3 Limitations

Our study has several limitations, most of which are common to similar types of research.

The first is the demographic bias among the users of Q&A sites. In general, the demographics depend on the type of service. One study that explored the demographics of active askers on Yahoo! Answers indicated that the user group was younger than the average population of web search users [19]. According to another study that explored the advice sources of young adults for digital media use, males with higher Internet skills were significantly more likely to ask questions to strangers on online [54]. Unlike Stack Overflow, which targets expert users (developers and programmers), Yahoo! Chiebukuro targets a wide range of users and is likely to attract many who are not familiar with information technology. Although Yahoo! Chiebukuro has not officially released the statistics of its active users, such demographic biases may also exist in our dataset to some extent.

The second limitation is that we analyzed only a Q&A site provided for a particular language. This means that the only people who ask questions are those who can use the language that the Q&A site supports. For example, we investigated Yahoo! Chiebukuro in this study, which only supports Japanese, and we acknowledge that non-expert users

from Japan may have different security and privacy attitudes compared to those from other countries due to differences in cultural factors or security and privacy literacy levels [26, 30, 56, 57, 80]. However, we believe that our findings identify the potential issues that researchers from other countries also need to resolve because most of the security and privacy technologies and concepts mentioned in our dataset are common to users worldwide.

The third limitation is the lack of profile analysis of the askers. We decided not to conduct such analysis (e.g., exploring the relationships between askers’ demographics and question topics) because we found in our preliminary investigation that a non-negligible number of users posted questions anonymously and did not publish their age and gender on their profile pages.

Fourth, our metric for question sensitivity (i.e., anonymous posts) may not exactly match askers’ perceived sensitivity, although it is a commonly used metric in the literature [23, 64]. Askers tend to post sensitive questions anonymously [23], but not every anonymous post is sensitive; i.e., there may be other reasons askers choose to post anonymously.

Lastly, because of the short sampling period (seven days), we do not claim the generalizability of our results. Instead, as we mentioned in Section 5.2, we recommend that the research community establish a research ecosystem that regularly extracts and clarifies the current user concerns from Q&A sites. We have contributed to this endeavor by demonstrating that analyzing Q&A sites for non-expert users can be a useful method for identifying their concerns at any given time.

6 Conclusion and Future Work

Research methodology to understand the concerns of non-expert users related to security and privacy in daily life is becoming increasingly important, as such concerns change over time with the evolution of technology and changes in lifestyles. We conducted an analysis of questions posts on a Q&A site for non-expert users and successfully identified their main concerns about security and privacy. Many users experienced issues related to incident identification and response, appropriate measures after being attacked, and usability of security software. Our analysis of question seriousness suggests that there is a strong demand for answers, especially for questions about privacy abuse and account/device management.

Future work should assess the answers given for the security- and privacy-related questions. We are interested in whether the askers received high-quality answers (i.e., comprehensive, actionable, and effective advice [71]) and whether they were satisfied. In future work, we aim to obtain a deeper understanding of askers and responders so as to design better social support for security and privacy.

References

- [1] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *Proceedings of the 37th IEEE Symposium on Security and Privacy, S&P'16*, 2016.
- [2] Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. Knowledge sharing and Yahoo Answers: everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web, WWW'08*, 2008.
- [3] Sara Albakry, Kami Vaniea, and Maria K Wolters. What is this URL's destination? empirical evaluation of users' URL reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI'20*, 2020.
- [4] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. Self-declared throwaway accounts on reddit: How platform affordances and shared norms enable parenting disclosure and support. In *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing, CSCW'19*, 2019.
- [5] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE Security & Privacy*, 10(2):71–75, 2012.
- [6] Antoaneta Baltadzhieva and Grzegorz Chrupała. Question quality in community question answering forums: A survey. *ACM SIGKDD Explorations Newsletter*, 17(1):8–13, 2015.
- [7] Bram Bonné, Sai Teja Peddinti, Igor Bilogrevic, and Nina Taft. Exploring decision making with {Android's} runtime permission dialogs using in-context surveys. In *Proceedings of the 13th Symposium on Usable Privacy and Security, SOUPS'17*, 2017.
- [8] Leanne Bowler, Jung Sun Oh, Daqing He, Eleanor Mattern, and Wei Jeng. Eating disorder questions in Yahoo! Answers: Information, conversation, or reflection? In *Proceedings of the 75th Association for Information Science and Technology Annual Meeting, ASIST'12*, 2012.
- [9] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- [10] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No one can hack my mind Revisiting a study on expert and non-expert security practices and advice. In *Proceedings of the 15th Symposium on Usable Privacy and Security, SOUPS'19*, 2019.
- [11] Jason Ceci, Hassan Khan, Urs Hengartner, and Daniel Vogel. Concerned but ineffective: User perceptions, methods, and challenges when sanitizing old devices for disposal. In *Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS'21*, 2021.
- [12] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The spyware used in intimate partner violence. In *Proceedings of the 39th IEEE Symposium on Security and Privacy, S&P'18*, 2018.
- [13] Erik Choi, Vanessa Kitzie, and Chirag Shah. Developing a typology of online Q&A models and recommending the right model for each question type. In *Proceedings of the 75th Association for Information Science and Technology Annual Meeting, ASIST'12*, 2012.
- [14] Erik Choi, Vanessa Kitzie, and Chirag Shah. A machine learning-based approach to predicting success of questions on social question-answering. In *Proceedings of the 2013 iConference, iConference'13*, 2013.
- [15] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "It's not actually that horrible": Exploring adoption of two-factor authentication at a university. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI'18*, 2018.
- [16] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Proceedings of the 12th Symposium on Usable Privacy and Security, SOUPS'16*, 2016.
- [17] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI'18*, 2018.
- [18] Kevin Gallagher, Sameer Patil, and Nasir Memon. New me: Understanding expert and non-expert perceptions and usage of the tor anonymity network. In *Proceedings of the 13th Symposium on Usable Privacy and Security, SOUPS'17*, 2017.

- [19] Giovanni Gardelli and Ingmar Weber. Why do you ask this? Using toolbar data to identify common patterns of Q&A users. In *Proceedings of the 21st International Conference on World Wide Web, WWW'12*, 2012.
- [20] Simson Garfinkel and Heather Richter Lipford. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust*, 5(2):1–124, 2014.
- [21] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J LaViola Jr, and Pamela J Wisniewski. Safety vs. surveillance: What children have to say about mobile apps for parental control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI'18*, 2018.
- [22] Maximilian Golla, Grant Ho, Marika Lohmus, Monica Pulluri, and Elissa M Redmiles. Driving 2FA adoption at scale: Optimizing two-factor authentication notification design patterns. In *Proceedings of the 30th USENIX Security Symposium, SEC'21*, 2021.
- [23] Cheng Guo and Kelly Caine. Anonymity, user engagement, quality, and trolling on Q&A sites. In *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing, CSCW'21*, 2021.
- [24] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. Away from prying eyes: Analyzing usage and understanding of private browsing. In *Proceedings of the 14th Symposium on Usable Privacy and Security, SOUPS'18*, 2018.
- [25] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An empirical analysis of data deletion and opt-out choices on 150 websites. In *Proceedings of the 15th Symposium on Usable Privacy and Security, SOUPS'19*, 2019.
- [26] Marian Harbach, Alexander De Luca, Nathan Malkin, and Serge Egelman. Keep on lockin' in the free world: A multi-national comparison of smartphone locking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI'16*, 2016.
- [27] F Maxwell Harper, Daniel Moy, and Joseph A Konstan. Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI'09*, 2009.
- [28] F Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A Konstan. Predictors of answer quality in online Q&A sites. In *Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems, CHI'08*, 2008.
- [29] F Maxwell Harper, Joseph Weinberg, John Logie, and Joseph A Konstan. Question types in social Q&A sites. *First Monday*, 15(7), 2010.
- [30] Ayako A Hasegawa, Naomi Yamashita, Mitsuaki Akiyama, and Tatsuya Mori. Why they ignore english emails: The challenges of non-native speakers in identifying phishing emails. In *Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS'21*, 2021.
- [31] Ayako Akiyama Hasegawa, Takuya Watanabe, Eitaro Shioji, and Mitsuaki Akiyama. I know what you did last login: Inconsistent messages tell existence of a target's account to insiders. In *Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC'19*, 2019.
- [32] Gary Hsieh and Scott Counts. mimir: A market-based real-time question and answer service. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI'09*, 2009.
- [33] Quora Inc. Quora. <https://www.quora.com>, 2022 (accessed January 14, 2022).
- [34] Stack Exchange Inc. Stack Overflow. <https://stackoverflow.com/>, 2022 (accessed January 14, 2022).
- [35] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...no one can hack my mind": Comparing expert and non-expert security practices. In *Proceedings of the 11th Symposium on Usable Privacy and Security, SOUPS'15*, 2015.
- [36] Yahoo! Japan. Yahoo! Chiebukuro (Yahoo! 知恵袋). <https://chiebukuro.yahoo.co.jp/>, 2021 (accessed December 28, 2021).
- [37] Yahoo! Japan. Yahoo! Chiebukuro Help center. <https://support.yahoo-net.jp/PccChiebukuro/s/article/H000008128>, 2022 (accessed February 12, 2022).
- [38] Yahoo! Japan. Transparency report. <https://about.yahoo.co.jp/common/transparencyreport/>, 2022 (accessed January 11, 2022).

- [39] Grace YoungJoo Jeon and Soo Young Rieh. The value of social search: Seeking collective personal experience in social Q&A. In *Proceedings of the 76th Association for Information Science and Technology Annual Meeting*, ASIST'13, 2013.
- [40] Grace YoungJoo Jeon and Soo Young Rieh. Social search behavior in a social Q&A service: Goals, strategies, and outcomes. In *Proceedings of the 78th Association for Information Science and Technology Annual Meeting*, ASIST'15, 2015.
- [41] Puneet Kaur, Amandeep Dhir, Anushree Tandon, Ebtesam A Alzeiby, and Abeer Ahmed Abohassan. A systematic literature review on cyberstalking. An analysis of past achievements and future promises. *Technological Forecasting and Social Change*, 163, 2021.
- [42] Takashi Koide, Daiki Chiba, Mitsuaki Akiyama, Katsunari Yoshioka, and Tsutomu Matsumoto. It never rains but it pours: Analyzing and detecting fake removal information advertisement sites. In *Proceedings of the 17th Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, DIMVA'20, 2020.
- [43] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. <http://taku910.github.io/mecab>, 2022 (accessed January 20, 2022).
- [44] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. Exploring the effects of incorporating human experts to deliver journaling guidance through a chatbot. In *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW'21, 2021.
- [45] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. "I hear you, I feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.
- [46] Zhe Liu and Bernard J Jansen. Questioner or question: Predicting the response rate in social question and answering on Sina Weibo. *Information Processing & Management*, 54(2):159–174, 2018.
- [47] Tamara Lopez, Thein Tun, Arosha Bandara, Levine Mark, Bashar Nuseibeh, and Helen Sharp. An anatomy of security conversations in Stack Overflow. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Society*, ICSE-SEIS'19, 2019.
- [48] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.
- [49] Xiao Ma, Jeff Hancock, and Mor Naaman. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, CHI'16, 2016.
- [50] Dominique Machuletz and Rainer Böhme. Multiple purposes, multiple problems: A user study of consent dialogs after GDPR. In *Proceedings of the 20th Privacy Enhancing Technologies Symposium*, PETS'20, 2020.
- [51] Shirang Mare, Franziska Roesner, and Tadayoshi Kohno. Smart devices in Airbnbs: Considering privacy and security for both guests and hosts. In *Proceedings on the 20th Privacy Enhancing Technologies Symposium*, PETS'20, 2020.
- [52] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. Characterizing the use of browser-based blocking extensions to prevent online tracking. In *Proceedings of the 14th Symposium on Usable Privacy and Security*, SOUPS'18, 2018.
- [53] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI'17, 2017.
- [54] Marina Micheli, Elissa M Redmiles, and Eszter Hargittai. Help wanted: Young adults' sources of support for questions about digital media. *Information, Communication & Society*, 23(11):1655–1672, 2020.
- [55] Najmeh Miramirkhani, Oleksii Starov, and Nick Niki-forakis. Dial one for scam: A large-scale analysis of technical support scams. In *Proceedings of the 24th Annual Network and Distributed System Security Symposium*, NDSS'17, 2017.
- [56] Masahiko Mizutani, James Dorsey, and James H Moor. The internet and japanese conception of privacy. *Ethics and Information Technology*, 6(2):121–128, 2004.
- [57] Keika Mori, Takuya Watanabe, Yunao Zhou, Ayako Akiyama Hasegawa, Mitsuaki Akiyama, and Tatsuya Mori. Comparative analysis of three language

- spheres: Are linguistic and cultural differences reflected in password selection habits? *IEICE Transactions on Information and Systems*, 103(7):1541–1555, 2020.
- [58] Mattia Mossano, Kami Vaniea, Lukas Aldag, Reyhan Düzgün, Peter Mayer, and Melanie Volkamer. Analysis of publicly available anti-phishing webpages: Contradicting information, lack of concrete advice and very narrow attack vector. In *Proceedings of the 5th European Workshop on Usable Security*, EuroUSEC’20, 2020.
- [59] Kevin Kyung Nam, Mark S Ackerman, and Lada A Adamic. Questions in, knowledge in? A study of Naver’s question answering community. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, CHI’09, 2009.
- [60] Lorenzo Neil, Elijah Bouma-Sims, Evan Lafontaine, Yasemin Acar, and Bradley Reaves. Investigating web service account remediation advice. In *Proceedings of the 17th Symposium on Usable Privacy and Security*, SOUPS’21, 2021.
- [61] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI’20, 2020.
- [62] Sanghee Oh, Yan Zhang, and Min Sook Park. Cancer information seeking in social question and answer services: Identifying health-related topics in cancer questions on Yahoo! Answers. *Information Research*, 21(3), 2016.
- [63] Nikhil Patnaik, Joseph Hallett, and Awais Rashid. Usability smells: An analysis of developers’ struggle with crypto libraries. In *Proceedings of the 15th Symposium on Usable Privacy and Security*, SOUPS’19, 2019.
- [64] Sai Teja Peddinti, Aleksandra Korolova, Elie Bursztein, and Geetanjali Sampemane. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *Proceedings of the 35th IEEE Symposium on Security and Privacy*, S&P’14, 2014.
- [65] Emilee Rader and Rick Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, 2015.
- [66] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Proceedings of the 8th Symposium on Usable Privacy and Security*, SOUPS’12, 2012.
- [67] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How I learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 23rd ACM Conference on Computer and Communications Security*, CCS’16, 2016.
- [68] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. Where is the digital divide? A survey of security, privacy, and socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI’17, 2017.
- [69] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How well do my results generalize? Comparing security and privacy survey results from mturk, web, and telephone samples. In *Proceedings of the 40th IEEE Symposium on Security and Privacy*, S&P’19, 2019.
- [70] Elissa M. Redmiles, Amelia Malone, and Michelle L. Mazurek. I think they’re trying to tell me something: Advice sources and selection for digital security. In *Proceedings of the 37th IEEE Symposium on Security and Privacy*, S&P’16, 2016.
- [71] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *Proceedings of the 29th USENIX Security Symposium*, SEC’20, 2020.
- [72] Robert W Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, CHI’18, 2018.
- [73] Robert W Reeder, Iulia Ion, and Sunny Consolvo. 152 simple steps to stay safe online: Security advice for non-tech-savvy users. *IEEE Security & Privacy*, 15(5):55–64, 2017.
- [74] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A usability study of five two-factor authentication methods. In *Proceedings of the 15th Symposium on Usable Privacy and Security*, SOUPS’19, 2019.
- [75] Joshua Reynolds, Deepak Kumar, Zane Ma, Rohan Subramanian, Meishan Wu, Martin Shelton, Joshua Mason, Emily Stark, and Michael Bailey. Measuring

- identity confusion with uniform resource locators. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.
- [76] Joshua Reynolds, Nikita Samarina, Joseph Barnes, Taylor Judd, Joshua Mason, Michael Bailey, and Serge Egelman. Empirical measurement of systemic 2FA usability. In *Proceedings of the 29th USENIX Security Symposium*, SEC'20, 2020.
- [77] Lynne D Roberts. Ethical issues in conducting qualitative research in online communities. *Qualitative Research in Psychology*, 12(3):314–325, 2015.
- [78] M Angela Sasse and Ivan Flechais. *Usable security: Why do we need it? How do we get it?* O'Reilly, 2005.
- [79] Toshinori Sato. mecab-ipadic-neologd : Neologism dictionary for mecab. <https://github.com/neologd/mecab-ipadic-neologd>, 2022 (accessed January 20, 2022).
- [80] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI'17, 2017.
- [81] Chirag Shah, Vanessa Kitzie, and Erik Choi. Modalities, motivations, and materials—investigating traditional and social online Q&A services. *Journal of Information Science*, 40(5):669–687, 2014.
- [82] Chirag Shah, Marie L Radford, Lynn Silipigni Conaway, Erik Choi, and Vanessa Kitzie. “how much change do you get from 40\$?” Analyzing and addressing failed questions on social Q&A. In *Proceedings of the 75th Association for Information Science and Technology Annual Meeting*, ASIST'12, 2012.
- [83] Wonsun Shin and Hyunjin Kang. Adolescents' privacy concerns and information disclosure online: The role of parents and the internet. *Computers in Human Behavior*, 54:114–123, 2016.
- [84] Yunpeng Song, Yun Huang, Zhongmin Cai, and Jason I Hong. I'm all eyes and ears: Exploring effective locators for privacy awareness in iot scenarios. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.
- [85] Ivan Srba and Maria Bielikova. A comprehensive survey and classification of approaches for community question answering. *ACM Transactions on the Web*, 10(3):1–63, 2016.
- [86] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. Awareness, adoption, and misconceptions of web privacy tools. In *Proceedings of the 21st Privacy Enhancing Technologies Symposium*, PETS'21, 2021.
- [87] Lisa Sugiura, Rosemary Wiles, and Catherine Pope. Ethical challenges in online research: Public/private perceptions. *Research Ethics*, 13(3-4):184–199, 2017.
- [88] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. Understanding privacy-related questions on Stack Overflow. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.
- [89] Jenny Tanga, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. Defining privacy: How users interpret technical terms in privacy policies. In *Proceedings of the 21st Privacy Enhancing Technologies Symposium*, PETS'21, 2021.
- [90] Robert S Tokunaga and Krystyna S Aune. Cyber-defense: A taxonomy of tactics for managing cyberstalking. *Journal of interpersonal violence*, 32(10):1451–1475, 2017.
- [91] Leanne Townsend and Claire Wallace. Social media research: A guide to ethics. *University of Aberdeen*, 1:16, 2016.
- [92] Emily Tseng, Diana Freed, Kristen Engel, Thomas Ristenpart, and Nicola Dell. A digital safety dilemma: Analysis of computer-mediated computer security interventions for intimate partner violence during covid-19. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI'21, 2021.
- [93] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un) informed consent: Studying GDPR consent notices in the field. In *Proceedings of the 26th ACM Conference on Computer and Communications Security*, CCS'19, 2019.
- [94] Artem Voronkov, Leonardo Horn Iwaya, Leonardo A Martucci, and Stefan Lindskog. Systematic literature review on usability of firewall configuration. *ACM Computing Surveys*, 50(6):1–35, 2017.
- [95] Charles G. Wilt. Japanese-English translation. <http://cw-translation.net/e/japanese-english-translation-industry-rates-comparison.html>, 2022 (accessed February 12, 2022).
- [96] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M Carroll. Parental control vs. teen self-regulation: Is there a middle ground

for mobile online safety? In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing, CSCW'17*, 2017.

- [97] Xin-Li Yang, David Lo, Xin Xia, Zhi-Yuan Wan, and Jian-Ling Sun. What security questions do developers ask? A large-scale study of Stack Overflow posts. *Computer Science and Technology*, 31(5):910–924, 2016.
- [98] Leah Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. Stop clicking on “update later”: Persuading users they need up-to-date antivirus protection. In *Proceedings of the 9th International Conference on Persuasive Technology, PERSUASIVE'14*, 2014.
- [99] Yiming Zhao, Linrong Wu, Jin Zhang, and Taowen Le. How question characteristics impact answer outcomes on social question-and-answer websites. *Journal of Global Information Management*, 29(6):1–21, 2021.
- [100] Yixin Zou, Allison McDonald, Julia Narakornpichit, Nicola Dell, Thomas Ristenpart, Kevin Roundy, Florian Schaub, and Acar Tamersoy. The role of computer security customer support in helping survivors of intimate partner violence. In *Proceedings of the 30th USENIX Security Symposium, SEC'21*, 2021.

Appendix

A Examples of Coder-Rated Seriousness

Table 2 shows some examples of question posts and the value of coder-rated seriousness reviewed by two coders. These

coders manually reviewed the seriousness of each question text using a 5-point Likert scale (1 is not serious; 3 is moderately serious; 5 is very serious), where a serious question can be defined as one that you believe the question asker really wanted an answer for [32].

Table 2: Examples of question posts and the value of coder-rated seriousness.

Question Texts	Ave.
When I was looking at an adult site, I mistakenly called the number. <i>I can't sleep because of anxiety.</i> Will my personal information be leaked due to my call? <i>I am also worried</i> that my parents will know about it because I have registered their credit card. <i>Please help me.</i>	5.0
<i>URGENT!</i> When I plugged the USB cable connected to my smartphone into the computer that my company owns, the message “Do you want to load images” was displayed. I immediately unplugged it. This doesn't leave any images of my smartphone on the computer, does it? I don't want my images to be leaked. <i>I'm very anxious.</i>	5.0
I'm a student. My device is restricted by <security software's name> that my parents set. Is there any way I can unlock it without using my parents' devices? If anyone knows, <i>please answer.</i>	4.0
I got this email. This is a scam email, right?	3.0
In general, are anti-virus apps needed for smartphones?	2.0
Who is making phishing emails that spoof credit card companies?	2.0

Words in italics indicate signals expressed by askers, such as expressions of urgency, anxiety, or a call for help, that would affect the coders' judgement. Note that coders did not rate seriousness based solely on the number of signals but rather did so comprehensively. Questions were originally posted in Japanese.

The Nerd Factor: The Potential of S&P Adepts to Serve as a Social Resource in the User’s Quest for More Secure and Privacy-Preserving Behavior

Nina Gerber

Technical University of Darmstadt

Karola Marky

Leibniz University Hannover, University of Glasgow

Abstract

There are several ways to inform individuals about secure and privacy-preserving behavior in private social environments. Experts who are versed in security and privacy (S&P), who might be social peers, such as family members or friends, can provide advice or give recommendations. In this paper, we specifically investigate how S&P adepts inform peers in their private social environment about security and privacy. For this, we first conducted thirteen in-depth interviews with S&P adepts, revealing 1) their own S&P behavior and strategies in their personal lives, 2) obstacles in S&P conversations with peers, 3) situations in which S&P adepts intervene in the behavior of others, and 4) the perception of S&P adepts and stereotypes. Based on the interview results, we conducted three co-design workshop sessions with S&P adepts to explore options to better support S&P adepts informing their peers about secure and privacy-preserving behavior.

1 Introduction

In 2022, more than 22 years after Adams and Sasse’s seminal paper “Users are not the enemy” [3], many users are still struggling to protect their IT security and privacy (S&P). Those of us who are relatively well versed in the subject know that users are indeed not the enemy, but we still struggle to help users in their efforts. Accordingly, while many researchers and developers are engaged in understanding lay users’ mental models and developing tools to help them protect their S&P; direct, interpersonal one-on-one help or influence among friends and family rarely happens.

Yet, from investigations in other domains, such as general technology support [52], home security [51, 54], or professional contexts [38, 62, 63], we learned that help from knowledgeable peers has a high potential to impact the behavior of lay users positively. The idea of helping lay users through the social influence of people with technical backgrounds is not novel: In 2012, Lipford and Zurko [48] proposed a new paradigm for influencing people to behave securely. Instead of focusing on the usability of security tools, they argued for using social processes (e.g., building a security “neighborhood watch”) where people from a user’s social network watch over their security decisions. Four years later, Redmiles et al. [56] stated that people with technical backgrounds should be supported in responding to security advice requests from their peers, since even a small set of essential security advice might have a large possible impact on lay users.

Still, little research has been conducted in this area to date. Findings from related studies tend to suggest that tech-savvy individuals have little interest in actively intervening in the security and privacy behavior of their social environment [52]. Our research addresses this issue and seeks to determine what barriers underlie this and how those can be overcome.

Our goal is to (1) investigate the status quo of S&P supporting in the private context (i.e., when, how and why do S&P adepts (not) support people in their private social environment), and (2) explore options to overcome existing barriers. To this end, we first conducted in-depth interviews with 13 S&P adepts, i.e., people who are fairly versed in IT security and privacy. Building on the results, we then conducted three co-creation workshops with another 11 S&P adepts.

We find that S&P adepts only try to educate people from their social environment about S&P with whom they have a close social relationship. This may be because a trusting relationship is essential for S&P adepts to feel able to address what they consider to be a sensitive topic, where the interlocutor may quickly feel criticized or lectured. Unsolicited advice is given mainly for S&P issues that require explicit interaction, such as passwords. One reason for this could be that for more complex technical issues a common terminology has

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

to be found first, and S&P adepts often have to struggle with users having wrong mental models of what they are trying to explain to them. Opportunities to promote exchange between experts might help them to build a better knowledge foundation for promising approaches in assisting lay users. Finally, we learned that S&P adepts require possibilities to improve their knowledge further (e.g., through open access publications) and that rewards might motivate them, such as recognizing support-giving as a professional achievement. Our paper makes the following contributions:

- We provide an in-depth investigation of S&P knowledge exchange and support between S&P-savvy individuals and their peers in a social context.
- We explore several avenues to overcome existing barriers to S&P support.
- We provide recommendations for S&P adepts and the research community that help to facilitate the development of S&P adepts as a social resource for the improvement of users' S&P behavior.

2 Related Work

To set the scene for our work, we report research about the social influence on S&P behavior, social support and S&P advice, as well as the perspective of S&P experts.

2.1 Social Influence on S&P Behavior

The role of social influence on people's S&P behavior has been extensively investigated by Das et al. [12–16]. They conducted a survey to investigate triggers that impact S&P behavior [12], and found that 39% of the triggers were social. The reported sharing rates were rather low and reasons included perceived obligations to protect others and noticing insecure actions. Primary reasons for not sharing were lack of desire and that others did not need to know about one's S&P practices. Das et al. [13] further explored under which circumstances and for which purposes people talk about S&P with others. Their results confirmed that social interactions, e.g., observing others, were powerful triggers for improving S&P behavior. Reasons for starting S&P conversations either focus on warning others or seeking advice. However, S&P experts are often considered paranoid, "hyper-secure", and behaving "above and beyond" (p.153) [13], a finding that has also been shown in previous studies, e.g., email encryption was considered paranoid [32]. Many security-savvy participants avoided the topic since they worried about being socially inappropriate or, e.g., too preachy [13]. This suggests a large untapped potential: if we better understand how S&P adepts can be motivated to share their knowledge with their social environment, this in turn, could act as an effective trigger to improve the S&P behavior of less tech-savvy individuals. This paper represents a first step towards achieving this goal.

In a subsequent survey study, Das et al. [16] focused on sharing S&P news. They found S&P experts want to share news, e.g., because they feel responsible. In two large-scale studies, Das et al. [14, 15] found that people were influenced by their (Facebook) friends in both directions when adopting or rejecting security features.

Other studies focus on social influence in the privacy context [2, 10, 22]. E.g., Emami-Naeini et al. [22] found in a vignette study with MTurkers that friends denying data collection, and privacy experts allowing data collection mostly influenced people's decisions when interacting with IoT devices. Social influence has also been proven effective in the nudging context, i.e., stating that a minority of users like themselves had accepted cookie use could nudge participants away from accepting cookies [10]. A very recent study by Krsek et al. [45] showed that that non-personal social influence has a high potential to motivate users to apply settings different from the defaults offered by Facebook.

A recent interview study shows that implicit social privacy norms on social media among young adults [55] exist. Yet, sanctions that follow violations are mostly indirect, non-confronting and consequently offer no possibility for violators to learn. Our participants may be particularly affected by this, as it can be assumed that they have particularly strict norms. Thus, they could benefit from solutions that address this issue, and, at the same time, add value to society as a whole by shifting the general social norms towards greater privacy protection.

2.2 Social Support and S&P Advice

Prior work showed that people rely on their social network for general tech and S&P support [18, 27, 29, 46, 51, 52, 56]. Using a combination of semi-structured interviews and a survey, Nthala and Flechais [51] found that users often seek advice or technical help from others they perceive as competent and trustworthy, mostly family and friends. Further, security support is sometimes delegated and occasionally knowledgeable participants offer unsolicited support, e.g., when noticing insecure behavior. Based on these findings, we chose to focus on relatives and friends as receivers of S&P support, and also include questions about responsibility, advice seeking, and intervening.

Two studies of privacy advice sharing among developers on online platforms (e.g., "Stack Overflow") show that privacy-related conversations are mostly motivated by external events, e.g., updates that require actions from developers [47] and advice is mostly shared as links to official documentation [63].

Poole et al. [52] conducted semi-structured interviews to investigate why and how tech-savvy people provide support for social peers. Usually, tech-savvy people are approached unsolicited and quickly gain a reputation. While most participants were happy to support as teenagers, it became increasingly difficult as they got older. Still, they continued to provide

support based on a sense of obligation which inspired our work. Consequently, we focus on (1) how to facilitate S&P behavior among people who are closest in our participants' social network, and (2) how this behavior influences the various aspects of the social relationship. Regarding security, Poole et al.'s participants reported engaging in "digital housekeeping" when visiting family members, e.g., updating software. Although helpers did not promote to be experts, they tried preserving that image by avoiding situations in which they cannot help. We also pick up on this in our interview guide.

In a representative US-survey, Redmiles et al. [56] found people with higher skills to be more likely to get S&P advice from work, whereas others get it from family, friends, and service providers. Fagan et al. [23] found that when deciding about whether to follow security advice, people tend to focus on individual aspects rather than social ones. Further, self-rated security expertise does not make a good predictor for security behavior, which we considered in our recruitment process.

Forget et al. [27] combined behavioral and configuration data with interviews with mostly older adults. System maintenance, including security, was often outsourced to "residual experts", usually family members. However, those were not always experts and sometimes had erroneous problem interpretations, leading to serious security threats. In another interview study with older adults, Frik et al. [29] confirmed that security and privacy settings are often delegated to others, like family or community members or technical experts.

2.3 The Perspective of S&P Experts

Few studies address the behavior and judgment of experts with respect to helping ordinary users in their S&P efforts. Ion et al. [41] compared security practices of experts and non-experts in a study combining interview and survey data. Not surprisingly, they found experts to show better security practices than non-experts. Further, non-expert users need advice with installing updates, password managers, and two-factor authentication (2FA). In a recent replication study, Busse et al. [7] identified password security, 2FA, links, attachments, and updates as topics that primarily call for expert advice.

Haney and Lutters [39] conducted interviews with security advocates, i.e., individuals who professionally promote security practices. An important aspect of this task is establishing trust. Tahaei et al. [62] investigated privacy-savvy developers in their professional context, identifying motivations, challenges, and strategies to promote privacy-friendly development. Collaborative solutions and guidelines from companies were identified as promising solutions. While Haney and Lutters [39] investigated professionals interacting with strangers, we focus on the potential of S&P-savvy individuals to motivate and facilitate secure and privacy-friendly behavior in their existing social network, where strong relationships of trust should already exist. Existing research on security

advocates [37, 38, 62] also confirms the importance of non-technical, interpersonal skills, including the need to make sure people do not feel stupid for knowledge gaps. Perhaps due to this fact, security advocates also have backgrounds in non-technical fields, such as psychology or education [37, 38]. This confirms findings about people seeking advice from others in their social network they consider experts, but not necessarily turning to those with a technical background [56]. Likewise, we focus on individuals knowledgeable in the fields of S&P to a certain degree and thus able to facilitate secure and private behavior of others.

Haney and Lutters [39] further identified techniques used by security advocates to overcome negative perceptions like being honest about risks, making one's language understandable, and engaging listeners through reward systems or relatable narratives, and metaphors. Haney and Lutters [39] focus on analyzing the status-quo since security advocates are already doing their best to promote secure behavior, whereas we aim to understand what we would need for S&P-savvy individuals to be tapped as a valuable social resource in the quest for more secure and more privacy-preserving behavior.

The importance of pursuing this line of research is further emphasized by findings of a survey study. Rader et al. [54] showed that stories have great potential to change security attitudes and behavior for the better. Stories told in the home context are more likely to change behavior compared to professional contexts. Yet, stories told by people knowledgeable in security are more likely to be retold, thus influencing more people. In a further analysis, Rader and Wash [53] found experts tend to focus on *how* an attack is conducted and prevented, whereas non-experts were mainly interested in *who* carried out an attack and *why*. The authors recommend experts should consider this in their communication with non-experts.

3 Study I: In-Depth Interviews

First, we wanted to gain a deeper understanding of the topic by conducting in-depth semi-structured interviews with S&P adepts and learn about their experiences with sharing their knowledge or motivating other people in terms of S&P in the private context. We conducted thirteen interviews until we reached data saturation. The interviews were held via a video-call tool, with an average duration of about an hour. All participants received a 20€ gift card for an online shop [34]. The interviews were audio-recorded and transcribed for analysis. We conducted two pilot interviews with experienced researchers to check our questions for clarity and comprehensibility and refined our interview guide based on the feedback.

3.1 Method

Participants. We recruited 13 participants by mailing lists and word-of-mouth. We used university mailing lists (also

addressing interested non-students) including those of collaborators, reached out to our professional contacts (researchers and practitioners from various institutions and organizations) and contacts of collaborators, and were open to snowballing. All participants first completed a screening survey to make sure they qualified as S&P adepts (for the detailed scores, the reader is referred to Table 2 in Appendix A.1). All participants had been working intensively on the topic for several years, either in the context of research activity or in another professional context. The participants were between 21 and 56 years old. Two of the participants self-identified as female, eleven as male. All participants were residing in the UK or Germany at the time the study was conducted. For detailed demographics, including occupation and highest education, the reader is referred to Table 1 in Appendix A.1.

Study Procedure. Prior to the interviews, participants were asked to complete a screening survey to ensure that they met the criteria for study participation. We used the Security Behavior Intentions Scale (SEBIS) [20] to measure security intentions, the six-item validated self-report measure of security attitudes (SA-6) [24] to measure security attitudes, the Internet Users' Information Privacy Concerns Scale (UIPC-8) [35,50] to measure privacy concerns, the Online Privacy Literacy questionnaire (OPLIS) [65] to measure privacy knowledge, the Affinity for Technology Interaction Scale (ATI scale) [28] to measure technical affinity, and two self-constructed items. We then contacted participants who qualified for study participation to set up an appointment and asked them to sign the informed consent form via email. The interviews consisted of six main parts (see Appendix A.2).

First, we thanked the participants, made sure they had signed the informed consent form, and gave them the opportunity to ask questions. We then asked about *their S&P behavior*, including social aspects such as whether they talked about this with others and how others might have reacted to their behavior in the past. Second, we asked about their experiences with observing insecure or privacy-unfriendly behavior of others, including their feelings on this topic and whether they had ever *interfered* in such a situation. Third, we asked whether other people usually *asked them for advice* on S&P issues, including who, on what issues, and how they responded to that. Fourth, we asked whether they *feel responsible* for the S&P of other people, including who, why, and how this manifests itself in their behavior, e.g., by sharing news about security incidents and data breaches, or doing digital housekeeping. Fifth, we asked about *bad experiences* with giving advice to others or interfering and the fears associated with it, such as being socially awkward or straining the relationship. Sixth, we asked whether and why others perceive the participants as *S&P experts* and whether they are afraid of coming across as paranoid or tech nerds. We also asked about gender effects and loosely relied on the repertory grid technique [60] to ask which characteristics of a person they

associate with IT security and privacy behavior. Finally, we asked the participants to complete another short questionnaire on their demographic information.

Data Analysis. We used thematic analysis [5] to analyze our transcribed data. The author that conducted most of the interviews first read through all transcripts multiple times and then coded all interviews at sentence level to develop a codebook, going back and forth several times to refine the codebook. That author then went through all transcripts and used the codebook for another round of coding. Next, another author went through the entire coding and marked all the codings they disagreed with, including passages where a code should be added (following recommendations for thematic analysis against conducting multiple independent codings and calculating ICR [6], p.278-279¹). The authors then came together to discuss the notes of the second author and agree on a final coding. After this, both authors grouped the codes into six main themes.

Ethics. The study met all requirements for studies with human participants given by our ethics commission. Before the study, all participants were informed about the study purpose and conditions, informed that they could quit the study at any time without any negative consequences, and asked to confirm their participation by signing an informed consent sheet. Although we used a video-call software, we only recorded the audio track by using another software, and stored it locally on the interviewer's computer. Further, all participants were free to turn off their cameras for the interview. All data was handled confidentially and any identifiable information was deleted in the transcribing process. We decided to compensate the participants with gift cards for the "Greenpeace Magazine Warehouse" webshop, which is associated with Greenpeace, to support a charitable organization, but reward the participants with a product of their own choice from the store.

Limitations. Like most qualitative and exploratory work, our study is subject to several limitations. First, we rely on self-reported data, which might be biased due to social desirability, availability bias, and wrong recalls or self-assessments. We focus on social aspects, which may be especially sensitive for this kind of bias. Still, we aimed to gain a first understanding of IT security and privacy adepts' mindsets and experiences. Further research is needed to explore this topic in more depth. Second, we used a convenience sample, using personal networks and those of colleagues, as well as word-of-mouth. We wanted to target, inter alia, experienced researchers and practitioners in the field without making our project too public in order to preserve anonymity for the publication process. We thus decided not to recruit participants at public events such as scientific conferences or fairs, as it has been done in some

¹ According to Braun and Clarke [6], qualitative research acknowledges the researcher's influence on the research process. Conducting ICR as a means to "prove" reliability is thus seen as not applicable for thematic analysis, as data should not only be described, but also interpreted.

prior studies which aimed to recruit IT security experts [7,41]. As a result, our sample is skewed towards male and young participants, rather homogeneous in terms of culture (UK and Germany) and background, and also includes university students who may be knowledgeable about security but have limited professional security experience. Hence, our sample is hardly representative of all S&P adepts but rather serves as a first step in shining light on this complex topic. Further research is needed to explore the perspective of S&P adepts with different professional and demographic backgrounds, particularly from non-western cultures. Third, we mainly focus on the status quo in the interviews, i.e., we asked how our participants currently interacted with others in the context of S&P, and explored possible obstacles for interfering or motivating others towards more secure and privacy-preserving behavior. Yet if we aim to use S&P adepts as a social resource, we also need to know how social interactions in this context could be facilitated. We address this point in Study II, drawing on a more solution-oriented, participatory approach, i.e., co-creation workshops.

3.2 Results

In the following, we describe our results and provide quotes where applicable. Considering the explorative nature of our study and the limited sample size, we deliberately refrain from reporting exact numbers to avoid the appearance of generalizability. Instead, we will mention a rough frequency frame to emphasize topics that were mentioned by many participants.

3.2.1 Own S&P Behavior

Protection Strategies. When asked about what protection measures they applied in a everyday context, most of our participants referred to using secure authentication mechanisms, i.e., unique, secure passwords, 2FA, and locking devices. Other important security measures include using antivirus programs, updates, and checking emails for phishing. Reported privacy protection measures focused on avoiding tracking (e.g., by blocking or deleting cookies, using private modes in browsers, or VPN), and minimizing data collection (e.g., refrain from using social media and soft- or hardware from certain vendors, and covering one's webcam).

Social Conflicts. Half of our participants mentioned to have experienced social conflicts due to their S&P behavior, mostly with friends and family members or significant others. These conflicts arose from our participants not wanting IoT devices in their homes due to privacy concerns or expressing these concerns when visiting other IoT-equipped households, not wanting to share their passwords or location, and not wanting to use social media although the significant other wanted to tag their spouse on Facebook. Further, very few participants reported foregoing security or privacy to avoid delaying others (e.g., in a meeting), and to avoid being socially excluded.

3.2.2 Intervening in Others' S&P Behavior

Less than half of our participants said they had ever actively interfered in someone's S&P behavior. Basically, our participants only get involved if it affects them (e.g., their own data is involved or they personally would suffer from the consequences) or if they feel a responsibility (professionally or privately, because people rely on their advice or are close to them) (the latter confirming [13, 16, 51, 52]). Only very few participants reported to have negative experiences with giving solicited or unsolicited advice, this includes recipients of the advice having problems with their OS after an update, and data loss after data encryption.

Raising Awareness. Almost all participants reported that at some point, they had tried to raise someone's S&P awareness. In most cases, the recipients of these efforts were family members or friends. Explicitly not addressed were persons with whom our participants have no close relationship. The topics addressed varied, and included data breaches, hacking attacks, scams, exploits, changes in privacy policies, eavesdropping, and new as well as established protection tools such as 2FA or the Tor browser. About half of our participants referred to possible consequences of neglecting S&P protection to make the importance of this protection clear. Other reported strategies were checking the recipient's email address on websites like "Have I Been Pwned", pranks, and trying to initiate a cost-benefit analysis for data sharing. Further, some participants emphasized that a negative framing should be avoided.

Motivating. Half of our participants reported efforts to motivate others in terms of S&P protection. Still, most of these efforts were limited to authentication (i.e., choosing secure passwords, keeping and entering them secretly, using 2FA). Only one participant each referred to the use of secure messenger apps, and operating systems, as well as doing updates.

Being Asked for Advice. Almost all participants reported to be asked for advice on S&P topics regularly. The most popular topics for advice focus on authentication (secure passwords, 2FA), which tools can be used for protection, and data collection (e.g., which services collect what kind of data, how does personalized advertisement work). Some participants also reported to be asked about whether it is advisable to use certain services and devices such as Google smartphones from a S&P point of view, and to give advice on (potential) spam and phishing emails. Most of our participants reported that family members, especially their parents, asked them for S&P advice frequently. Also, about a third said they were asked for advice by friends, and only a few mentioned acquaintances, colleagues, or others (confirming earlier findings [51, 52]). Most of our participants liked to be asked for advice as they feel valued and enjoy being perceived as an expert in this field, while others are primarily pleased that their social network is dealing with the topic at all. Still, about a third of our participants also mentioned negative aspects of being asked

for advice, such as the pressure of giving good advice, being asked too often about the same topics, and being asked questions without clear answers (e.g., P5: “*Especially the question ‘Is it really secure then?’ I mean, nothing is ever secure.*”)

Feeling Responsible. Half of our participants felt responsible for the S&P behavior of their parents, mainly because they had more expertise in this area, had given them S&P advice previously, and their close relationship. Most of these participants also said they would engage in “digital housekeeping” [52]. Only one participant said they also felt responsible for the S&P behavior of other family members and close friends, and another participant who worked as IT admin for their customers’ S&P behavior. Yet, these results should be taken with a grain of salt, because our sample was rather young, which might be a reason why no (older) children were considered, for whom they might also would feel responsible. Interestingly, most participants said they did not feel responsible for the S&P behavior of their social network since they felt that the decision to (not) act securely and privacy-preservingly was a personal one that they had no right to interfere with. Seemingly, this did not apply to their parents, perhaps because adult children also frequently interfere in other areas of their parents’ lives, e.g., in medical matters. Very few participants also said that they did not feel responsible for others’ S&P behavior as this would involve too much effort.

3.2.3 Conversations about S&P

Trigger. Our participants reported several opportunities that sparked conversations about S&P: sitting together with others in front of the computer, which offers the additional opportunity for others to observe one’s own S&P behavior (e.g., using 2FA, tracing blocker, the Tor browser) and ask questions, if they were using someone else’s computer and thus saw their security and privacy settings (both confirming [13]), giving general technical support, or if others received spam emails, were asked to take security measures by their provider, or saw news about current hacking attacks or scams.

Topics. Most participants reported talking with others about S&P-related topics, mainly to share experiences with protection measures and tools, discuss the pros and cons of not using social media and messengers, and inform others about data breaches or security incidents. More than half of our participants said they would primarily talk to other knowledgeable, tech-savvy people about S&P, as these were – in contrast to less tech-savvy people – interested in these topics.

3.2.4 Obstacles

All participants mentioned obstacles to improving the S&P behavior of their social network, e.g., by giving solicited or unsolicited advice.

Lack of Interest. More than a third of our participants complained about a lack of genuine interest in these topics (P2: “*It’s like when someone tells me something about brass band music. I would nod my head and say ‘That’s interesting’. But that wouldn’t really interest me, and that’s exactly how I feel the other way around.*”), confirming prior work in the professional context [39].

Social Aspects. Others referred to social aspects, such as not wanting to bother others, wanting to avoid negative reactions, not wanting to criticize others, and avoid being perceived as preachy or intrusive.

Lack of Resources and Opportunities. A few participants also mentioned a lack of resources, i.e., facilitating others’ S&P behavior being too time-consuming or too much effort, and triggers, mainly because S&P behavior is not directly observable in most cases.

Lack of Legitimacy. Some participants were also reluctant to give advice or interfere in others’ behavior since they themselves did not always act as securely and privacy-preservingly as they want. Regarding privacy, one participant each also explained that there is no “right” level of privacy and thus people have to make their own decisions, and that privacy is an especially sensitive topic as some people may be quickly offended because you imply that they are trying to hide something.

3.2.5 Reactions

Others’ Reactions. Overall, our participants reported more positive than negative reactions when they gave solicited or unsolicited S&P advice. Positive reactions included interest, gratitude, sympathy, and acceptance, whereas negative reactions mainly refer to disinterest. If others observed their S&P-aware behaviors, our participants mainly got neutral reactions, i.e., others were non-judgmentally surprised about their behavior. Still, like in [13], a few also reported being smiled upon (e.g., P3: “*I think if you’re interested in data security, you always get these joking sayings that you’re one of the tin foil hatters or paranoid people.*”) Most participants said their advice had not brought about any long-term change in the recipients. Only one participant reported to have had a lasting influence on others S&P behavior.

Own Reactions. About half of our participants said they understood that other people’s behavior was not always secure and privacy-preserving, because they also knew the costs of such behavior and could well understand if other people were not willing to accept them. Accordingly, most people tend not to take it personally when other people ignore their (solicited or unsolicited) advice. Very occasionally, however, participants reported that such ignorance of the topic was perceived as a personal attack, as it was “*part of one’s own identity*” (P1) and that they felt thus somewhat “*affronted*” (P4).

3.2.6 Perception as Expert and Stereotypes

Expert. Prior research [27, 51, 56] indicates that people do not only delegate their S&P to people who have a professional background in IT security or computer science, but also to knowledgeable people with a non-technical background. This also applies to “security advocates” [37–39] who deliberately chose this path for themselves. Still, about two-third of our participants reported being considered an expert in S&P due to a technical study or profession. Some participants, however, attributed their expert status to their private interest in the topic, and on support provided in the past. Most participants, self-identifying as female and male, thought gender has an impact on whether someone is perceived as an expert, with all agreeing that it tends to be more difficult for females (who have the same knowledge as males) to be seen as experts. This is attributed to common stereotypes, e.g., P9: “*The technology nerd is imagined as an overweight, male basement dweller.*”

Stereotypes. S&P behavior of non-experts was often associated with age. Our participants tended to rate younger people (i.e., teenagers) as oversharing on social media and thus privacy-unfriendly and older people (i.e., over 50) as insecure due to lacking technical knowledge. Further, technical expertise and awareness of possible consequences were associated with adequately secure and privacy-preserving behavior. Very secure and privacy-friendly behavior, on the other hand, was associated with anxiety.

3.2.7 Summary of Interview Findings

S&P adepts mostly only try raising awareness of people they are close to (friends and family), but enjoy being asked for advice. Negative aspects of being asked for advice, however, include the pressure of giving good advice, being asked too often about the same topics, and being asked questions without clear answers. While reactions to (un)solicited S&P advice are mostly positive to neutral, S&P adepts are nevertheless afraid of negative reactions and struggle with the fact that primarily unsolicited meddling can lead to socially awkward situations. One difficulty is also getting started on the topic, since there are not many triggers for talking about S&P. Communication on S&P topics, therefore, takes place primarily between S&P adepts, as it is assumed that others are not interested in the topic. In general, privacy in particular is seen as a matter for everyone to decide for themselves. An exception seems to be parents, for whom S&P adepts feel responsible.

4 Study II: Co-Creation Workshops

While the interviews in Study I primarily focused on the status quo and aimed to identify potential barriers to S&P support from social peers, we took a more solution-oriented perspective in the second study. We conducted three co-creation workshops with three to four S&P adepts each to explore how S&P

adepts can be supported to improve the S&P behavior of people in their social environment. The co-creation workshops were held via a video-call tool, with an average duration of about two hours. We used a Mural whiteboard² for facilitating the collaboration. Participants were offered compensation of €25 or £20, however, eight of the eleven participants chose not to be paid as their primary interest in participating was to support research in this area. The workshops were audio-recorded and transcribed for analysis. We conducted a pilot workshop with experienced researchers to check the procedure and materials and refined our workshop guide based on the feedback.

4.1 Method

Participants. We recruited 11 participants for three co-creation workshops by mailing lists and word-of-mouth. Like in the interviews, all participants had been working intensively on the topic for several years, either in the context of a research activity or as practitioners, and were currently residing in the UK or Germany. Four of the participants self-identified as female, seven as male. For detailed demographics and the screening data, the reader is referred to Appendix B.1.

Workshop Procedure. The co-creation workshop followed the first steps of a design sprint [44]. Before the study, participants were asked to complete a shorter version of the screening survey from the interview study (based on the interview participants’ feedback that the survey was too long, we removed the SEBIS and all OPLIS scales except for the most relevant *technical aspects*) to ensure that they met the criteria for study participation. The screening survey started with a consent form that covered the entire co-creation workshop. The workshop started with an icebreaker session (approx. 7 min) where participants were asked to draw their mood to familiarize themselves with the Mural board and then introduce themselves to the others. During that time, small talk about social S&P situations was possible.

Map and Target. Next, the participants were introduced to the scenario by watching a 2-minute presentation held by the moderator. After the presentation, the participants’ attention was drawn back to the Mural board. Their first task was brainstorming facilitators and obstacles of supporting their social peers in behaving securely and privacy-preservingly. This task was meant to provide a neutral introduction to the topic and initiate an exchange between the participants. In the further course, the brainstorming results served as a source of inspiration for the development of the co-creation solutions. To support the brainstorming, we used the miracle question [17], which originates from systemic therapy and in which clients adopt a solution-oriented perspective in which they are asked to imagine that a particular problem no longer exists. Furthermore, the S&P adepts were instructed to write down their

²<https://www.mural.co/> last-accessed Feb. 16 2022

thoughts on sticky notes and discuss them. After the brainstorming, the S&P adepts were asked to agree on a common goal for the remainder of the workshop by collaboratively formulating the problem as a “How Might We” question [11].

Sketch. Once the goal was clear, we used the 5-3-4 method [58] to co-create solutions. Using this method, each participant first wrote three ideas on sticky notes. Second, the S&P adepts shifted clockwise and could either add three new ideas or extend the ideas from their successor. This was repeated until the S&P adepts fully rotated once. Each rotation was limited to three minutes. After that, each participant gave a short presentation (approx. 1 min) of their ideas. The others were allowed to ask questions.

Decide. Once the ideas were clear, the participants were asked to vote for ideas following the how-now-wow principle [66]. This approach addresses the issue that people tend to brainstorm highly original ideas, but usually settle on well-known solutions in the further development process. Brainstorming ideas are evaluated on two dimensions, originality and easiness to implement: now-ideas are normal ideas that are easy to implement, how-ideas are original ideas that are (too) hard to implement, and wow-ideas are original ideas that are easy (enough) to implement. For this, each participant received 15 dots (5 per category) and was asked to place these dots on the sticky notes from the previous round. Next, the ideas were sorted into the how-now-wow matrix based on the voting. After the sorting, the S&P adepts discussed the results. They were specifically asked to explain their voting and discuss how wow-ideas could be realized as solutions.

Finally, the adepts were thanked for participation, could ask questions, provide comments and were reimbursed.

Data Analysis. To analyze the results, we had two iterations – one for the data collected on the Mural board and one for the transcribed recordings. Similar to the interview analysis, we used thematic analysis [5]. First, one author, who was present at all workshops, familiarized with the Mural boards and then developed a codebook based on the sticky notes. The codebook was then applied to all data collected. Next, a second coder went through the coding and marked disagreements that were discussed later on. To analyze the transcripts, both authors coded all transcripts at the sentence level to iteratively develop a codebook considering also the first round of coding the board. One author then went through all transcripts once more to apply the codebook. The coding was verified by the second author. The authors then came together to discuss the notes of the second author and agree on a final coding considering the data on the board and the transcripts. After this, both authors grouped the codes into four main themes.

Ethics. For the co-creation workshops, we took the same precautions like in the interview study. Since the co-creation workshop involved other participants, the co-creation participants were informed about that before the study. During the

workshop, participants were encouraged to have the camera on but were not required to do so. With a screen-recording software, we captured the Mural board but not the video call.

Limitations. There are several limitations based on the used method and sample. First, since the sample is rather small and homogeneous, our results should be considered as first insights that should be validated and broadened by future investigations. Considering the sample composition, it was biased towards researchers because more researchers than practitioners participated. Practitioners might struggle with other issues than researchers who are used to teaching. Yet Usable S&P researchers are particularly versed in the topic and thus might come up with a plethora of solutions in a shorter amount of time compared to other S&P adepts. Furthermore, the sample was slightly skewed towards male and young participants, of whom all had a university degree and were currently residing in the UK or Germany. Due to COVID-19, we opted for an online workshop, since capturing qualitative data online can be suitable [49]. This allowed for a more diverse and international sample. Still, in-person co-creation workshops are better for designing solutions using paper and sketching, while online workshops lead to more text-based co-creation. Our workshops thus focus more on generating concepts and identifying obstacles and facilitators. As a consequence, the co-creation serves rather to generate research data instead of designing solutions for actual use. The results hence must be enhanced by in-person workshops in the future.

4.2 Results

Four themes for the co-created solutions emerged during the analysis, which are described below.

4.2.1 Set a Constructive Dialogue Space

Create a Constructive Atmosphere. A reoccurring theme during the workshops reflected in various co-created solutions is the issue of addressing S&P topics in “normal” conversations without being judgmental or preachy. Some participants, hence, thought it would be helpful to establish social norms for such discussions. An important point here is finding a balance between creating awareness and accepting the user’s S&P attitude, which often differs from that of the S&P adepts. In addition, the use case of the person seeking help and their lack of knowledge should be accepted. Furthermore, users seeking support should be able to address their problem without the S&P adept making a big deal out of it, e.g., P1WS1: “[It works] as soon as someone has the feeling, I can now also ask a question and say I have but only five minutes time. And somehow you get an answer in five minutes that you can work with.” Last but not least, the recipient should be given time to reflect on what has been said, i.e., the S&P adept should take up the topic again after a period of reflection if necessary, but

should not push it too hard in the initial conversation.

Establish Contact Between S&P Adepts and Users. Among the most important issues for our participants was how to establish contact between supporters and recipients. Like in Study I, our participants were reluctant to raise S&P issues themselves for fear of violating social norms. The challenge of not being in a position to give advice because of doubts about having enough expertise (security) or not always behaving optimally (privacy), an aspect that was also mentioned in Study I, also plays a role here. Accordingly, some of the considerations in the co-creation process related to how to make it clear to the outside world that one is a suitable contact person for the topic. There are officially defined roles and responsibilities for this in the professional context, e.g., P1WS1: *“For example, I was never asked about data protection until I was a data protection officer, because then it was clear that I am in charge.”*, which could possibly also be transferred to the private context. One option for this, which has already proven itself unofficially in practice [27, 29, 51], is to officially delegate S&P. Like P1WS1 reported on his experience as a data protection officer: *“People were much more likely to bring things up to me if they felt it wasn’t their problem afterward.”* Another co-created solution that might fit this point was to offer oneself as an S&P adept in clubs and communities, as this reaches a lot of people, and word of availability as an advisor spreads quickly in the local environment.

If S&P adepts do want to proactively approach people, finding an entry point is challenging, as it is not a common topic when talking to one’s social environment, e.g., P2WS2: *“I don’t usually come into situations where my friends or family ask me about like security and privacy behavior and when I speak to those people I usually have other topics in mind than just randomly starting giving advice on internet security.”* Certain publicly effective events, such as changes in legislation [59], changes in terms and conditions (T&Cs) of popular products like WhatsApp [40], or contact tracing apps [8] reported in the media can serve as an icebreaker for conversations. In this case, either people approach the S&P adepts, or the topic serves as a reminder for the S&P adepts that they could approach people. Movies could also be a good conversation starter. Although they often paint an unrealistic picture of S&P [30], they can serve as a starting point to explain how something actually works. A broader approach to this would be to conduct awareness campaigns. Although this has already been suggested in the literature [33], the goal would not be to raise S&P awareness per se, but, as P2WS2 put it: *“It is probably mostly about giving experts a stage. This one is mostly about providing like the pressing issue are for the public to actually be motivated to learn about privacy and security issues or find their experts of trust and to ask questions.”* One suggestion was to do this in conjunction with action days such as Safer Internet Day or Password Day [25, 64].

Build Trust. Many S&P adepts were concerned with the ques-

tion of how they could give those seeking help confidence in their abilities as supporters. Possible solutions for this were discussed, e.g., a kind of certification or score that changes depending on the quality of the help provided or the advice given. This would not only strengthen the subjective trust of the user, but also of the S&P adepts in their own abilities. Tools that enable users to experience S&P settings, such as AmIUnique [26], could also be used to enable users to judge the quality of advice themselves. One participant also noted that S&P are sensitive topics that are better discussed in (confidential) face-to-face conversations rather than, for example, via texts or in public. Still, some participants thought that credibility could only be achieved if there are large, trustworthy institutions, such as courts, that back up certain statements on S&P, as these are often difficult to believe: P4WS1: *“[They need to say] ‘That’s how tracking works on the internet. Yeah, that’s super creepy. I’m sure you guys don’t want that.’ If you do it as individuals, doesn’t matter how great people think you are, then you sound like a conspiracy theorist.”*

4.2.2 Harness the Potential of Exchange

Promote Exchange Between S&P Adepts. Another topic that came up multiple times during the workshops was the desire to share successful tactics and strategies for support with other S&P adepts. This wish, however, remained on the surface, the participants had no concrete idea of how such an exchange could be designed, except for the vague idea of a platform. Still, a concrete co-creation idea, which aims in a similar direction, is to refer people seeking help to other specialists who are more knowledgeable in the area concerned, similar to the way it is done in medicine. This approach would offer the possibility to admit without loss of face that one does not know something and still have the feeling that one has helped the person seeking help (at least a little) by referring them to the right place. On the other hand, this could lead to helping people with whom one does not have a close relationship. Since most S&P adepts tend to help out of a sense of responsibility for their immediate social environment [51, 52], different facilitating conditions might have to be created at this point. One example of this was to give the S&P adepts the opportunity to offer consultation hours during their working hours.

Promote Exchange Between Users. Another idea of relieving the S&P adepts was to refer people seeking help to other users who had already been helped with the same problem, in a kind of snowball system, e.g., P1WS3: *“That you say, hey, I’ve already explained something similar to this guy, go see him. If this guy then explains it again to a buddy, then it becomes even clearer for him.”* This idea also emerged in a more institutionalized context such as a school, where existing peer systems such as dispute resolution or mentoring programs are maintained by having each new generation step up and pass on their knowledge to the next generation.

4.2.3 Facilitate Knowledge Transfer

Find Common Ground. Two challenges that can arise when communicating knowledge are the question of a shared language, which must form the basis for successful communication, and the fact that users often have incorrect mental models of the matter concerned. The first problem has been addressed in several co-creation solutions, e.g., via the development of a dictionary that translates terms between S&P adepts and users, but also via the specification of certain terms, such as virus/malware, that do not keep changing over the years. With regard to the latter point, e.g., the development of metaphors was suggested (confirming earlier findings [39]), which offers great potential for presenting technical issues such as end-to-end encryption (E2EE) in a comprehensible way [61]. There was also a frequent request for training S&P adepts in explaining facts in a popular scientific way. Since S&P adepts are often asked for advice on specific problems, a flowchart, for example, would be a promising tool for the S&P adepts to use as a basis for deciding what knowledge they need to convey in order to understand the actual matter of interest.

Show S&P Relevance. Several co-creation proposals were aimed at making users aware of the relevance of S&P to create a basis for conversation. The S&P adepts found it most promising to illustrate the possible consequences of IT security and privacy violations, e.g., through real-life stories. In addition, the participants reported that it is easier to discuss S&P tools with users that require explicit interaction, such as passwords, than those that primarily take place in the background. This could be helped by tools that visualize the influence of different settings on a device or in a program, so that users can try out what influence a certain setting has, for example, on the collection of their data.

Enable Remote Access. Some participants dealt with the problem of helping someone across a distance with a technical S&P problem. While this works easily via screen sharing and remote access on some devices, such solutions are lacking to date, e.g., for mobile devices. Still, this option should be taken with caution, as it tempts S&P adepts to “*just do things quickly themselves*” (P1WS3), although it would be more sustainable to explain the solution to the person seeking help.

4.2.4 Strengthen Capabilities and Opportunities

Improve Expert Knowledge. Practitioners face the challenge of keeping up-to-date with the latest findings from S&P research. This is not only a question of the time required, which could be minimized by a convenient news ticker that summarizes the latest research results, but also of paywalls behind which many research papers are hidden. Although open access publications are already an existing solution to this problem, many publishers require a publication fee from the authors, which cannot always be raised by the institutions concerned. To better assess one’s skills and knowledge gaps,

a “test your knowledge quiz” would also be helpful.

Reward Support-Giving. Ultimately, it should also be worthwhile for the S&P adepts to provide support. One possible way of doing this would be to integrate the support into everyday working life, e.g., by making working time available for this purpose or by recognizing the support as a professional achievement in the context of a scientific or industrial career. A less formalized reward system would be the development of a gamification solution, e.g., P3WS2: “*So, gamification would already like covering ninety percent of all the security experts, because they are all children and want to play games.*” Intrinsic motivation, on the other hand, can come from the social relationship itself, e.g., P3WS2: “*I feel stronger about the need of giving friends and family security advice, because I feel socially obliged to help them to prevent mistakes if I can. If a complete stranger maybe would have the same issue, I wouldn’t bother to go the extra mile.*” An implicit solution would therefore be to emphasize the social aspect of the support, for example, by providing support in a nice setting like a café or by providing some kind of exchange of support in another field where the user is an expert.

4.2.5 Summary of Co-Creation Workshop Findings

It is important to set a constructive, trusting dialogue space to avoid socially awkward situations. As S&P adepts are reluctant to raise the topic themselves for fear of disinterest, media reports, movies, and awareness campaigns could serve as conversation triggers. S&P adepts often do not feel they are in the moral position to give advice, hence, it could be helpful for users to officially delegate privacy and security to S&P adepts in the private context. Encouraging exchange has the potential to counteract the nagging aspects of being asked for advice by referring users to other S&P adepts on topics where they struggle to give good advice, and to other users they have helped before on topics where they are always being asked. Support could be made easier and less time-consuming by facilitating free and easy access to materials, such as flowcharts for knowledge transfer, metaphors, and research results. By rewarding and recognizing support in a professional or social context, motivation can be maintained even in the absence of direct positive responses from users.

5 Discussion

Below, we first recap our findings and then build on them to provide recommendations for S&P adepts who want to support other people as well as the S&P research community.

5.1 Summary of Main Findings

Prior research showed that social triggers have great potential of influencing people’s security and privacy behavior for the better [10, 12–16, 22, 45, 55]. Indeed, people tend to rely on

their social network for tech, but also for S&P advice and support [18, 27, 29, 46, 51, 52, 56]. Few studies have focused on the support-givers' perspective so far. We fill in this gap by adding knowledge about when, how, and why S&P adepts give advice and support to people from their social circle, and how they could be supported in this task.

It has been shown that when giving advice, it is especially important to establish trust, hence social skills are important [38, 39, 62]. We confirm these findings and enhance them by showing that S&P adepts often struggle to comment on or intervene in others' S&P behavior due to fear of negative reactions. Consequently, although they are often *asked by relatives and friends for advice*, they primarily proactively *talk to other S&P-savvy individuals* about S&P-related topics.

A major obstacle to communication between S&P adepts and users is that S&P adepts do not feel in the (*moral*) *position to judge* the behavior of others. If we want to use the potential of S&P adepts to improve the S&P behavior of users in their social environment, we have to find solutions that create a conversation where users ask the S&P adept directly for advice. This could be facilitated, e.g., via the official delegation of S&P in a private context to the S&P adepts. S&P adepts should also be supported in *finding the right tone* for such conversations for which no social norms exist yet, i.e., positive, non-judgmental, and non-moralizing.

It is further important to consider the *users' mental models* in conversations, which may differ from those of the S&P adepts [43, 53, 54]. This can be supported by prepared materials that use metaphors to explain complex issues in a comprehensible way. It could also be helpful for different S&P adepts to *exchange information about strategies and explanations* that have been used successfully – in cases where one is not familiar enough with the topic, one might even *refer the person seeking help to another S&P adept*. To reduce the sometimes daunting effort of S&P adepts, which is not always rewarded by the gratitude of those seeking help, it should be as easy as possible for S&P adepts to *obtain information*. Another option would be to *reward support-giving* in the official context or to *highlight the social aspects* of the process.

5.2 Recommendations: S&P Adepts

We first give four recommendations for S&P adepts based on the results reported in Section 4.2.1 and 4.2.3 considering methods to establish a constructive dialogue space and facilitate knowledge transfer.

Signal Availability. Our participants stated that if someone wants to help people, it must first be clear that this is the appropriate contact person for the topic. To achieve this, we recommend S&P adepts to *be approachable* and address the problem without opening a huge can of worms. This could be realized by transferring the principle of official roles, like data protection officers, from professional to private contexts.

For example, S&P adepts could ask their peers whether they would like to *delegate their S&P* to them. To reach a large number of people seeking help and having word of their expertise spread quickly, S&P adepts should actively *offer their skills in associations and communities*, e.g., by having a special badge on social media profiles, or organizing workshops. S&P adepts should also *openly admit when their knowledge in a particular area is not sufficient* and, if possible, put the person seeking advice in touch with a more suitable S&P adept, e.g., by forwarding a message.

Use Conversation Starters to Talk about S&P. S&P are usually not common topics when talking to people from one's social circle. To find an easy entry into the topic, we recommend S&P adepts to *use media coverage of events*, e.g., legislative changes or the introduction of new technologies such as contact tracing apps as an icebreaker for conversations. They should further *rely on action days*, such as the Safer Internet Day as an occasion and reminder to raise the issue in their social environment. Another suitable conversation starter is to *use popular, unrealistic movies and TV shows* to explain how something really works. To realize this, experts could be supported by publicly available online collections of movie clips for different topics that they could either show their peers or share with them. In the interests of sustainability, if an S&P adept is asked for help with a technical problem, one should also *not just make all the settings themselves*, but explain to the person seeking help what they are doing and why.

Stay Positive. Talking about others' S&P behavior can be socially difficult, because one does not want to criticize the other person as stated in both studies. To address this issue, we recommend S&P adepts to stick to *positive, non-judgmental language*, *do not moralize*, and give users *time to reflect*, i.e., by revisiting the topic after a while. S&P adepts could be supported here by publicly available informative materials and educational videos that help them strengthening their communication skills. Since not all S&P adepts have the capabilities for this, it would also make sense to integrate communication training into curricula for technical subjects.

Establish a Common Ground. Another challenge is to find a common basis for discussion. For this purpose, we recommend S&P adepts to be aware that people asking for help might have *erroneous mental models*. To address this, S&P adepts should use *metaphors* to explain the technical background of solutions, such as E2EE. To lay the groundwork for this, security curricula should include human factors to strengthen understanding of the lay user perspective. Also, they should use *consistent (technical) terms* in the long term. As stated above, S&P adepts could be supported by informative materials or educated within specific workshops. Such materials would ideally be standardized and use a common terminology. Further, awareness for existing materials, such as those offered by the National Cybersecurity Alliance

(U.S. [4]), the National Cyber Security Centre (U.K., [9]), or the BSI (Germany, [31]) should be raised, e.g., by using professional networks or mailing-lists.

Considering the four recommendations above, there are several obstacles and challenges. Since security experts might *not agree on what the most important, useful tips for non-tech-savvy users are* [57], there is a risk that they will not even recognize where their knowledge is insufficient or where an acute need for action is. To address this, an exchange between S&P adepts is also important with regard to sensible security and privacy advice (see Section 5.3). Further, S&P adepts must ensure that they *do not refer help seekers to malicious actors* who, for example, want to gain access to systems or passwords. Since S&P adepts first have to find the necessary *time* and *motivation*, the following recommendations refer to how the research community can support them in doing so.

5.3 Recommendations: Research Community

We further propose to implement the following measures in the S&P research community to strengthen the capabilities and opportunities for S&P adepts and harness the potential of exchange (see Section 4.2.4 and 4.2.2):

Reward Support-Giving. S&P adepts' desire to support others often conflicts with other commitments and goals. To address this, we recommend that commitment to user support should somehow be *recognized as a career achievement*, e.g., by awarding specific certifications or social media badges. To further enhance the S&P adepts' motivation, *gamification solutions* should not only be developed for the S&P behavior of users, but also for the support giving of S&P adepts. For this, one could also introduce a *rating system for advice quality* similar to online platforms like StackOverflow or as part of existing social networks, which strengthens user trust in the S&P adept and the adepts' trust in their own abilities.

Facilitate Access to Information. In order to facilitate access to research results also for S&P adepts from industry and researchers from institutions with less funding, *open access* should be specifically promoted since participants in our study who were practitioners voiced interest in research papers that is hindered by closed access. This could be realized by primarily applying for projects with funding for open access fees or by preferentially publishing at conferences and journals that publish the publications free of charge under an open-access license. Furthermore, the research community could offer a *newsletter* summarizing the most important recent research findings. Another option is strengthening the link between practitioners and researchers, for instance, by offering practitioner tracks at scientific conferences and publishing in practitioner magazines.

Establish a Peer-System. Furthermore, to increase outreach and relieve S&P adepts, a *peer system* could be introduced

by passing on knowledge to the next generation, similar to mentoring programs. For this purpose, people seeking advice could also be referred directly to other users who have already been helped on the same topic. There are several ways to realize this: a standalone online platform, as part of a social network, or within organizations and schools.

Create a Platform for Professional Exchange. A platform should be created that enables the *exchange between S&P adepts* on this topic, e.g., in the context of workshops or as a Slack channel. This could also serve as a discussion space to gain consensus on what is effective and actionable S&P advice.

6 Conclusion and Future Work

We add to existing work on social support in the S&P context by investigating how and under which circumstances S&P adepts support people in their private social environment, the challenges they face and ways to overcome them. For this, we first analyze the status quo by conducting in-depth interviews with 13 S&P adepts, and then explore options to assist S&P adepts in their efforts to help others in three co-creation workshops with 11 S&P adepts. We find that S&P adepts often struggle finding the right tone in conversations with lay users, partly because they do not see themselves in a moral position to give advice. Once contact is established, another challenge is to find a shared language. Since lay users often have different mental models than S&P adepts, it can be helpful to use metaphors for this purpose.

Some of the findings from our exploratory studies need to be confirmed and analyzed in more detail, such as what obstacles S&P adepts face in improving others' behavior. The effectiveness of the recommendations proposed by us should be investigated in field studies. For this, the introduction of a peer system and a platform for professional exchange would be a good idea. Another possible next step is the creation and evaluation of guidelines and training for the social aspects of conversations. Focusing on risks of S&P failures seems promising for emphasizing the relevance of S&P. Further research is needed to understand how such risks should be communicated [1, 21, 42]. Since it is easier to communicate about S&P tools that are observable, research should identify solutions that improve the visibility of S&P issues, such as privacy icons [19, 36]. Finally, since the lack of motivation is a main obstacle for providing S&P advice, future work should identify and validate techniques that motivate S&P adepts.

Acknowledgments

This work has been co-funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, grant number 251805230/GRK 2050) and by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research, Science and the Arts within

their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

- [1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Secrets and likes: the drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, 30(4):736–758, 2020.
- [2] Alessandro Acquisti, Leslie K. John, and George Loewenstein. The impact of relative standards on the propensity to disclose. *Journal of Marketing Research*, 49(2):160–174, 2012.
- [3] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Commun. ACM*, 42(12):40–46, dec 1999.
- [4] National Cybersecurity Alliance. Stay safe online, 2022. <https://staysafeonline.org> (Accessed 30-May-2022).
- [5] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
- [6] Virginia Braun and Victoria Clarke. *Successful qualitative research: A practical guide for beginners*. SAGE Publications, 2013.
- [7] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No one can hack my mind revisiting a study on expert and Non-Expert security practices and advice. In *Fifteenth Symposium on Usable Privacy and Security*, SOUPS 2019, pages 117–136, Santa Clara, CA, August 2019. USENIX Association.
- [8] Rory Cellan-Jones and Leo Kelion. Coronavirus: The great contact-tracing apps mystery, 2021. <https://www.bbc.com/news/technology-53485569> (Accessed 16-February-2022).
- [9] National Cyber Security Centre. Information for individuals & families, 2022. <https://www.ncsc.gov.uk/section/information-for/individuals-families> (Accessed 30-May-2022).
- [10] Lynne M. Coventry, Debora Jeske, John M. Blythe, James Turland, and Pam Briggs. Personality and social framing in privacy decision-making: A study on cookie acceptance. *Frontiers in Psychology*, 7, 2016.
- [11] Rikke Friis Dam and Teo Yu Siang. Define and frame your design challenge by creating your point of view and ask "how might we", 2021. <https://www.interaction-design.org/literature/article/define-and-frame-your-design-challenge-by-creating-your-point-of-view-and-ask-how-might-we> (Accessed 02-February-2022).
- [12] Sauvik Das, Laura A. Dabbish, and Jason I. Hong. A typology of perceived triggers for End-User security and privacy behaviors. In *Fifteenth Symposium on Usable Privacy and Security*, SOUPS 2019, pages 97–115, Santa Clara, CA, August 2019. USENIX Association.
- [13] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A. Dabbish, and Jason I. Hong. The effect of social influence on security sensitivity. In *10th Symposium On Usable Privacy and Security*, SOUPS 2014, pages 143–157, Menlo Park, CA, July 2014. USENIX Association.
- [14] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 739—749, New York, NY, USA, 2014. Association for Computing Machinery.
- [15] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. The role of social influence in security feature adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1416—1426, New York, NY, USA, 2015. Association for Computing Machinery.
- [16] Sauvik Das, Joanne Lo, Laura Dabbish, and Jason I. Hong. Breaking! A typology of security and privacy news and how it's shared. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1—12, New York, NY, USA, 2018. Association for Computing Machinery.
- [17] Steve De Shazer, Yvonne Dolan, Harry Korman, Terry Trepper, Eric McCollum, and Insoo Kim Berg. *More than miracles: The state of the art of solution-focused brief therapy*. Routledge, 2021.
- [18] Paul Dourish, Rebecca E. Grinter, Jessica Delgado de la Flor, and Melissa Joseph. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8:391–401, 2004.
- [19] Serge Egelman, Raghudeep Kannavara, and Richard Chow. Is this thing on? Crowdsourcing privacy indicators for ubiquitous sensing platforms. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1669—1678, New York, NY, USA, 2015. Association for Computing Machinery.

- [20] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15*, pages 2873—2882, New York, NY, USA, 2015. Association for Computing Machinery.
- [21] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an iot privacy and security label? In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 447–464, 2020.
- [22] Pardis Emami Naeini, Martin Degeling, Lujo Bauer, Richard Chow, Lorrie Faith Cranor, Mohammad Reza Haghghat, and Heather Patterson. The influence of friends and experts on privacy decision making in iot scenarios. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), Nov 2018.
- [23] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth Symposium on Usable Privacy and Security, SOUPS 2016*, pages 59–75, Denver, CO, June 2016. USENIX Association.
- [24] Cori Faklaris, Laura A. Dabbish, and Jason I. Hong. A Self-Report measure of End-User security attitudes (SA-6). In *Fifteenth Symposium on Usable Privacy and Security, SOUPS 2019*, pages 61–77, Santa Clara, CA, August 2019. USENIX Association.
- [25] Better Internet for Kids. Safer internet day. <https://www.saferinternetday.org/> (Accessed 16-February-2022).
- [26] Inria National Institute for Research in Digital Science and Technology. Amiunique. <https://amiunique.org/> (Accessed 16-February-2022).
- [27] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. Do or do not, there is no try: User engagement may not improve security outcomes. In *Twelfth Symposium on Usable Privacy and Security, SOUPS 2016*, pages 97–111, Denver, CO, June 2016. USENIX Association.
- [28] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human-Computer Interaction*, 35(6):456–467, 2019.
- [29] Alisa Frik, Leysan Nurgalieva, Julia Bernd, Joyce Lee, Florian Schaub, and Serge Egelman. Privacy and security threat models and mitigation strategies of older adults. In *Fifteenth Symposium on Usable Privacy and Security, SOUPS 2019*, pages 21–40, Santa Clara, CA, August 2019. USENIX Association.
- [30] Kelsey R. Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L. Mazurek. The effect of entertainment media on mental models of computer security. In *Fifteenth Symposium on Usable Privacy and Security, SOUPS 2019*, pages 79–95, Santa Clara, CA, August 2019. USENIX Association.
- [31] Bundesamt für Sicherheit in der Informationstechnik. Digitaler Verbraucherschutz, 2022. https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/verbraucherinnen-und-verbraucher_node.html (Accessed 30-May-2022).
- [32] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. Secrecy, flagging, and paranoia: Adoption criteria in encrypted email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06*, pages 591–600, New York, NY, USA, 2006. Association for Computing Machinery.
- [33] Nina Gerber, Benjamin Reinheimer, and Melanie Volkmann. Investigating people’s privacy risk perception. *Proceedings on Privacy Enhancing Technologies*, 2019(3):267–288, 2019.
- [34] Greenpeace Media GmbH. Greenpeace magazin. warenaus. <https://warenhaus.greenpeace-magazin.de/> (Accessed 16-February-2022).
- [35] Thomas Gross. Validity and reliability of the scale internet users’ information privacy concerns (IUIPC). *Proceedings on Privacy Enhancing Technologies*, 2021:235–258, 2021.
- [36] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. Toggles, dollar signs, and triangles: How to (in)effectively convey privacy choices with icons and link texts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, New York, NY, USA, 2021. Association for Computing Machinery.
- [37] Julie Haney and Wayne Lutters. Cybersecurity advocates: Discovering the characteristics and skills for an emergent role. *Information & Computer Security*, 29, 2021.
- [38] Julie Haney, Wayne Lutters, and Jody Jacobs. Cybersecurity advocates: Force multipliers in security behavior change. *IEEE Security Privacy*, 19(4):54–59, 2021.

- [39] Julie M. Haney and Wayne G. Lutters. "it's Scary... It's Confusing... It's dull": How cybersecurity advocates overcome negative perceptions of security. In *Fourteenth Symposium on Usable Privacy and Security*, SOUPS 2018, pages 411–425, Baltimore, MD, August 2018. USENIX Association.
- [40] Alex Hern. Whatsapp to try again to change privacy policy in mid-may, 2021. <https://www.theguardian.com/technology/2021/feb/22/whatsapp-to-try-again-to-change-privacy-policy-in-mid-may> (Accessed 16-February-2022).
- [41] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...No one can hack my Mind": Comparing expert and Non-Expert security practices. In *Eleventh Symposium On Usable Privacy and Security*, SOUPS 2015, pages 327–346, Ottawa, July 2015. USENIX Association.
- [42] Timo Jakobi, Sameer Patil, Dave Randall, Gunnar Stevens, and Volker Wulf. It Is About What They Could Do with the Data: A User Perspective on Privacy in Smart Metering. *ACM Trans. Comput.-Hum. Interact.*, 26(1):2:1–2:44, 2019.
- [43] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "My data just goes Everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security*, SOUPS 2015, pages 39–52, Ottawa, July 2015. USENIX Association.
- [44] Joonas Koivumaa. *Sprint: How to Solve Big Problems and Test New Ideas in just Five Days*. Lapin ammattikoulu, 2017.
- [45] Isadora Krsek, Kimi V Wenzel, Sauvik Das, Jason I. Hong, and Laura A. Dabbish. To self-persuade or be persuaded: Examining interventions for users' privacy setting selection. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.
- [46] Amanda Lenhart, Mary Madden, Sandra Cortesi, Urs Gasser, and Aaron Smith. Where teens seek online privacy advice, 2013. <https://www.pewresearch.org/internet/2013/08/15/where-teens-seek-online-privacy-advice/> (Accessed 17-January-2022).
- [47] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. How developers talk about personal data and what it means for user privacy: A case study of a developer forum on reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), jan 2021.
- [48] Heather Richter Lipford and Mary Ellen Zurko. Someone to watch over me. In *Proceedings of the 2012 New Security Paradigms Workshop*, NSPW '12, pages 67–76, New York, NY, USA, 2012. Association for Computing Machinery.
- [49] Bojana Lobe, David Morgan, and Kim A. Hoffman. Qualitative data collection in an era of social distancing. *International Journal of Qualitative Methods*, 19:1609406920937875, 2020.
- [50] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet users' information privacy concerns (UIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.
- [51] Norbert Nthala and Ivan Flechais. Informal support networks: an investigation into home data security practices. In *Fourteenth Symposium on Usable Privacy and Security*, SOUPS 2018, pages 63–82, Baltimore, MD, August 2018. USENIX Association.
- [52] Erika Shehan Poole, Marshini Chetty, Tom Morgan, Rebecca E. Grinter, and W. Keith Edwards. Computer help at home: Methods and motivations for informal technical support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 739 — 748, New York, NY, USA, 2009. Association for Computing Machinery.
- [53] Emilee Rader and Rick Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, 12 2015.
- [54] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS 2012, New York, NY, USA, 2012. Association for Computing Machinery.
- [55] Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. "It's easier than causing confrontation": Sanctioning strategies to maintain social norms and privacy on social media. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020.
- [56] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How i learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 666—677, New York, NY, USA, 2016. Association for Computing Machinery.
- [57] Robert W. Reeder, Iulia Ion, and Sunny Consolvo. 152 simple steps to stay safe online: Security advice for non-tech-savvy users. *IEEE Security Privacy*, 15(5):55–64, 2017.

- [58] Curedale Robert. *Design Thinking: Process and Methods Manual*. Design Community College Inc, 2013.
- [59] Adam Satariano. U.s. news outlets block european readers over new privacy rules, 2021. <https://www.nytimes.com/2018/05/25/business/media/europe-privacy-gdpr-us.html> (Accessed 16-February-2022).
- [60] Luis Angel Saúl, M. Angeles López-González, Alexis Moreno-Pulido, Sergi Corbella, Victoria Compañ, and Guillem Feixas. Bibliometric review of the repertory grid technique: 1998–2007. *Journal of Constructivist Psychology*, 25(2):112–131, 2012.
- [61] Leonie Schaewitz, David Lakotta, M. Angela Sasse, and Nikol Rummel. Peeking into the black box: Towards understanding user understanding of E2EE. In *European Symposium on Usable Security, EuroUSEC '21*, pages 129—140, New York, NY, USA, 2021. Association for Computing Machinery.
- [62] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Privacy champions in software teams: Understanding their motivations, strategies, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, New York, NY, USA, 2021. Association for Computing Machinery.
- [63] Mohammad Tahaei, Tianshi Li, and Kami Vaniea. Understanding privacy-related advice on Stack Overflow. *Proceedings on Privacy Enhancing Technologies*, 1:18, 2022.
- [64] National Today. World password day. <https://nationaltoday.com/world-password-day/> (Accessed 16-February-2022).
- [65] Sabine Trepte, Doris Teutsch, Philipp K. Masur, Carolin Eicher, Mona Fischer, Alisa Hennhöfer, and Fabienne Lind. *Do People Know About Privacy and Data Protection Strategies? Towards the “Online Privacy Literacy Scale” (OPLIS)*, pages 333–365. Springer Netherlands, Dordrecht, 2015.
- [66] Ramon Vullings, Godelieve Spaas, and Igor Byttebier. *Creativity Today*. BIS Publishers, 2009.

A Study I: Interview Study

A.1 Interview Demographics and Screening

In this section, we provide detailed demographics from the interviews (Table 1) and screening data (Table 2).

Table 1: Demographics of the interview sample.

ID	age	gender	occupation	highest education
P1	29	f	PhD Student	Master Degree
P2	43	m	Computer Scientist	Diploma
P3	34	m	Researcher at University	PhD or higher
P4	56	m	Production Plant Manager	Master Degree
P5	25	m	Student	Bachelor Degree
P6	23	m	Student	Bachelor Degree
P7	22	m	Intern in a company & student	Bachelor Degree
P8	53	m	Head of IT	Apprenticeship
P9	21	f	Student	High School Diploma
P10	22	m	Student	High School Diploma
P11	25	m	Development Engineer	Master Degree
P12	48	m	System Administrator	High School Diploma
P13	21	m	Student	School Student

Table 2: Screening of the interview sample.

	median	minimum	maximum	percentile		
				25	50	75
SEBIS_Device Securement	4.5000	4.00	5.00	4.2500	4.5000	4.7500
SEBIS_Password Generation	4.2500	3.00	5.00	3.5000	4.2500	4.3750
SEBIS_Proactive Awareness	3.6000	2.80	4.60	3.4000	3.6000	4.3000
SEBIS_Updating	4.3333	3.00	5.00	3.8333	4.3333	4.6667
SA6	3.8333	3.00	4.83	3.3333	3.8333	4.2500
ATI	5.1111	3.00	5.78	4.7222	5.1111	5.3333
IUIPC8_Control	6.5000	3.50	7.00	5.7500	6.5000	6.7500
IUIPC8_Awareness	7.0000	4.00	7.00	6.2500	7.0000	7.0000
IUIPC8_Collection	6.2500	3.75	7.00	5.0000	6.2500	6.8750
OPLIS_Knowledge	4.0000	2.00	5.00	3.5000	4.0000	5.0000
OPLIS_Technical Aspects	5.0000	4.00	5.00	5.0000	5.0000	5.0000
OPLIS_Law	3.0000	0.00	4.00	2.0000	3.0000	4.0000
OPLIS_Protection	4.0000	1.00	5.00	2.0000	4.0000	4.0000

Note: Cut off scores were SEBIS (average of all scales): 3.7, SA6: 3, ATI: 3, IUIPC-8 (average of all scales): 3.7, OPLIS (average of all scales): 2.5. Please note that these are the absolute lowest limits, but the values of most participants were significantly higher. We also considered the whole picture and made sure that someone scoring low on one scale (e.g., privacy concerns) scored considerably higher on other scales, e.g., security behavioral intention.

A.2 Interview Guide

- Welcoming the participant
- Security and Privacy Behavior
 - What do you do to protect your S&P in everyday life?
 - Do you share this behavior in interactions with others? Do others recognize your S&P in everyday life?
 - How do others react to it?
 - How do you feel when others respond to your behavior?
 - Is this a topic of conversation? (E.g., refraining from using social networks, or use of certain messengers)
 - Have there ever been situations where others reacted with surprise to your S&P behavior?
 - Have there ever been situations where you felt uncomfortable acting according to your S&P ideas?
 - * How did you resolve this?
 - * What did you take away from this for future situations?

- In conversations, do you generally hold an S&P opinion, e.g., about using certain products and services?
 - * (If yes:) How do you feel about this?
 - * (If not:) Why not?
- Interference
 - Do you ever observe insecure or privacy-unfriendly behavior in others?
 - * Could you give an example?
 - * What are your thoughts and feelings about it?
 - * Do you intervene in the other person's behavior?
 - (If yes:) How and why?
 - (If no:) Why not?
 - * Do you potentially fear of coming across as arrogant?
 - * Do you potentially fear of coming across as preachy?
- Advice
 - Do others ask you for advice about S&P?
 - * Who?
 - * About what specifically?
 - * How do you respond to that?
 - * Can you give examples for what advice you give?
- Responsibility
 - Do you feel responsible for the S&P of others?
 - * Of whom?
 - * Why?
 - * How does this affect your behavior?
 - Do you engage in "digital housekeeping" with relatives or friends?
 - Do you share news about changes (e.g., new privacy regulations, E2E encryption), data breaches, or security incidents?
 - * (If yes:) With whom and why?
 - * (If not:) Why not?
- Bad experiences
 - Have you ever had a bad experience giving S&P advice to someone?
 - Have you ever had a bad experience when intervening without being asked?
 - Are you afraid that intervening might somehow be socially inappropriate/awkward?
 - Are you afraid that it might strain social relationships if
 - * People are annoyed by interventions?
 - * You notice that people don't follow your advice?
 - Are you afraid that helping might be too much work/you will be asked all the time in the future?
 - Are you afraid that you don't know some things and that this will damage your reputation as an expert?
- Perception
 - Would you say that others think you are an S&P expert?
 - Why do others think you are an S&P expert?
 - Generally regarding own behavior or when interfering/advising others ...
 - * ... are you sometimes afraid of coming across as paranoid/as a "tin foil hat wearer"?

- * ... are you sometimes afraid of coming across as a tech nerd?
 - * Would you say that S&P expertise has anything to do with gender? Is being an S&P expert different for women than for men?
- Repertory Grid: With which characteristics would you describe someone who ...
- * ... behaves too insecurely?
 - * ... behaves in just the right secure way?
 - * ... behaves too securely?
 - * ... behaves too privacy-unfriendly?
 - * ... behaves in exactly the right privacy-friendly way?
 - * ... behaves too privacy-friendly?
- End and reimbursement

A.3 Codebook

The codebook is available at https://www.arbing.psychologie.tu-darmstadt.de/media/ag_arbeits_und_ingenieurpsychologie/responsive_design/forschungsergebnisse_1/TheNerdFactor_Codebooks.pdf.

B Study II: Co-Creation

B.1 Workshop Demographics and Screening

In this section, we provide detailed demographics from workshops (Table 3) and screening data (Table 4).

Table 3: Demographics of the co-creation sample. Please note that P1WS1 refers to Participant 1 from the first workshop, P1WS2 to Participant 1 from the second workshop etc.

ID	age	gender	occupation	highest education
P1WS1	31–35	m	Employed Full-time & Privacy Officer	Bachelor Degree
P2WS1	26–30	f	Scientific Employee	Master Degree
P3WS1	31–35	m	Post-doc Researcher	PhD or higher
P4WS1	26–30	m	Usable Security Researcher (PhD)	Master Degree
P1WS2	31–35	f	Post-doc Researcher	PhD or higher
P2WS2	26–30	m	Software Developer	Master Degree
P3WS2	31–35	m	Post-doc Researcher	PhD or higher
P1WS3	26–30	m	Doctoral Candidate	Master Degree
P2WS3	31–35	f	Post-doc Researcher	PhD or higher
P3WS3	36–40	m	Professor, computer science	PhD or higher
P4WS3	26–30	f	Usable Security Researcher (PhD)	Master Degree

Table 4: Screening of the workshop sample.

	median	minimum	maximum	percentile		
				25	50	75
SA6	3.6667	3.00	4.67	3.1667	3.6667	4.3333
ATI	4.5556	2.89	5.78	4.1111	4.5556	5.2222
IUIPC8_Control	6.0000	5.00	7.00	5.5000	6.0000	7.0000
IUIPC8_Awareness	7.0000	6.00	7.00	6.5000	7.0000	7.0000
IUIPC8_Collection	6.5000	3.25	7.00	5.5000	6.5000	6.7500
OPLIS_Technical Aspects	5.0000	4.00	5.00	4.0000	5.0000	5.0000

Note: Cut off scores were SA6: 3, ATI: 2.7, IUIPC-8 (average of all scales): 3.7, OPLIS: 4. Please note that these are the absolute lowest limits, but the values of most participants were significantly higher. We also considered the whole picture and made sure that someone scoring low on one scale (e.g., privacy concerns) scored considerably higher on other scales, e.g., knowledge about technical privacy aspects.

B.2 Workshop Guide

- Welcome (3 min)
- Icebreaker (7 min):
 - paint your current mood on Mural (paint & introduce)
 - meanwhile small talk about social S&P situations
- Introduction to the topic (2 min)
- Brainstorming: Facilitators and obstacles, using the miracle question
- Formulate "How Might We" question (2 min)
- Develop solutions: 5-3-4
 - everyone creates 3 ideas on sticky notes (writing and/or drawing)
 - after 3 min: go clockwise, everyone creates 3 more ideas (inspired by existing or new), repeat until rotation is complete
 - Remember: The target group is you! You should develop solutions that help you to support others.
- Short presentation of your own thoughts (1 min each), everyone can ask comprehension questions until all ideas are reasonably clear
- Decision: Dotmocracy with How-Now-Wow matrix
- Discussion: everyone explains what they think is good and why and what is difficult and why, then open discussion where wow ideas can be developed into solutions
- Wrap up, farewell, payment

B.3 Codebook

The codebook is available at https://www.arbing.psychologie.tu-darmstadt.de/media/ag_arbeits_und_ingenieurpsychologie/responsive_design/forschungsergebnisse_1/TheNerdFactor_Codebooks.pdf.

“I don’t know why I check this ...” – Investigating Expert Users’ Strategies to Detect Email Signature Spoofing Attacks

Peter Mayer¹, Damian Poddebniak², Konstantin Fischer³, Marcus Brinkmann³, Juraj Somorovsky⁴,
Angela Sasse³, Sebastian Schinzel², and Melanie Volkamer¹

¹SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology

²Münster University of Applied Sciences

³Ruhr University Bochum

⁴Paderborn University

Abstract

OpenPGP is one of the two major standards for end-to-end email security. Several studies showed that serious usability issues exist with tools implementing this standard. However, a widespread assumption is that expert users can handle these tools and detect signature spoofing attacks. We present a user study investigating expert users’ strategies to detect signature spoofing attacks in Thunderbird. We observed 25 expert users while they classified eight emails as either having a legitimate signature or not. Studying expert users explicitly gives us an upper bound of attack detection rates of all users dealing with PGP signatures. 52% of participants fell for at least one out of four signature spoofing attacks. Overall, participants did not have an established strategy for evaluating email signature legitimacy. We observed our participants apply 23 different types of checks when inspecting signed emails, but only 8 of these checks tended to be useful in identifying the spoofed or invalid signatures. In performing their checks, participants were frequently startled, confused, or annoyed with the user interface, which they found supported them little. All these results paint a clear picture: Even expert users struggle to verify email signatures, usability issues in email security are not limited to novice users, and developers may need proper guidance on implementing email signature GUIs correctly.

1 Introduction

Signatures can provide end-to-end protection of the authenticity and integrity of email messages. Yet, Müller et al. [19] showed that verifying email signatures and displaying the

result of the verification in a graphical user interface (GUI) is very challenging. Among others, they described *weak signature forgeries* that mimic GUI elements of a valid signature closely, but not perfectly. Upon close inspection, the user can detect that the GUI elements are fake. For example, they used HTML and CSS to include a green signature validation banner, but the fake banner was positioned incorrectly and did not provide the interactivity of the original. They classify the forgery as *weak* because they argue that vigilant users can detect it. In this work, we examine a subset of their attacks and attempt to answer the following question: *Which strategies do users employ to detect email signature spoofing, and how susceptible do these strategies leave them to these attacks?*

We answer this question by interviewing *expert users* of Thunderbird, who frequently use signatures and are familiar with public-key cryptography, digital signatures, and their email clients. We conducted a user study with participants drawn from attendees of FOSDEM 2020 – a European open source developer conference that also hosted an OpenPGP key signing party. Two pre-studies at the Chaos Communication Camp and Congress in 2019 informed the design of this study.

Our study participants were asked to use Thunderbird and its OpenPGP-plugin Enigmail to inspect eight semantically identical emails. Four of these emails contained a valid signature, and four contained an invalid signature. The invalid signatures were forgeries similar to the weak forgeries in [19]. The participants had to decide whether a signature was legitimate or not. As they were expert users, this gives an upper bound on how well users can detect such attacks.

Our results indicate that even expert users have no effective strategies to detect email signature spoofing attacks, leading to 52% of our participants failing to detect at least one out of four forged email signatures. Our participants’ checks were diverse: They applied 23 different checks when inspecting the attack emails. Of these checks, only 8 tended to be helpful to identify spoofed signatures. Also, the GUI often startled or perplexed the participants.

To counter the lack of effective user strategies to detect email signature spoofing attacks and the resulting suscepti-

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

bility to these attacks, email clients should offer guidance to users so they perform the most effective checks and are deterred from making ineffective ones. The GUI should make affordances [21] immediately apparent. We believe a way forward is to follow the results pertaining to supporting developers [18] and offer guidance to developers of email clients in creating GUIs that actively support users in detecting attacks. The core contributions of our research are:

- We give an overview of the checks expert users apply in their strategies to verify email signatures (section 5.1) and assess the usefulness of these checks to detect weak forgery attacks (section 5.2).
- We present the first upper bound baseline regarding expert users' performance in email signature spoofing detection (section 5.3).
- We present an overview of usability issues and how these prevent effective detection of spoofed signatures and instead increase user risk and uncertainty (section 5.5).
- We make the study materials, research artifacts, and evaluation tools available as open-source.¹

2 Background on Email and OpenPGP

Emails [5] consist of two parts, a header and a body, where the header is a list of (name, value) fields and the body is ASCII text. The header contains the sender address, recipient address, and other metadata, while the body contains the actual content of the message. With MIME [9], emails internally become a tree-structure that can contain not only text but also other data types such as images, attachments, and digital signatures as defined in the OpenPGP standard [3, 7].

Verifying Email Signatures When rendering an email, the client has to clearly communicate each signature's validity, origin, and scope through the GUI to the user. This can be very difficult. Most email clients do not attempt to handle all signed parts at any layer but instead support only a single signed element, omit the scope of the signature, omit information about the signer, or otherwise simplify the process. Such clients require additional security checks.

OpenPGP Signer vs. Email Sender OpenPGP does not require that the signer's identity is identical to the sender in the email header. A secure email client should either enforce that the sender and signer have the same email address, in which case they can omit the signer identity from the signature verification result, or include the signer identity in the result, in which case the user is responsible for checking it.

OpenPGP Key Management Any digital signature could have been generated by anyone at first sight. To make signatures useful in the context of email, the signing key has to be bound to a user identified by an email address.

¹<https://github.com/SECUSO/email-signature-expert-study>

Early OpenPGP implementations favored decentralized key management requiring manual validation, either directly or through the Web of Trust. Today, many users expect automatic key validation, and the most popular solution is the centralized key server *keys.openpgp.org* with 290k keys (Feb. 2022), where the email address is validated by sending a registration link. In addition, various domain-based proposals exist, such as DNS TXT records, DNSSEC/DANE [40], or HTTPS via the Web Key Directory (WKD).

3 Related Work

Human Aspects of Secure Email In their seminal work in 1999, Whitten and Tygar [38] evaluated the usability of PGP 5.0 in the Eudora email client with a cognitive walk-through and user test (12 novice users). They demonstrated several serious usability issues. Follow-up works by other authors have studied PGP 9.0 in Outlook Express (pilot study with six novice users) [32], PGP support in Mailvelope (20 student participants) [28], and PGP support in Outlook 2016, Thunderbird and Maildroid (12 participants) [23], as well as with Enigmail and Mailvelope (52 non-technical participants) [16]. Due to these usability issues, it was found that while users want to use secure email [25] and find it important [22], adoption of email standards like OpenPGP and S/MIME is low.

Besides usability issues, the key management is often identified as a reason for the low adoption [22, 35]. One proposed mitigation of these key management issues is the automation of the related tasks [2, 11, 27]. For example, Garfinkel et al. [10] propose to accept all keys and only notify the user if the key differs from a previously used one. However, automation can have negative effects, as Ruoti et al. [29] note. Another solution proposed by Roth et al. [24] is rather to use in-person verification than trust certificate authorities. Lerner et al. [15] proposed combining this social approach with automation using Keybase, a service allowing users to link their public keys and social media accounts. Their proposal "Confidante" was well received by the study participants and reduced the time spent on key management while reducing the number of critical errors. Unfortunately, none of these proposals have been adopted, which means that the key management issues remain. The focus of our paper, however, is on the potential usability issues for expert users.

Several researchers have investigated how the usability issues can be addressed. Tolsdorf and Lo Iacocno [37] proposed to use persuasive design to improve the design of secure email GUIs. Ruoti et al. [26] found several ways to increase understanding of email encryption: a short delay and dialogue when encrypting or decrypting emails, a dedicated composer for encrypted emails (separate from the composer for unencrypted emails), and tutorials. Lausch et al. [14] analyzed the usability of novel security indicators in email clients and identified envelopes, torn envelopes, and postcards as promising

candidates for future designs. However, text indicators might be enough: Stransky et al. [34] found in their comparison of several security indicators that simple text labels such as “encrypted” are as effective. Furthermore, their results indicate that icons can even lead to negative perceptions of the users. Gaw et al. [12] give an example of how a bad design can lead to annoyance. They found that the practice of connecting the encryption status of an email to the urgency status of that email led users to avoid encryption for regular emails.

One may argue that McGregor et al. [17] already studied expert users in the context of secure email communication. However, their focus was on encryption, while our focus is on signed emails and expert users’ ability to detect email signature spoofing attacks. In their investigation of the tools used by journalists in the year-long “Panama Papers” project, they found that the tools used were perceived as highly usable and useful by the involved journalists, allowing them to meet confidentiality goals for the entire duration of the project.

Email Signature Spoofing Research indicates that signatures might be at least as desirable for users as encryption. Reuter et al. [22] found that the primary concern in terms of secure email is protection against others impersonating a trustworthy sender. The authenticity provided by digital signatures can fulfill exactly this role.

Müller et al. [19] describe three classes of weak signature spoofing attacks that can be detected by users of email clients: (1) *UI Attacks* are directed at the presentation of signature validation results in the email client. The attacker crafts an email containing an image that mimics a legitimate signature validation. (2) *ID Attacks* are directed at the potential mismatch between the sender and the signer of an email. An attacker creates a legitimately signed email with the attacker’s key and then manipulates the email headers such that the signature looks like it was made by the sender instead. (3) *MIME Attacks* are directed at the complex MIME processing in email clients. The attacker gets a legitimately signed email from the victim and then constructs a new email that shows the same signature for a different content.

Other attacks on email signatures include covert content attacks [20], where an attacker attempts to acquire legitimate signatures unbeknownst to the signer, and spoofing attacks at the transport level for DKIM signatures [4]. However, these two types are not in the scope of our research.

4 Methodology

This research aims to investigate the strategies of expert users when deciding whether a signature is legitimate (i.e., a valid signature from the correct sender) and which individual checks these strategies comprise.

4.1 Research Questions

Our investigation is guided by five research questions:

RQ1 [Checks & Strategies]

- (a) Which checks do experts of OpenPGP email signatures in Thunderbird apply to discern legitimate from illegitimate signatures?
- (b) How does the participants’ overall strategy for the application of the checks look like?

RQ2 [Usefulness of Checks]

- (a) Which checks helped participants to correctly discern legitimate from illegitimate email signatures?
- (b) Which checks did not help participants to correctly discern legitimate from illegitimate email signatures?
- (c) Which checks used by the participants pushed participants to incorrect decisions when discerning legitimate from illegitimate email signatures?

RQ3 [Performance of Participants]

Were experts successful in detecting attacks, i.e., discerning legitimate from illegitimate email signatures?

RQ4 [Predictability of Success]

- (a) Is it possible to predict the outcome of discerning legitimate from illegitimate email signatures based on the outcome of the SA-6 scale?
- (b) Is it possible to predict the outcome of discerning legitimate from illegitimate email signatures based on the outcome of the RSeBIS scale?
- (c) Is it possible to predict the outcome of discerning legitimate from illegitimate email signatures based on self-reported expertise with email signatures?
- (d) Is it possible to predict the outcome of discerning legitimate from illegitimate email signatures based on the self-reported frequency of OpenPGP usage in Thunderbird?

RQ5 [User Perceptions]

How did participants perceive the process of investigating the legitimacy of message signatures?

4.2 Study Design

Two pre-studies informed the design of our main study.

4.2.1 Ethics

While our institutions did not mandate ethical approval for this study, our study fulfills all requirements of our institutions regarding studies with humans. The study procedure and data collection was approved by the data protection authority, also ensuring data minimization. The study had an informed consent form (see appendix A.1), explaining how to withdraw from the study and including a privacy policy. Participants received a debriefing, where the attacks were explained and any remaining concerns or questions of the participants were addressed. Additionally, we provided our contact data to participants in case of further questions or concerns. The video and audio recordings, as well as the questionnaire responses,

were encrypted when stored or in transit. Only researchers approved by our data protection authority had access.

4.2.2 First Pre-Study

The goal of the first pre-study was to identify the most widely used email clients and signature standards among our target participant group of expert users. This study was held at the summer camp 2019 of the Chaos Computer Club (CCC) in Germany, attended mainly by IT security enthusiasts and internet activists. Participants were recruited by approaching them on the campground and asked on the spot how often they use OpenPGP and S/MIME signatures and which email clients they use. We collected data from 23 participants. The results showed that Thunderbird with Enigmail was the most popular option, with 12 participants stating to use it regularly. Other options might have yielded insufficient sample sizes in our main study. Therefore, we focused on OpenPGP signatures in Thunderbird with Enigmail for our study.

4.2.3 Selection of Attacks for the Study

Based on the decision to focus on OpenPGP signatures in Thunderbird (68.4.1) with Enigmail (2.1.5), we designed a set of four attack emails with illegitimate signatures for this specific scenario. These four attack emails are loosely based on the “weak forgery” class described in [19]. Our study covered the UI redressing and ID attacks, and we added a new typo-domain case. Due to time constraints, MIME attacks were excluded. Other attacks in [19] were perfect forgeries at the cryptographic API layer and not relevant to our study. In detail, the eight used emails were as follows (cf. figure 1):

Legitimate (4x) Email with a legitimate signature. Enigmail shows a green bar “Good signature from Bob <bob@code-audit.org>”. An extended view of this email using all GUI elements is depicted in figure 2.

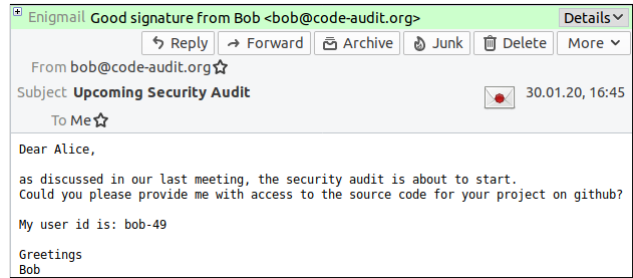
Broken-Signature (1x) Email with a broken signature. Enigmail shows a yellow bar “Unverified signature”.

Redressing (1x) Email with an inline image of Enigmail’s original green bar at the top of the email body. The simulated bar shows a green bar “Good signature from Bob <bob@code-audit.org>” and scales with the window but does not react to mouse clicks.

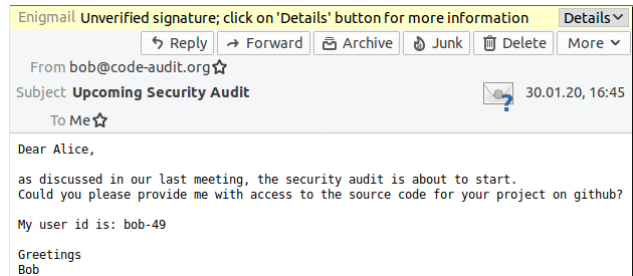
Conflicting-Signer (1x) Email signed by a different, easy to spot identity. Enigmail shows a green bar “Good signature from Celine <celine@example.org>”.

Conflicting-Signer-Subtle (1x) Email signed by a different, hard to spot identity. Enigmail shows a green bar “Good signature from Bob <bob@code-audit.org>”.

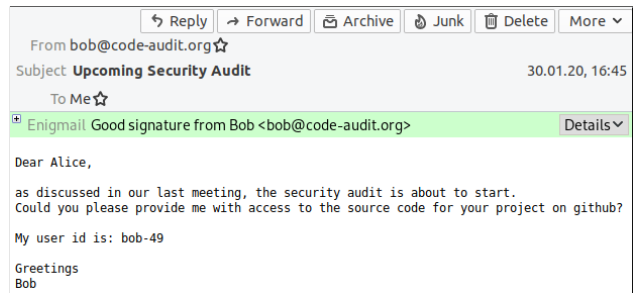
These messages are meant to imitate the work of an attacker, who can send and arbitrarily modify email messages. They can also create new identities and have public keys for these new identities placed as trusted in Alice’s keychain (to emulate key validation automation like WKD).



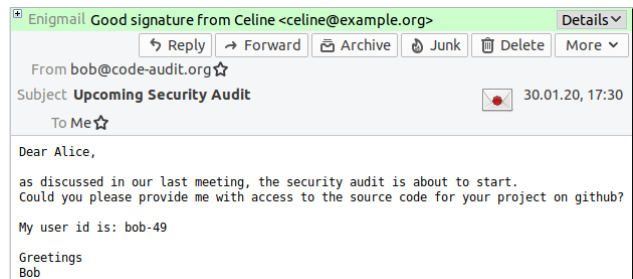
(a) Legitimate



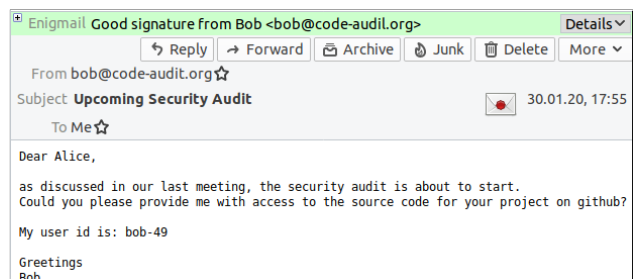
(b) Broken-Signature



(c) Redressing



(d) Conflicting-Signer



(e) Conflicting-Signer-Subtle, note the ‘l’ instead of ‘t’

Figure 1: Legitimate email and attack emails as displayed in Thunderbird 68.4.1 using Enigmail 2.1.5.

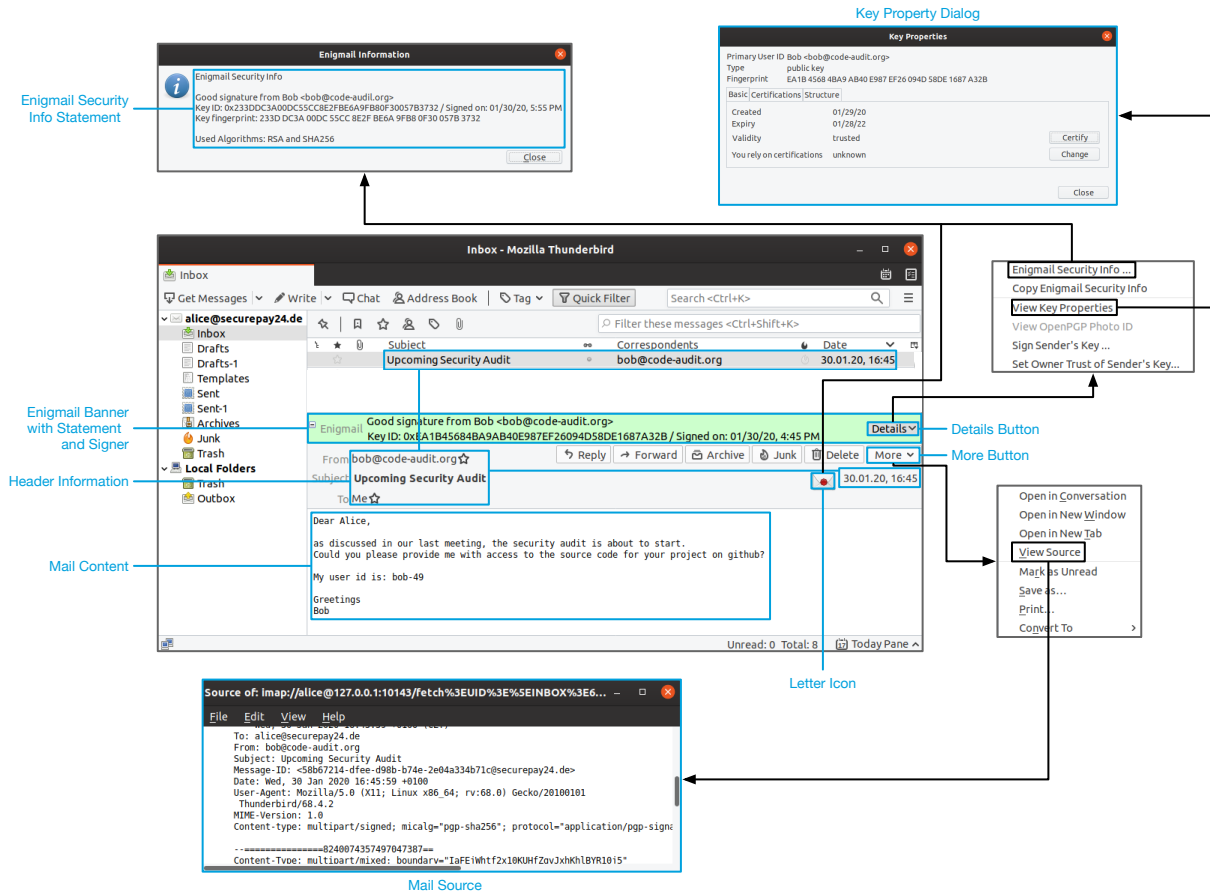


Figure 2: Overview of the Thunderbird 68.4.1 and Enigmail 2.1.5 GUI.

Our study aimed to evaluate the email client security indicators. All test emails were constructed such that their legitimacy could be deduced by only using the GUI elements. It was not required to inspect the source code or make further assumptions about the email context to identify the invalid signatures. All email messages had identical headers and text to ensure that our expert users focus on the available GUI elements when evaluating email signature’s legitimacy. The GUI elements are depicted in figure 2. Additional technical descriptions of the attacks can be found in appendix C.

To our knowledge, there are no well-established strategies what security checks users should perform and in which order. The following strategy would at least uncover the attacks in this study: First, check that the banner shows a valid signature. Second, check that the banner is the correct indicator in this email client and that the banner is at the right location. Third, check that the signer and the sender are identical.

4.2.4 Study Procedure

Our goal in designing the study procedure was to allow the participants as much freedom as possible and to perform all the checks they normally would and capture their thoughts.

Therefore, we decided to use a think-aloud protocol [39] and have the participants perform all study tasks on a prepared study laptop, where they could inspect all emails in a fully functional Thunderbird instance. All instructions and questionnaires were shown in a Firefox browser on this laptop and were implemented as surveys on the SoSciSurvey² platform.

Our study consisted of four parts (see figure 3). A Python script automated progression between the parts and started the screen and audio recording at the start of the third part.

Part 1 - Informed Consent and Explanations The participants had to consent to their participation and the analysis of their data (cf. appendix A.1.1). They received the instructions (cf. appendix A.1.2), including that their task would be to assess the legitimacy of email signatures. They were thus fully primed and the detection rates represent upper bounds. We discuss this design decision in section 4.4. To progress to the second part and start the actual decision tasks, the participants had to close the Firefox browser (cf. figure 3).

Part 2 - Introductory Questionnaire Participants had to fill an introductory questionnaire (see appendix A.2). It

²<https://www.sosicisurvey.de/>

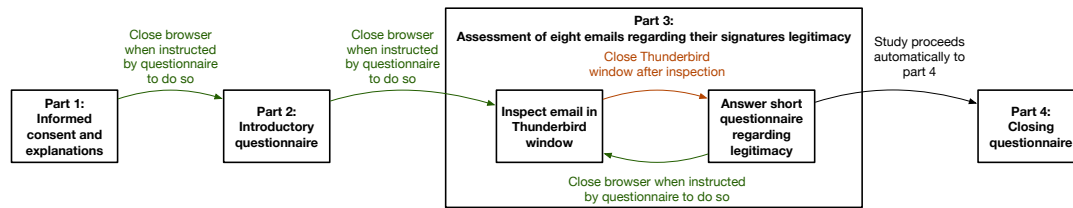


Figure 3: Overview of the four parts of our study.

queried whether the participant had participated in one of our previous studies. Those participants were not eligible to proceed. Furthermore, the questionnaire included five questions to measure the participants’ self-reported expertise with email encryption and signatures and one question on how often they use OpenPGP encrypted and signed emails in Thunderbird on average. The questionnaire ended with instructions for the third part. Therein, participants were asked to vocalize all thoughts and describe what they are doing during the assessment tasks, beginning from the moment they see the first email up to the point when they have made their decision regarding the legitimacy of the last of the emails. The detailed instructions can be found in appendix A.1.2. Again, participants would proceed to the next part by closing the Firefox browser displaying the questionnaire (cf. figure 3).

Part 3 - Assessment of Emails The participants were instructed to judge if a given email message was legitimately signed by bob@code-audit.org. All participants saw all eight emails listed in section 4.2.3 in a random order. The email messages were shown one at a time. The random order served to minimize any ordering bias. The Thunderbird interface was reset after each email message so that there was only one message in the inbox at any time, and participants could not jump back and forth between messages. During this part of the study, we captured the laptop screen and recorded audio of the participants thinking aloud. If the participants did not say anything, the experimenters reminded the participants to vocalize and explain their thoughts and actions. After inspecting each email message, a short questionnaire popped up, in which the participants could indicate their decision about this message and optionally note any issues they encountered.

Part 4 - Closing Questionnaire The closing questionnaire included the Refined Security Behavior Intentions Scale (RSe-BIS [30]) and the Security Attitudes scale (SA-6 [8]).

4.2.5 Second Pre-Study

We performed a second pre-study to pilot the study procedure described above. This second pre-study was held at the Chaos Communication Congress 2019, attended by an audience very similar to the summer camp. Participants were recruited by approaching attendants directly and handing out leaflets. Overall, we performed nine full runs of the OpenPGP study procedure. These runs allowed us to improve the setup and the

emails the participants inspected. For example, some participants falsely classified emails as illegitimate due to missing trace headers, i.e., *Received*, or other artifacts that we did not anticipate. These issues were corrected for the final study and steered the participants towards focusing on the graphical security indicators. Also, we addressed a data recording issue preventing full recordings for the think-alouds.

4.2.6 Main Study

We conducted our main study at the Free and Open Source Developer Meeting (FOSDEM) in February 2020 in Brussels.³ Like the CCC venues, this event is attended by IT specialists, but with a focus on Open Source rather than IT security.

Participants were recruited similarly to the second pre-study by approaching attendees directly and using leaflets. Additionally, FOSDEM 2020 hosted a room for Mozilla with a scheduled talk on Thunderbird development, and one of the co-located events at FOSDEM was a large OpenPGP key signing party. We used both of these opportunities for recruiting. If an attendee was interested in participating in our study, they were asked if they frequently use Thunderbird with Enigmail (OpenPGP) and if they already had participated in our pre-studies. Those that had were excluded. Similarly, participants who stated not to use Thunderbird with OpenPGP frequently were excluded. We then explained the study’s goal and the task participants would have to perform.

Overall, we conducted think-aloud sessions with 33 participants. Of these 33 participants, two had to be excluded since their questionnaire data indicated they did not use Thunderbird, three were excluded since they could not be considered expert users (scored lower than 2 on average in our self-reported expertise questions with no individual value larger than 2), two were excluded due to interruptions by third-party attendees, and one was excluded due to missing consent (presumably in error). This left us with 25 valid recordings of think-aloud sessions of expert users classifying signatures.

4.3 Analysis

Qualitative Analysis The think-aloud recordings were transcribed, including the mouse cursor actions and dialogues appearing on screen. Then qualitative analyses were performed

³Note that this was before the COVID-19 pandemic started, and in-person studies were still unproblematic: <https://archive.fosdem.org/2020/>

using an inductive coding approach [36] with two independent coders. The two coders created the codebook based on the research questions (cf. section 4.1) and an initial coding of three transcripts. Both coders coded five more transcripts to ensure inter-rater reliability (IRR). As a measure for IRR, Krippendorff’s α was used. The value of $\alpha = 0.71$ indicates a moderate IRR, which is acceptable given our unstructured think-aloud data and the exploratory nature of our study. The remaining 17 transcripts were coded independently, eight by one coder and nine by the other. The coders met to discuss changes or additions to the codebook as they arose from newly coded transcripts. The final codebook contained 69 codes in seven categories (see appendix B for the full codebook).

Quantitative Analysis For the SA-6 and RSeBIS scales and our self-reported expertise questions, the mean of all values for each participant was used in the correlation analyses. For the frequency of use, the answer for each participant was normalized to days per year.

4.4 Limitations

Our participants were sampled from a non-diverse group of people attending FOSDEM in person. As the conference was in Brussels, we expect the participants to be primarily from Belgium and adjacent countries. Therefore, our results might not generalize to other populations. The quantitative results would benefit from a larger sample. Yet, further data collection was prevented by the onset of the COVID-19 pandemic.

Participants were self-selected based on leaflet advertising and word of mouth. We asked participants to not share details of the study with others, but could not control communication between the participants and attendees outside the study. We did not observe any reactions of one participant to another.

Our participants were explicitly tasked with identifying whether a given email signature was legitimate or not. Thus, they were likely to check more thoroughly than under real-world circumstances. Priming our participants in this way was intentional. We wanted to capture our participants’ strategies validly even in the first email. We decided that priming our participants to use these strategies throughout the study would be the prudent way to collect this data. Our findings can thus be seen as an upper bound of the expert users’ capabilities. Six participants even mentioned at least once during the experiments, after identifying an attack, that they might fall for this in real life: “*But I don’t usually do these checks unless I know I am actively being targeted, like right now.*” -P6.

For our qualitative analysis, we rely on think-aloud data, which does not guarantee a complete insight into our participants’ minds and reasoning. We cannot rule out that some checks were not verbalized and are thus missing in our data.

Finally, the delay in our research due to the COVID-19 pandemic has seen Enigmail being integrated into Thunderbird, and as a result, the GUI changed. We describe these differences and discuss their impact on our results in section 6.

#	Checks	n
<i>Not related to PGP signatures</i>		
1	Header Information	113
2	Mail Content	41
3	Mail is Classified as Junk	1
4	Mail is Encrypted	1
<i>Related to Redressing Attacks</i>		
5	GUI Behaves Unexpectedly	31
6	Alternative Message Views	8
<i>Related to Enigmail GUI</i>		
7	Banner Indicator	116
8	Compare Signer to Sender	54
9	Security Info Statement	43
10	Fingerprint	39
11	Letter Icon Status	17
12	Banner Position	15
13	Banner Signer	15
14	Signature Date	14
15	Crypto Algorithms	4
<i>Related to Key Management</i>		
16	Sender’s Key	22
17	Signer’s Key is Signed with Own Key	11
18	Key Property Trust Statement	8
19	Key Validity	6
20	Key is in Keyring	5
21	Keyring	5
22	Key Creation Date	3
<i>Mail Source</i>		
23	Mail Source	86
<i>Proposed Checks</i>		
1*	Compare Fingerprint to Known One	11
2*	Mail Source	7
3*	Fingerprint	6
4*	Out of Band Verification	4
5*	Recheck with GPG on Command Line	3
6*	Keyring	2
7*	Key Revocation	1
8*	Signature Date	1

Table 1: Overview of the checks applied by our participants and how often they were applied. The checks are grouped regarding the five categories that emerged from the coding and sorted in descending order of their frequencies. The “Proposed Checks” at the bottom of the table are the checks that participants talked about but did not perform.

5 Results

In the following, we present the results of our study regarding the five research questions outlined in section 4.1.

5.1 RQ1: Checks & Strategies

5.1.1 Checks Applied by Participants

We identified 23 distinct checks in the transcripts of the 25 think-aloud sessions (cf. table 1). The checks are generally named after the information or GUI element that is inspected by the participant. See figure 2 for an overview of Thunderbird’s GUI. Participants applied on average 9.8 distinct

checks (median: 10, sd: 2.5) across all emails. Overall, the 23 checks were applied 659 times by our participants, with an average of 3.3 (median: 3, sd: 2.0) checks per email.

From the qualitative coding, five categories of checks emerged: (1) checks based on information not related to OpenPGP signatures, (2) checks based on information related to redressing attacks, (3) checks based on information related to the Enigmail GUI, (4) checks related to key management, and (5) checks based on inspecting the email source code. In the following, we discuss the checks in each category. In addition to the checks applied during the study, participants also proposed additional ones. These proposed checks will be discussed at the end of this section.

Checks Not Related to OpenPGP Signatures These checks are not related to the signatures at all. They are based on inspecting information even found in emails without signatures. Checking the *Header Information* is the most frequently applied check in this category with 113 applications. It includes all the header information displayed in Thunderbird's GUI, e.g., sender, recipient, subject, or date and time. The *Mail Content* was inspected 41 times. The remaining two checks, i.e., whether the *Mail is Classified as Junk* and whether the *Mail is Encrypted* were both applied only once. None of these checks can detect the attacks in our study.

Checks Related to Redressing Attacks The following two checks are not related to email signatures but allow detecting the *Redressing* attack. The most frequently applied check is a reaction when the *UI Behaved Unexpectedly*, which occurred 31 times. Usually, the first encounter with the fake banner prompted this check. Most participants correctly interpreted the unresponsive GUI and adopted these checks for subsequent emails. The second check in this category is inspecting the message in an *Alternative Message View*, such as in plaintext, in simple HTML, or looking at what a reply would contain. This check was applied only eight times.

Checks Related to Enigmail GUI These checks directly relate to the information displayed by the Enigmail GUI. The most frequently applied check (116 times) is checking the *Banner Indicator*, i.e., the statement ("Good signature", "Unverified signature", etc.) and color of the Enigmail banner, which both essentially communicate the same information to the user. This check can detect the *Broken-Signature* attack. The second most frequently used check is to *Compare Signer and Sender* (54 applications). This check can detect two of the four attacks in our study (i.e., *Conflicting-Signer* and *Conflicting-Signer-Subtle*). Another two checks that were somewhat similarly often applied are inspecting the *Security Info Statement* (applied 43 times) and checking the *Fingerprint* (applied 39 times). The *Security Info Statement* displays information similar to the Enigmail banner and allows detection of the same attack. All remaining checks were applied less than 20 times. Of these, checking the *Letter Icon Status* is the most useful, allowing to identify the redressing attack

with a first-level GUI element. However, this check is only possible if the user spots that this indicator is missing.

Checks Related to Key Management Some participants ventured beyond Thunderbird to perform checks related to their keychain. However, these checks were not used very frequently. Inspecting the *Sender's Key* (e.g., the existence of subkeys) is the most frequent check (22 applications) and checking whether the *Signer's Key is Signed with Bob's Key* is the second most frequently applied check (11 applications). All other checks were applied less than 10 times.

Checking the Mail Source Another relatively popular check (86 applications) was inspecting the *Mail Source*. While some participants just screened it in general, some inspected specific information, such as *Received* headers.

Proposed Checks Participants proposed several checks which they did not perform. Inspecting fingerprints is the most common theme, with 11 participants stating that outside the study setting they would try to *Compare the Fingerprint to a Known One* and another six participants stating that they might check the *Fingerprint* in more detail (without specifying how they would perform this check). Some of the checks were mentioned as potential further avenues but did not seem to be required at the time, for example: "So, I see some stuff that I could look at if I was at all suspicious that I probably haven't been looking at before." -P14. Other checks were not possible in the study, such as an *Out of Band Verification*: "In this case I would call Bob on the phone." -P10.

5.1.2 Overall Strategy for Application of Checks

Figure 4 exemplifies how our participants transitioned from one check to the next for the *Redressing* email. While a path with just two checks emerges, when following the transitions with the highest probabilities (participants realize that the *GUI Behaves Unexpectedly* and then inspect the *Mail Source*), figure 4 illustrates how participants did not seem to follow a pre-determined path for every mail. Only three of our 25 participants took this direct path, as expected from the probabilities. Instead, each new email was the start of a treasure hunt, as we watched our participants explore the Enigmail GUI. Consequently, there was a great variety in the order that checks were applied. Among the 42 transitions between checks we observed for the *Redressing* email, only in four instances are the transition probabilities above 50%. Often it seemed that participants were not sure what they were looking for next, as illustrated by P6 when they, after opening a dialog containing details of the signing key, uttered "I don't know why I check this. . .". This lack of a common strategy is consistent across all attacks (cf. appendix F). There is one exception to this though: if a participant checks the *Mail Source*, it is most frequently the last check they perform. This is, however, contrasted by many checks with a high fan-out and similar probabilities for the subsequent checks.

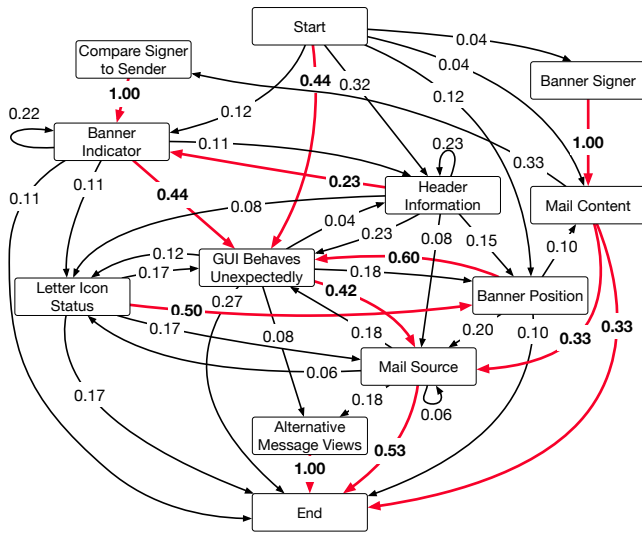


Figure 4: Participants’ transition probabilities from one check to another for the *Redressing* email. The most likely transition after each check is drawn in bold and red. Due to rounding, the probabilities for each node might not add up to 100%.

From inspecting the transition graphs of each email, a few additional relevant observations become apparent. For the *Broken-Signature* email, participants did not seem to trust the *Banner Indicator* check. Instead, when following the path of highest probabilities, participants would also check the *Security Info Statement*. This is noteworthy since the wording regarding the signature’s validity differs slightly in these two checks. While the Enigmail banner reads “Unverified signature”, the statement in the security info dialogue reads “Bad signature”. The latter seems to have had a much stronger impact on the participants’ decision. Participants’ suspicion might also have been caused by them being primed.

For the *Redressing* attack, both first-level Enigmail GUI elements are present among the checks: The banner and the letter icon. However, while the (spoofed) banner is among the first elements our participants checked, the *Letter Icon Status* is only ever checked after other checks were performed. This illustrates how a missing security indicator poses problems and might not be recognized by the participants, which replicates findings from other domains [6, 31].

For the *Conflicting-Signer* and *Conflicting-Signer-Subtle* email, several participants needed only one check, namely to *Compare Signer to Sender*. For the *Conflicting-Signer* attack this even represents the path with the highest probability (cf. figure 11 in appendix F).

5.2 RQ2: Usefulness of Checks

In order to understand which checks contributed most to participants’ detection of the attacks, we coded each of the checks regarding whether it pushed them towards the right decision, towards the wrong decision, or did seemingly not contribute

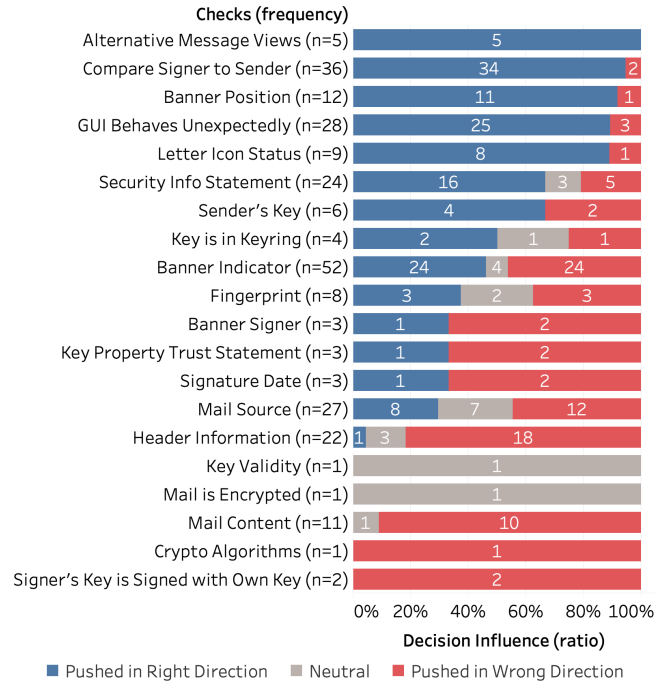


Figure 5: How the checks influenced participants’ decision for the emails with illegitimate signatures. Checks from table 1 not appearing here were only applied in the legitimate case.

to the decision (neutral). The latter case occurred in particular when participants could not interpret the information they checked, e.g.: “I don’t know what it what it [sic.] means, does it mean the signature does not match the content of the body or does it mean there is no trust part. That’s unclear.” -P16. We leave instances where participants did not comment on a certain check out of the analysis to not introduce unnecessary interpretation and bias into our results.

Figure 5 gives an overview of how each check influenced the decision of the participants when inspecting the emails with illegitimate signatures. Unsurprisingly, the checks based on information not related to email signatures are among the least effective. Two of the checks based on information provided by Enigmail or information found in the keyring, i.e., checking the *Crypto Algorithms* and checking whether the *Signer’s Key is Signed with Bob’s Key*, did not prove useful to the participants either. However, they were rarely used.

The most frequently applied check *Banner Indicator* pushed our participants as often towards a correct decision as it did towards an incorrect decision. This points towards severe issues with this most prominent part of Enigmail’s first-level GUI elements. The issues arise when we look at the decisions for the *Redressing*, *Conflicting-Signer*, and *Conflicting-Signer-Subtle* emails. In these attacks, the banner color is green, and the banner contains the text “Good signature”. For the *Redressing* email, the *Banner Position* is the better check, but it requires the participants to know how the

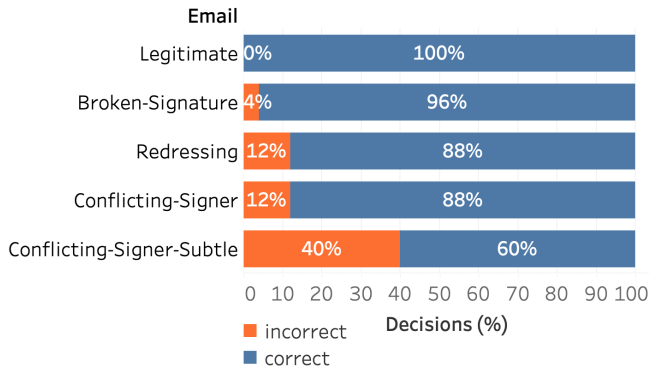


Figure 6: Overview of correct and incorrect classifications for the valid and each of the four attack emails.

interface is supposed to look and to recognize the difference. For the other two emails, the “Good signature” statement is insofar misleading as it does not reflect the expectations of the participants as will be further discussed in section 5.5.

When participants *Compare Signer and Sender* it mostly guides them towards the correct decision. Making this comparison as easy as possible would greatly benefit detecting the corresponding attacks. The more reliable but far less frequently used first-level GUI element is the *Letter Icon*. It is placed in the header area, and a click on it leads to the also relatively helpful *Security Info Statement*. Thus it might provide a better template for future designs.

Checking the *Mail Source* also stands out: It pushed more participants towards an incorrect decision than a correct one. Specifically, participants misinterpreted the information they saw or inspected header information that did not help them. Also, this check exhibits the highest number of neutral ratings where participants could not interpret the information they saw, e.g., P17 pondered : “Another weird segment. I am probably not knowledgeable enough regarding MIME parts.”

5.3 RQ3: Performance of Participants

Figure 6 shows an overview of the correct and incorrect responses for the valid mails as well as each of the attacks. It becomes apparent that our participants were fairly successful in discerning legitimate and illegitimate emails. However, overall 52% of participants failed to detect at least one of the attacks (average 0.76 attacks per participant, *median* = 1, *sd* = 0.97). Thus, these misclassifications are not due to repeated failures by a few participants, but they seem to be fairly evenly distributed among the participants.

When looking at the attacks individually, figure 6 clearly shows that the more intricate the attacks become, the more difficulties even expert users have. The *Conflicting-Signer-Subtle* attack was the most successful, with 40% of participants falling for it. This is likely due to two issues. Firstly, this attack is presumed to take place *after* the corresponding key for *bob@code-audil.org* was imported into the victim’s

		RSeBIS	SA-6	TE	FoU	CR
RSeBIS	ρ	1	** .760	.341	.226	-.195
	Sig.	-	< .001	.095	.278	.351
SA-6	ρ	** .760	1	** .586	.304	.014
	Sig.	< .001	-	.002	.139	.946
TE	ρ	.341	** .586	1	* .464	-.201
	Sig.	.095	.002	-	.019	.336
FoU	ρ	.226	.304	* .464	1	-.091
	Sig.	.278	.139	.019	-	.665
CR	ρ	-.195	.014	-.201	-.091	1
	Sig.	.351	.946	.336	.665	-

Table 2: Overview of the investigated Pearson correlations ρ . In all cases $n = 25$. Calculations are 2-tailed. ** marks significance at the .05/.01 level. TE = Technical Expertise, FoU = Frequency of Usage, CR = Ratio of Correct Responses

keyring, e.g., by automated key retrieval such as WKD, leading to a green Enigmail banner signaling a valid signature to the victim. Secondly, the discrepancy between signer and sender was minimal, with the two differing in only one letter, which even looks similar at first glance. The effect of a more obvious discrepancy between signer and sender can be seen in our easier *Conflicting-Signer* case: only 12% of participants fall for this attack. Similarly, 12% of participants fall for the *Redressing* attack. The *Broken-Signature* email was still classified as legitimate by one participant because they found the key which was used to originally sign the (subsequently manipulated) email in their keyring.

5.4 RQ4: Predictability of Success

We wanted to investigate whether the participants’ *security behavior intention* (RSeBIS scale), *security attitude* (SA-6 scale), *self-reported technical expertise*, or the *frequency of use* (uses of OpenPGP and Thunderbird per year) might be used as predictors of the ratio of correct responses for each participant. This investigation is considered exploratory, with correlations between RSeBIS, SA-6 and the participants’ performance deemed not unlikely and the other two constructs being completely exploratory. Yet, from the correlation analysis with Pearson’s ρ (cf. table 2), it becomes quickly apparent that none of the measures can serve as a meaningful predictor. In contrast, the measures seem to be predictors for each other, particularly for RSebis and SA-6 as expected [8].

5.5 RQ5: User Perceptions

We observed many participants blaming themselves for any possible errors or slips that might have decreased their success in labeling the messages correctly. E.g., P18 already took it onto themselves to write down the key fingerprint of Bob, but then still said they did not do enough due diligence: “I

did write down the SHA256 signature, and they didn't match. Something is fishy. And now I regret that I didn't write down the creation and expiration dates. Insufficient due diligence.” -P18 From a usable security perspective, this seems absurd – a tool should never expect the user to write down or compare dates or long strings when it could just do it itself.

P24 was able to point out that in our Conflicting-Signer-Subtle email the signer's address had a typo. However, they decided to label the message as legitimately signed anyway, since they perceived Enigmail's statement “valid signature” to be trustworthy, outweighing their concerns about address inconsistencies: “I think this email signature is legit [selects “legitimate signature”]. However, the email header was somehow, um... worked with. It's just a guess, because I assume that Enigmail has also correctly verified the signature that is shown as correct. Where the discrepancy with the email address From-field comes from, I just don't know.” -P24. This was, however, the only instance where a participant noticed the address inconsistency but still went on to label the message as legitimately signed.

Four participants were upset when they realized that Enigmail was not pointing out that that signer was different from the sender, e.g.: *I would say this one is legit. [Pause] Except that it is signed by Celine... What?! OK. That is quite strange that Thunderbird does not claim anything about, like, it's signed by a different guy than the sender!* -P3. The remaining participants who noticed inconsistencies between signer and sender were mostly confused or insecure and could not pinpoint where the issue was exactly. However, they were able to recognize that *something* was off and thus labeled the message as illegitimate, accordingly.

As mentioned in section 5.1.2, our participants did not seem overly determined by neither having a certain, strict click-path nor a set of known indicators to look for. Instead, our participants were often startled, confused, or even annoyed when navigating the GUI elements offered to them by Enigmail and Thunderbird. P23 simply gave up on finding more information on the keys in their key chain after looking for, but not finding it, in three different places: “[After a very long search through Thunderbird's settings, looking for PGP Keys] Personally this is taking too long for me right now. [closes settings] That's why I cancel this [clicks on Inbox in folder selection] and would claim the email is just not trustworthy and stick to my first impression.”

In summary, even for our sample of expert users, the task of recognizing illegitimate OpenPGP signatures is generally accompanied by haphazardness and uncertainty.

6 Changes in Newest Thunderbird Version

Our study was conducted in 2020 with Thunderbird 68.4.1 and Enigmail 2.1.5. Since then, Thunderbird 78.2.1 has been released with built-in OpenPGP support [33]. Thus, we re-evaluated the presented attacks with the newest version of

Thunderbird (91.5.0) and discuss which study results are relevant to the newest version, or for email signatures in general.

Overall Assessment of Changes Figure 7 shows a valid email in Thunderbird 91.5.0. It uses new design elements and the signature validity status is not as prominent as in previous versions (cf. figure 1). The Enigmail banner and the letter icon were replaced with one button in the header area. The button is labeled “OpenPGP,” and an icon shows the signature status in green color. Upon pressing the button, a new dialogue appears. It contains a short signature status statement, the signer key ID, and a button that allows inspecting the key and the encryption status. While the newer interface is cleaner and contains just one first-level and one second-level GUI element, we also see negative properties. For example, the colored area in previous versions was much larger (cf. figure 1), which made the email validity more immediately apparent.

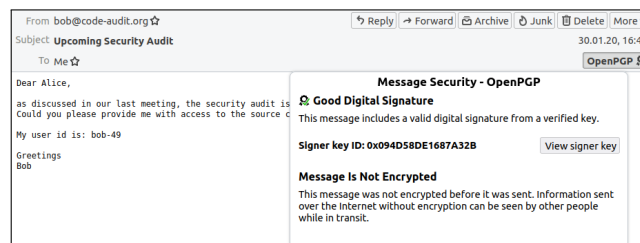


Figure 7: Legitimate email displayed in Thunderbird 91.5.0.

Relevance for Broken-Signature Email (Figure 9) The validity status became clearer. The wording is now “Invalid” email instead of “Unverified” email as it was previously. Also, the red color in the icon signifies the email invalidity. On the negative side, the colored area is much smaller.

Relevance for Redressing Email Our Redressing email has a now obsolete design and would not work in current versions of Thunderbird. Yet, since the validity indicator is not being displayed for unsigned emails in newer Thunderbird versions as well, the base issue remains. More research is needed to determine the viability of such attacks in the new GUI.

Relevance for Conflicting-Signer and Conflicting-Signer-Subtle Email (Figure 8) Detection of conflicting signer and sender got easier in newer Thunderbird versions. A red lock directly shows if the signer is not equal to the sender. Clicking the OpenPGP symbol reports *Uncertain Digital Signature* and allows the user to review the signer's key, which can make the detection of these attacks easier for users.

Yet, a bug allows bypassing this security indicator. By using two From headers (a technique from [19]), we were able to have Thunderbird 91.5.0 display a green icon. Our analysis revealed that the first From header was used to display the message sender. The second From header was used for the message sender validation and for displaying the name of the user in the list of emails. Thus, the second From header also

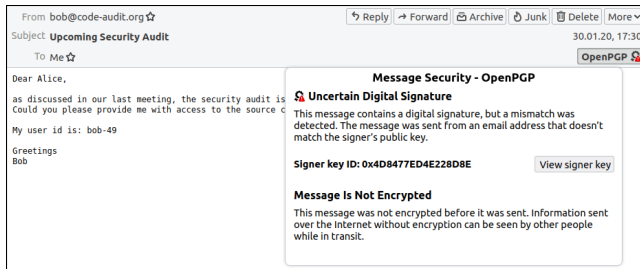


Figure 8: Thunderbird 91.5.0 notifies users about an *uncertain* digital signature when signer and sender differ.

includes a display name of Bob. This attack is only detectable by investigating the message signer in the third-level GUI (after clicking on the OpenPGP icon and the icon *View signer key*). We reported this issue to the Thunderbird developers, who plan a fix in the upcoming release.

Thunderbird 91.5.0 implements a custom key management instead of relying on the local GPG keyring. Thus, possibly insecure configurations where keys are side-loaded and trusted by a mechanism like WKD are less likely. Yet, OpenPGP key validation remains an open problem for Thunderbird users.

7 Discussion

Implications Thunderbird and Enigmail as used in our study have since been replaced with a built-in solution. Yet, our most important result remains unimpaired by this change: Even our expert user participants had no effective strategies to assess whether email signatures were legitimate or not. Instead, participants explored the interface as they went and exhibited much uncertainty about what to check. Some of the most frequently performed checks are of questionable usefulness, e.g., inspecting the *Mail Source*.

However, the users are not to blame here. We saw users baffled by the GUI or overwhelmed by the complexity of the necessary checks. The GUI needs to give meaningful support to the user when they need to perform these complex checks with obvious affordances [21] that invite to perform useful checks and deter from performing unuseful ones. This becomes particularly apparent when sender and signer differ. This discrepancy is not highlighted in the *Conflicting-Signer* and *Conflicting-Signer-Subtle* emails. This was perplexing for users, and we agree with this assessment. The problem seems to be that the signature is technically valid (i.e., no manipulation of the email), but the email context carries the additional expectation that it is only legitimate if it was signed by the sender. Honoring these expectations is what developers should strive for, and supporting developers in achieving this task by mapping out these expectations in an easily digestible way is the future work ahead of us as research community. Our work also highlights that email client GUIs need trustworthy zones where security status indicators can reside to

impact the viability of *Redressing* attacks. Future work is needed to formulate proper guidelines in this respect.

Also, our research supports the results of earlier studies. Most checks relating to key management, e.g., checking the *Key Validity*, leave at least as many participants in uncertainty or lead them to incorrect decisions as they helped participants. Similarly, the signature GUI seems geared towards checking for simple manipulations, not more sophisticated attacks. In both cases (key management and usability issues), our work extends the existing research, which reports on the usability issues surrounding encryption and digital signatures. Yet, the “upper bound” detection rates in our results due to the priming of our participants underlines the severity of these issues.

Recommendations We believe that the proper long-term solution for end-to-end email security is shifting the ecosystem from *indicating secure messages* to *warning about (potentially) insecure messages*. This is important, because the absence of security indicators is often overlooked by users [6, 31]. The security of systems should not rely on users checking for the presence of indicators. However, to avoid warning fatigue [1], this shift can only happen after end-to-end secured email has become the default for email communication. In this chicken-and-egg problem, it is up to current tools to help adoption by implementing these security features as usable as possible.

Overcoming the complexity of checking a signature’s legitimacy before hand-off to the user plays a key role. We believe an approach based on allowlists of secure MIME structures, as described in [13] to classify emails is key to achieving this. In particular harnessing the power of crowd-sourcing to maintain and extend such allowlists seems like a desirable approach. Based on these allowlists, we envision that email clients automate as many checks as possible and that interfaces distinguish four cases: (a) the *legitimate case*, where the signed email’s structure is in the allowlist and the signature is validly signed by the sender; (b) the *illegitimate case*, where the signed email’s structure is in the allowlist, but the signature is not valid or not from the sender; (c) the *check case*, where the signed email’s structure is not in the allowlist and the GUI has to support the user in performing useful checks; and (d) the *unsigned case*, where the email is not signed.

Coloration to distinguish the cases may support users. Yet, the colors should be chosen to be accessible by users with colorblindness.⁴ Also, the fourth case should not be skipped as is currently the case for most email clients. Such “missing indicators” rely on the user realizing that the indicator is not there, which has been proven to be problematic. This would introduce a source of conflicting information for *Redressing* attacks and thus make them easier to spot for users.

⁴E.g. <https://davidmathlogic.com/colorblind/> can be used to choose suitable colors.

Acknowledgements

We thank the organizers of the Free and Open Source Developer Meeting 2020 for providing an interview location on the premises of the event and allowing us to recruit participants from among the attendees. We also thank Imke Ines Klatt for her feedback on the referee instructions, recruiting, and execution of participant sessions during the 36c3. We thank Tobias Kappert and Christian Dresen for the execution of the first pre-study at the Chaos Communication Camp 2019, Martin Grothe for his support during the interview process at FOSDEM, Jonathan von Niessen for his security analyses of signature validation in Thunderbird, as well as Christiane Rosa, Marie-Claire Thiery, and Fabian Ballreich for their help in transcribing the think-alouds.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2092 CASA – 390781972. Marcus Brinkmann was partially supported by the German Federal Ministry of Economics and Technology (BMWi) project “Industrie 4.0 Recht-Testbed” (13I40V002C). Damian Poddebniak was supported by a research grant of the Münster University of Applied Sciences. This research was further supported by funding from the topic Engineering Secure Systems, subtopic 46.23.01 Methods for Engineering Secure Systems, of the Helmholtz Association (HGF) and by KASTEL Security Research Labs.

References

- [1] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A Large-Scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 257–272, Washington, D.C., August 2013. USENIX Association.
- [2] Wei Bai, Doowon Kim, Moses Namara, Yichen Qian, Patrick Gage Kelley, and Michelle L. Mazurek. Balancing Security and Usability in Encrypted Email. *IEEE Internet Computing*, 21(3):30–38, 2017.
- [3] J. Callas, L. Donnerhackle, H. Finney, D. Shaw, and R. Thayer. OpenPGP Message Format. RFC 4880 (Proposed Standard), November 2007. Updated by RFC 5581.
- [4] Jianjun Chen, Vern Paxson, and Jian Jiang. Composition kills: A case study of email sender authentication. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2183–2199. USENIX Association, August 2020.
- [5] D. Crocker. STANDARD FOR THE FORMAT OF ARPA INTERNET TEXT MESSAGES. RFC 822 (Internet Standard), August 1982. Obsoleted by RFC 2822, updated by RFCs 1123, 2156, 1327, 1138, 1148.
- [6] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems, pages 581–590, 2006.
- [7] M. Elkins, D. Del Torto, R. Levien, and T. Roessler. MIME Security with OpenPGP. RFC 3156 (Proposed Standard), August 2001.
- [8] Cori Faklaris, Laura A. Dabbish, and Jason I. Hong. A Self-Report Measure of End-User Security Attitudes (SA-6). In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, Santa Clara, CA, August 2019. USENIX Association.
- [9] N. Freed and N. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. RFC 2045 (Draft Standard), November 1996. Updated by RFCs 2184, 2231, 5335, 6532.
- [10] Simson L. Garfinkel, David Margrave, Jeffrey I Schiller, Erik Nordlander, and Robert C Miller. How to make secure email easier to use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems, pages 701–710, 2005.
- [11] Simson L. Garfinkel and Robert C. Miller. Johnny 2: A User Test of Key Continuity Management with S/MIME and Outlook Express. In *Proceedings of the 2005 Symposium on Usable Privacy and Security*, SOUPS '05, page 13–24, New York, NY, USA, 2005. Association for Computing Machinery.
- [12] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. Secrecy, flagging, and paranoia: adoption criteria in encrypted email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems, pages 591–600, 2006.
- [13] Daniel Gillmor. Guidance on end-to-end e-mail security. <https://datatracker.ietf.org/doc/draft-ietf-lamps-e2e-mail-guidance/>, 2022. Online; accessed 13 February 2022.
- [14] Joscha Lausch, Oliver Wiese, and Volker Roth. What is a secure email? In *European Workshop on Usable Security (EuroUSEC)*, 2017.
- [15] Ada (Adam) Lerner, Eric Zeng, and Franziska Roesner. Confidante: Usable Encrypted Email. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE European Symposium on Security and Privacy, pages 385–400, 2017.

- [16] Juan Ramón Ponce Mauriés, Kat Krol, Simon Parkin, Ruba Abu-Salma, and M. Angela Sasse. Dead on Arrival: Recovering from Fatal Flaws in Email Encryption Tools. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2017)*, The LASER Workshop, pages 49–57. USENIX Association, 2015.
- [17] Susan E. McGregor, Elizabeth Anne Watkins, Mahdi Nasrullah Al-Ameen, Kelly Caine, and Franziska Roesner. When the Weakest Link is Strong: Secure Collaboration in the Case of the Panama Papers. In *Proceedings of the 26th USENIX Security Symposium*, USENIX Security Symposium, pages 505–522. USENIX Association, 2017.
- [18] Azadeh Mokhberi and Konstantin Beznosov. Sok: Human, organizational, and technological dimensions of developers’ challenges in engineering secure software. In *European Symposium on Usable Security 2021*, page 59–75. Association for Computing Machinery, 2021.
- [19] Jens Müller, Marcus Brinkmann, Damian Poddebniak, Hanno Böck, Sebastian Schinzel, Juraj Somorovsky, and Jörg Schwenk. “Johnny, you are fired!” – Spoofing OpenPGP and S/MIME Signatures in Emails. In *28th USENIX Security Symposium, USENIX Security 2019.*, 2019.
- [20] Jens Müller, Marcus Brinkmann, Damian Poddebniak, Sebastian Schinzel, and Jörg Schwenk. Re: What’s Up Johnny? – Covert Content Attacks on Email End-to-End Encryption. <https://arxiv.org/ftp/arxiv/papers/1904/1904.07550.pdf>, 2019.
- [21] Don Norman. Affordances and design. *Unpublished article, available online at: http://www.jnd.org/dn.mss/affordances-and-design.html*, 2004.
- [22] Adrian Reuter, Ahmed Abdelmaksoud, Karima Boudaoud, and Marco Winckler. Usability of End-to-End Encryption in E-Mail Communication. *Frontiers in Big Data*, 4:568284, 2021.
- [23] Adrian Reuter, Karima Boudaoud, Marco Winckler, Ahmed Abdelmaksoud, and Wadie Lemrazzeq. Secure email - a usability study. In Matthew Bernhard, Andrea Bracciali, L. Jean Camp, Shin’ichiro Matsuo, Alana Maurushat, Peter B. Rønne, and Massimiliano Sala, editors, *Financial Cryptography and Data Security*, pages 36–46, Cham, 2020. Springer International Publishing.
- [24] Volker Roth, Tobias Straub, and Kai Richter. Security and usability engineering with particular attention to electronic mail. *International Journal of Human-Computer Studies*, 63(1-2):51–73, 2005.
- [25] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O’Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. “We’re on the Same Page”: A Usability Study of Secure Email Using Pairs of Novice Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems, pages 4298–4308. ACM, 2016.
- [26] Scott Ruoti, Jeff Andersen, Travis Hendershot, Daniel Zappala, and Kent Seamons. Private Webmail 2.0: Simple and Easy-to-Use Secure Email. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, Annual Symposium on User Interface Software and Technology, pages 461–472. ACM, 2016.
- [27] Scott Ruoti, Jeff Andersen, Tyler Monson, Daniel Zappala, and Kent Seamons. A Comparative Usability Study of Key Management in Secure Email. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 375–394, Baltimore, MD, August 2018. USENIX Association.
- [28] Scott Ruoti, Jeff Andersen, Daniel Zappala, and Kent Seamons. Why Johnny Still, Still Can’t Encrypt: Evaluating the Usability of a Modern PGP Client, 2015.
- [29] Scott Ruoti, Nathan Kim, Ben Burgon, Timothy van der Horst, and Kent Seamons. Confused Johnny: when automatic encryption leads to confusion and mistakes. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, Proceedings of the Ninth Symposium on Usable Privacy and Security - SOUPS ’13, pages 69–88, Ottawa, 2013. USENIX Association.
- [30] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-Confidence Trumps Knowledge: A Cross-Cultural Study of Security Behavior. *International Conference on Human Factors in Computing Systems*, pages 2202 – 2214, 2017.
- [31] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The Emperor’s New Security Indicators. In *2007 IEEE Symposium on Security and Privacy (SP ’07)*, IEEE Symposium on Security and Privacy, pages 51 – 65, 2007.
- [32] Steve Sheng, Levi Broderick, Colleen Alison Koranda, and Jeremy J. Hyland. Why johnny still can’t encrypt: evaluating the usability of email encryption software. In *Poster session of the Second Symposium On Usable Privacy and Security*, 2006.
- [33] Ryan Sipes. Openpgp in thunderbird 78. <https://blog.thunderbird.net/2020/09/openpgp-in-thunderbird-78/>, 09 2020. Online; accessed 18 February 2022.

- [34] Christian Stransky, Dominik Wermke, Johanna Schrader, Nicolas Huaman, Yasemin Acar, Anna Lena Fehlhaber, Miranda Wei, Blase Ur, and Sascha Fahl. On the Limited Impact of Visualizing Encryption: Perceptions of E2E Messaging Security. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, Symposium on Usable Privacy and Security, pages 437–454. USENIX Association, 2021.
- [35] Christian Stransky, Oliver Wiese, Volker Roth, Yasemin Acar, and Sascha Fahl. 27 Years and 81 Million Opportunities Later: Investigating the Use of Email Encryption for an Entire University. In *Proc. 43rd IEEE Symposium on Security and Privacy (SP'22)*, Symposium on Security and Privacy. IEEE, 2022.
- [36] David R. Thomas. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2):237–246, 2006.
- [37] Jan Tolsdorf and Luigi Lo Iacono. Vision: Shred If Insecure – Persuasive Message Design as a Lesson and Alternative to Previous Approaches to Usable Secure Email Interfaces. In *Proceedings of the 5th European Workshop on Usable Security (EuroUSEC 2020)*, European Workshop on Usable Security, pages 172–177, 2020.
- [38] Alma Whitten and J. D. Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Proceedings of the 8th Conference on USENIX Security Symposium - Volume 8, SSYM'99*, page 14, USA, 1999. USENIX Association.
- [39] Christopher D Wickens, Sallie E Gordon, Yili Liu, and J Lee. *An introduction to human factors engineering*, volume 2. Pearson Prentice Hall Upper Saddle River, NJ, 2004.
- [40] P. Wouters. DNS-Based Authentication of Named Entities (DANE) Bindings for OpenPGP. RFC 7929 (Experimental), August 2016.

A Procedure

Original instructions as presented to all participants of the study. Text enclosed by “<” and “>” denotes comments that were not contained in the original material.

A.1 Informed Consent and Explanation

A.1.1 Informed Consent (Page 1)

Dear participant, thank you for taking part in this study! Your participation will take only about 20-30 minutes of your time, but will help us tremendously in understanding the usage of email encryption and signatures.

We are researchers from the University of Applied Science Münster and Karlsruhe Institute of Technology. Our goal in this study is to investigate the usage and perception of e-mail encryption and signatures. This study comprises the following steps:

1. Informed consent (this page)
2. Instructions and introductory questions
3. Assessment of eight (8) emails regarding their signatures' legitimacy
4. Closing questionnaire

We will ask you to vocalize your thoughts while looking at the emails. In order to get a better understanding of your perceptions of the emails, both, your voice and your interaction on the screen will be recorded. The basis for the collection and analysis of the data is our data privacy policy [DE/EN].

Your participation is voluntary. If you wish to withdraw your participation before, during or after the completion of the survey, you can do so. If withdrawing, all data recorded up until this point will be discarded and deleted. For withdrawal from the survey once you have completed it, you will need to provide the participant code that you see below. Please write it down on the piece of paper provided to you.

Participant code: <randomly generated>

Please check the box below to indicate your agreement to participate in the study.

I am at least 18 years old and agree to participate in the study under the conditions as stated above.

A.1.2 Explanation (Page 2)

Dear participant, thank you for agreeing to participate in this study! To complete this study, you have to progress through four parts which are described in the following.

Note: Throughout the study, you will be asked to close program windows in multiple instances. This is an essential part of the study and represents having completed the respective task. Therefore, be sure to close the program windows only, once you have completed the respective task. In particular, when answering questionnaires only close the windows, once the questionnaire instructs you to do so or your answers might be lost.

Part 1: Informed consent and explanation This part is the one you currently see. It comprised your agreement to participate in this study (previous page) and explains the tasks comprised in this study (this page).

Part 2: Introductory questionnaire In this part, the task is to fill a questionnaire with questions about your usage of email encryption and signatures as well as questions regarding your IT background.

Part 3: Assessment of eight emails regarding their signature's legitimacy This part comprises the tasks of rating the legitimacy of the signatures of eight emails. Each rating task encompasses the following two steps:

1. Each email will be opened in a dedicated Thunderbird window, allowing you to check the legitimacy of the signature. Once you have decided whether or not the signature is legitimate, you will have to close the Thunderbird window.
2. Once the Thunderbird window is closed, a short questionnaire in which you have to indicate whether the signature is legitimate or not will be opened automatically. Having completed the questionnaire for the respective email, you will be instructed to close the respective browser window in order to advance to the next email. However, after completing the questionnaire of the 8th (last) email, you will automatically continue with part 4 of the study, without the need to close the window.

These two steps are repeated for each of the eight emails. Note that each email should be rated in isolation, i.e. the emails do not reference each other and should be rated on its own.

In order to get a better understanding of your perceptions of the emails, we will ask you to vocalize your thoughts and to explain your actions while looking at the emails. Talk out loud constantly, telling everything you are thinking beginning from the moment you see the first email up to the point when you have made your decision regarding the legitimacy of the last of the emails. Please try to not plan out what you are going to say and do not try to explain your thoughts. Just act as if you were alone in the room and talking to yourself. Your voice and your interaction on the screen will be recorded. The recording of the screen and your voice will start automatically at the beginning of this part.

Part 4: Closing questionnaire

This part comprises a final questionnaire.

This procedure with all four parts is illustrated in the following: <Inline image of procedure, see Figure 3.>

A.2 Introductory Questionnaire

1. Have you participated in a study on email encryption and signatures at either the Chaos Communication Camp 2019 or the Chaos Communication Congress 2019?
 Yes No
2. Please describe how you usually check if an email you received is legitimate or was sent by a scammer.
Please do not include any sensitive information about other people in your answer.
<Multiline free text form>

3. Are there any additional checks you would perform on all incoming emails if you knew you were at risk of being specifically targeted?

Please do not include any sensitive information about other people in your answer.

<Multiline free text form>

4. Please indicate to what extent the following statements apply to you. <Likert items from (1) “does absolutely not apply to me” to (5) “absolutely applies to me”>

- I use email encryption and signatures regularly
- I am confident in my ability to use email encryption and signatures (PGP, S/MIME, etc.)
- I feel confident in being able to explain how to operate the email encryption and signature scheme I use (PGP, S/MIME, etc.) to others
- When encountering problems handling encrypted or signed emails I usually know what the problem is
- I believe I would recognize emails with invalid signatures

5. I handle PGP encrypted and signed emails in Thunderbird on average about <dropdown>

- once
- twice
- three times
- four times
- five times
- more than five times
- I don't handle PGP encrypted / signed emails

per <dropdown>

- hour
- day
- week
- month
- year
- I don't handle PGP encrypted / signed emails

<new page>

For this part of the study, please assume the following:

- You are Alice, a software developer at SecurePay24.
- Your email address is alice@securepay24.de.
- Your company has authorised an external security audit of the software you are currently working on.
- The security audit is performed by Code Audit Inc.
- You know Bob, the contact person at Code Audit Inc., from a conference call meeting.
- Bob's email address is bob@code-audit.org.
- You have exchanged keys with Bob, i.e. you have his public key in your keychain.

Please click “Next” to continue with the study.

<new page>

In the next step of this study you will see an email opened in Thunderbird. Inspect it to determine whether its signature is legitimate or not. After having inspected the email, please close the Thunderbird window to proceed with the study. Remember: all emails are independent of each other and should be rated in isolation.

In this part of the study, please vocalize your thoughts and explain your actions while looking at the emails. Talk out loud constantly telling everything you are thinking, beginning from the moment you see the first email up to the point when you have made your decision regarding the legitimacy of the last of the emails. Please try to not plan out what you are going to say and do not try to explain your thoughts. Just act as if you were alone in the room and talking to yourself.

Please close this browser window now to proceed to the email. This will also start the recording of your voice and the interaction on screen.

A.3 Assessment of Eight (8) Emails Regarding Their Signatures’ Legitimacy

1. Is the signature of the previously inspected email legitimate?

- Yes, the signature is legitimate
- No, the signature is not legitimate

2. Is there anything else you want to tell us with respect to the email you saw?

Note here e.g. if you have closed the windows prematurely (i.e. before finishing inspecting the email).

<Multiline free text form.>

A.4 Closing Questionnaire

A.4.1 SA6

3. On a scale of “Strongly Disagree” to “Strongly Agree”, rate your level of agreement with the following statements. <Likert items from (1) “Strongly Disagree” to (5) “Strongly Agree”>

- I am extremely knowledgeable about all the steps needed to keep my online data and accounts safe.
- I am extremely motivated to take all the steps needed to keep my online data and accounts safe.
- I often am interested in articles about security threats.
- I seek out opportunities to learn about security measures that are relevant to me.
- Generally, I diligently follow a routine about security practices.
- I always pay attention to experts’ advice about the steps I need to take to keep my online data and accounts safe.

A.4.2 RSebis

4. To what extent do following statements apply to you? <Likert items from (1) “Never” to (5) “Always”>

- I use a PIN or passcode to unlock my mobile phone.
- I include special characters in my password even if it’s not required.
- When browsing websites, I mouseover links to see where they go, before clicking them.
- If I discover a security problem, I fix or report it rather than assuming somebody else will.
- When I’m prompted about a software update, I install it right away.
- I use different passwords for different accounts that I have.
- I set my computer screen to automatically lock if I don’t use it for a prolonged period of time.
- I try to make sure that the programs I use are up-to-date.
- When I create a new online account, I try to use a password that goes beyond the site’s minimum requirements.
- I manually lock my computer screen when I step away from it.
- I change my passwords even if it is not needed.
- I use a password/passcode to unlock my laptop or tablet.
- I know what website I’m visiting by looking at the URL bar, rather than by the website’s look and feel.
- I verify that information will be sent securely (e.g., SSL, “https://”, a lock icon) before I submit it to websites.
- I verify that my anti-virus software has been regularly updating itself.
- When someone sends me a link, I open it only after verifying where it goes.

A.4.3 Debriefing

Thank you for participating in this study!

The study is now finished, please contact the experimenter to receive a debriefing and ask any potential questions you might have.

B Code Book

USED CHECKS: *Alternative Message Views, Banner Indicator, Banner Position, Banner Signer, Compare Signer to Sender, Crypto Algorithms, Fingerprint, GUI Behaves Unexpectedly, Header Information, Key Creation Date, Key is in Keyring, Key Property Trust Statement, Key Validity, Keyring, Letter Icon Status, Mail Content, Mail is Classified as Junk, Mail is Encrypted, Mail Source, Security Info Statement, Sender's Key, Signature Date, Signer's Key is Signed with Own Key;*

PROPOSED CHECKS: *Compare Fingerprint to Known One, Fingerprint, Key Revocation, Keyring, Mail Source, Recheck with GPG on Command Line, Out of Band Verification, Signature Date;*

USEFULNESS: *Indeterminate, Neutral, Right Direction, Wrong Direction;*

DECISION: *False Illegitimate, False Legitimate, True Illegitimate, True Legitimate;*

PERCEPTION: *Email is encrypted, I might fall for this in real life, I might fall for this in a study, No distinction of Thunderbird and Enigmail, Not sure why Thunderbird trusts signature, Uncertainty leads to mistrust;*

PROBLEM: *Bad GUI design, Does not know what to do, GUI target too small, Misleading GUI, Unable to locate desired option, Unhelpful information;*

VALIDITY: *Check possible due to study setting, Check potentially failed due to study setting, Checks intensively*

C Email Test Cases

Any email consist of a set of headers and a payload. The test emails had the following headers: Received, To, From, Subject, Message-ID, Date, User-Agent, MIME-Version, and Content-Type. From these headers, only To, From, Subject, and Date, are used by Thunderbird in the graphical UI. Other headers are only available by additional configuration or when viewing the raw email source.

In the eight provided test cases, only the Content-Type could differ, in order to use different body payloads. In other words, only the email body is relevant to discern legitimate from illegitimate emails in our study. All (irrelevant) headers were set to “sane” defaults, such that no participant focused (nor were misguided) by missing or incorrect headers.

The following email serves as a template for all email test cases:

```
Received: ... // irrelevant
To: alice@securepay24.de
From: bob@code-audit.org
Subject: Upcoming Security Audit
Message-ID: ... // irrelevant
Date: Wed, 30 Jan 2020 16:45:59 +0100
User-Agent: ... // irrelevant
MIME-Version: ... // irrelevant
Content-Type: {}

{}
```

Although the actual payload differed, the Thunderbird UI always showed the following text in its main window:

```
Dear Alice,

as discussed in our last meeting, the security audit is
about to start.
Could you please provide me with access to the source
code for your project on github?

My user id is: bob-49

Greetings
Bob
```

C.1 Email Test Case: Legitimate

A legitimately signed email. The root Content-Type is multipart/signed and the payload was correctly signed with the key of bob@code-audit.org.

C.2 Email Test Case: Broken Signature

This test recreates an email with a broken signature. It only differs from the legitimate email by a non-functional change to the MIME boundary. In effect, Thunderbird is not able to correctly verify the signature anymore, but the signer is still bob@code-audit.org.

C.3 Email Test Case: UI Redressing

An email without a cryptographic signature at all. HTML and CSS were used to mimic Enigmail's “green bar.” The bar is not clickable, but otherwise a pixel-perfect copy of the original Enigmail bar. It resizes when Thunderbird is resized. However, the position of the bar differs from Enigmail. In Enigmail versions below 2.0.8 the bar was below Thunderbird's header area, and this placement is used in this test mail. However, Enigmail has since changed the position of the green bar to be above Thunderbird's header area. The source code was obfuscated to hide the HTML and CSS elements (via base64), and the MIME boundary was set to --PGP SIGNED MESSAGE--- to pretend that OpenPGP was used in some form. Due to the required images, the source code was substantially longer compared to the legitimate email.

Reasoning We obfuscated the source code to redirect our participants to focus on the Enigmail elements, since prior participants at 36c3 classified the email as illegitimate as soon as they saw the HTML source code.

C.4 Email Test Case: Sender is not Signer

This email is equal to the legitimate email except for the OpenPGP signature. Here, the email was signed with the key of celine@example.org, instead of bob@code-audit.org. However, the From header still states that the email is from bob@code-audit.org. This test is motivated by the fact that OpenPGP signatures are traditionally not bound to the

From header. In S/MIME, this check is conducted by the email client, and we anticipated this discrepancy as a potential source of confusion.

C.5 Email Test Case: Sender is not Signer 2

This email is equal to the signer-vs-sender email but the signer uses a typo-domain bob@code-audil.org, instead of bob@code-audit.org (note the ‘l’ instead of ‘t’).

This test is motivated by the fact, that newer technologies such as Autocrypt, WKD, etc. may automatically import OpenPGP keys into the local key ring. Here, we test the phase *after* automatic inclusion. In other words, the key of bob@code-audil.org is trusted. To prevent that a participant spots this key in a prior test, this key is only contained in the key ring during the period of this test.

D GnuPG Key Ring

The GnuPG keyring contained the following trusted keys:

- Alice <alice@securepay24.de>
- Bob <bob@code-audit.org>
- Celine <celine@example.org>
- David <david@example.org>
- Ezra <ezra@code-audit.org>
- Farah <farah@example.org>
- Garrett <garrett@code-audit.org>
- Hoy <hoy@example.org>
- Iva <iva@example.org>
- Joon <joon@code-audit.org>
- Kemina <kemina@example.org>

Additionally, during runtime of the Sender is not Signer2 case, Bob <bob@code-audil.org> (note the “l”) was added as a trusted key to the key ring.

E Screenshot of the new Thunderbird interface for the Broken-Signature case

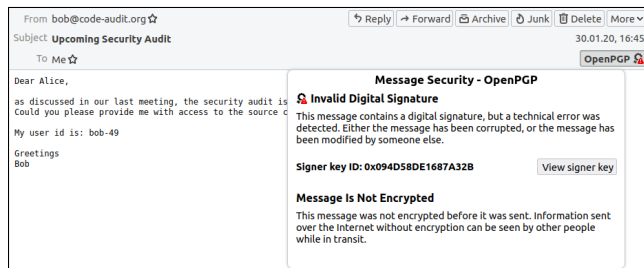


Figure 9: *Broken-Signature* email in Thunderbird 91.5.0.

F Additional Transition Graphs

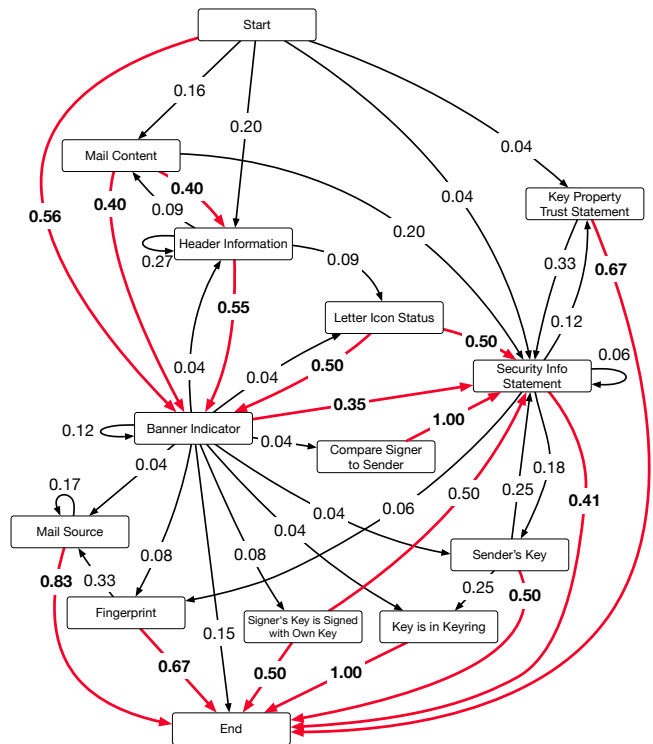


Figure 10: Overview of our participants’ transition probabilities from one check to another for the *Broken-Signature* email. The most likely transition after performing each of the checks is marked in red. Due to rounding the probabilities for each node might not add up to 100%.

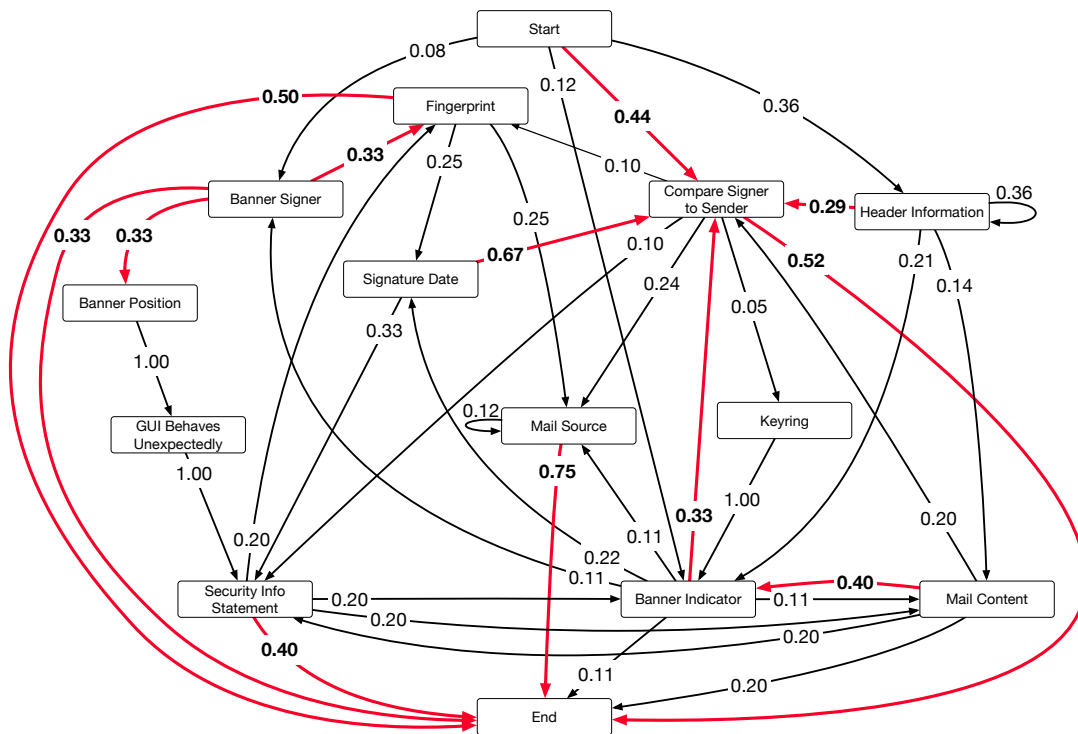


Figure 11: Overview of our participants' transition probabilities from one check to another for the *Conflicting-Signer* email. The most likely transition after performing each of the checks is marked in red. Due to rounding the probabilities for each node might not add up to 100%.

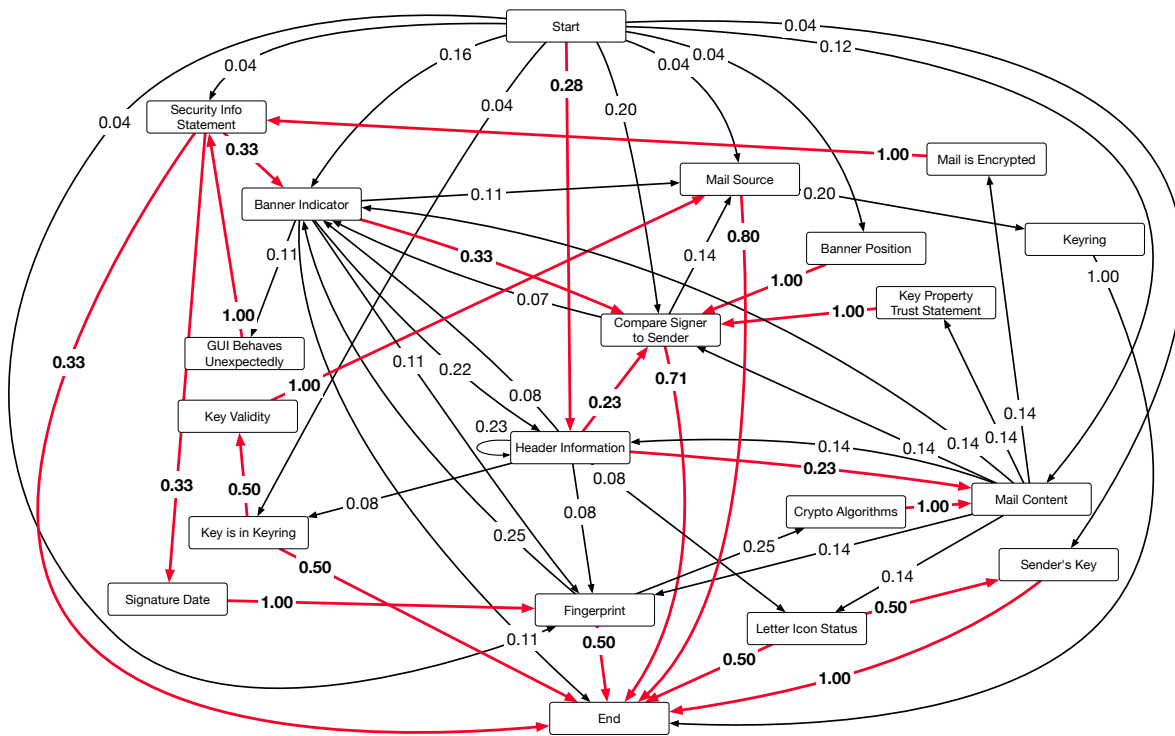


Figure 12: Overview of our participants' transition probabilities from one check to another for the *Conflicting-Signer-Subtle* email. The most likely transition after performing each of the checks is marked in red. Due to rounding the probabilities for each node might not add up to 100%.

Industrial practitioners' mental models of adversarial machine learning

Lukas Bieringer*
QuantPi

Kathrin Grosse*
University of Cagliari

Michael Backes
*CISPA Helmholtz Center
for Information Security*

Battista Biggio
*University of Cagliari,
Pluribus One*

Katharina Krombholz
*CISPA Helmholtz Center
for Information Security*

Abstract

Although machine learning is widely used in practice, little is known about practitioners' understanding of potential security challenges. In this work, we close this substantial gap and contribute a qualitative study focusing on developers' mental models of the machine learning pipeline and potentially vulnerable components. Similar studies have helped in other security fields to discover root causes or improve risk communication. Our study reveals two facets of practitioners' mental models of machine learning security. Firstly, practitioners often confuse machine learning security with threats and defences that are not directly related to machine learning. Secondly, in contrast to most academic research, our participants perceive security of machine learning as not solely related to individual models, but rather in the context of entire workflows that consist of multiple components. Jointly with our additional findings, these two facets provide a foundation to substantiate mental models for machine learning security and have implications for the integration of adversarial machine learning into corporate workflows, decreasing practitioners' reported uncertainty, and appropriate regulatory frameworks for machine learning security.

1 Introduction

Adversarial machine learning (AML) studies the reliability of learning based systems in the context of an adversary [5, 11, 67]. For example, tampering with some features often

suffices to change the classifier's outputs to a class chosen by the adversary [8, 23, 78]. Analogously, slightly altering the training data enables the attacker to decrease performance of the classifier [10, 70]. Another change in the training data allows the attacker to enforce a particular output class when a specified stimulus is present [19, 35]. Most state-of-the-art attacks and mitigations are in an ongoing arms race [4, 18, 80].

Although machine learning (ML) is increasingly used in industry, very little is known about ML security in practice. At the same time, previous works show that practitioners are concerned about AML [45, 58], and failures already occur [51], very little is known about ML security in practice. To tackle this question, we conduct a first study to explore mental models of AML. Mental models are relatively enduring, internal conceptual representations of external systems that originated in cognitive science [29]. In other security related areas, correct mental models have been found to ease the communication of security warnings [15] or enable users to implement security best-practices [79]. Mental models also serve to enable better interactions with a given system [85], or to design better user interfaces [28].

Our methodology builds upon these previous works by using qualitative methods to investigate the perception of vulnerabilities in ML applications. More concretely, we conducted 15 semi-structured interviews and drawing tasks with industrial practitioners from European start-ups and coded both drawings and the transcripts of the interviews. As the first work in this direction, we lay the foundations for practitioners' mental models of AML by describing two facets of these models. The first concerns the separation of ML related security (AML) and security unrelated to ML (non-AML security). In many cases, the borders between these two fields are blurry: a participant may start talking about evasion and finish the sentence with a reference to cryptographic keys. The second facet concerns the view of the ML model within a project. In contrast to the focus on an isolated model in AML research [4, 5, 11, 18, 67], our practitioners often describe one or more pipelines with potentially several applications of ML. Finally, we found more facets which are left for an in-depth

*First two authors contributed equally.

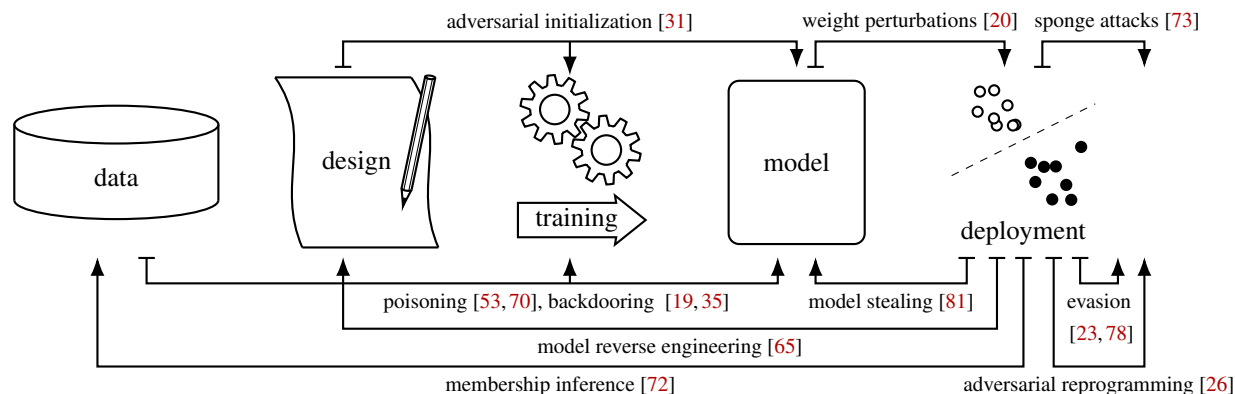


Figure 1: AML threats within the ML pipeline. Each attack is visualized as an arrow pointing from the step controlled to the point where the attack affects the pipeline.

investigation by future work. These include the application setting, prior education, and the perceived relevance of AML.

Our interviews showed that most of our participants lack an adequate and differentiated understanding to secure ML systems in production. At the same time, more than a third of our participants feels insecure about AML. These concerns seem justified as we found evidence for semi-automated fraud on ML systems in the wild. However, our findings have more practical implications. Our results allow us to address the current lack of understanding by (I) increasing awareness for AML and decreasing uncertainty about AML, (II) developing tools that help practitioners to assess and evaluate security of ML applications, and (III) drafting regulations that contain adequate security assessments and reduce insecurity about AML. However, more work is needed to understand the individual and shared mental models of practitioners and assess the real world security risks when applying ML.

2 Background and related work

In this section, we review related work on AML and recall different attacks that have recently been discussed. We also review literature on mental models with regard to human-computer interaction, usable security and ML.

2.1 Adversarial machine learning

AML studies the security of ML algorithms [5, 11, 67]. We attempt to give an informal overview of all attacks in AML, and additionally illustrate them in Figure 1.

Poisoning/backdooring. Early works in poisoning altered the training data [70] or labels [10] to decrease accuracy of the resulting classifier, for example SVM. For deep learning, due to the flexibility of the models, introducing backdoors is more common [19, 35]. Backdoors are chosen input patterns that reliably trigger a specified classification output. Defending such backdoors has led to an arms race [80].

Evasion/adversarial examples. Early work in evasion decreased the test-time accuracy of spam classification [23]. It was later shown that also more complex models change their output for small, malicious input perturbations [8, 78]. Albeit all classifiers are principally vulnerable towards evasion, recent works focus on the arms race in deep learning [4, 18].

Membership inference. After first inferring attributes [3] of the training data, research later showed that entire points can be leaked from a model [72]. More concretely, the attacker deduces, given the output of a trained ML model, whether a data record was part of the training data or not. As for other attacks, numerous defenses are being proposed [36, 62].

Model stealing. Tramèr et al. [81] recently introduced model stealing. During this attack, the attacker copies the ML model functionality without consent of the model’s owner. The attacker, given black box access to the original model, tries to reproduce a model with similar performance. As for the previous attacks, mitigations have been proposed [37, 66].

Weight perturbations. Fault tolerance of neural networks has long been studied in the ML community [16, 63]. Recently, maliciously altered weights are used to introduce a specific backdoor [34]. Few works exist to defend malicious change to the weights in general, not only related to backdoor introduction [76, 86].

For the sake of completeness, we conclude with a description of additional, recent attacks, some of which are part of our questionnaires (see Appendix D.3). In **adversarial initialization**, the initial weights of a neural network¹ are targeted to harm convergence or accuracy during training [31, 52]. In **adversarial reprogramming**, an input perturbation mask forces the classifier at test time to perform another classification task than originally intended [26]. For example, a cat/dog classifier is reprogrammed to classify digits. In **model reverse engineering**, crafted inputs allow to deduce from a trained model the usage of dropout and other architectural choices [65]. Fi-

¹Classifiers with convex optimization problems (for example SVM) cannot be targeted, as the mathematical solution to the learning problem does not depend on the initial weights.

nally, **sponge attacks** aim to increase energy consumption of the classifier at test time [73].

Practical Relevance of AML. In general, AML research has been criticized for the limited practical relevance of its threat models [27, 30]. A possible reason is our lack of knowledge about AI security in practice. Few works attempt to tackle this gap, including for example Lin and Biggio [51]. They give an overview about AI attacks that were carried out in practice based on AI related incidents covered in newspapers. Furthermore, Boenisch et al. [13] conducted a survey and developed an awareness score, which however encompasses AML, privacy, and non-AML security. Concerning which threats are relevant in practice in industry, Kumar et al. [45] and Mirsky et al. [58] found that practitioners are most concerned about model theft and poisoning. Yet, in academia, most work focused on evasion so far. To shed more light on AML in practice, we interview industrial practitioners and take a first step towards a theory of mental models of AML. To this end, we now introduce and review mental models.

2.2 Mental models

Mental models are relatively enduring and accessible, but limited, internal conceptual representations of external systems [25] that enable people to interact with given systems. Hence, the field of human computer interaction (HCI) studied this concept quite early [71]. Mental models, most recently, saw an increasing relevance in usable security. We now recall prior application scenarios and highlight relevant conceptual contributions in the context of security and ML.

Mental models in HCI and usable security. The relevance of mental models has been subject to a lengthy debate in HCI research [74, 83]. In many cases, the focus was to capture, depict and analyze mental models of specific objects of investigation. Examples of topics include, but are not limited to, the design of online search applications [6], interface design [42], and interfaces for blind people [24]. Research in usable security has recently focused on mental models of security in general [1, 85], privacy in general [69], security warnings [15], incident response [68], the internet [39], the design of security dashboards [56], the Tor anonymity network [28], privacy and security in smart homes [79, 88], encryption [87], HTTPS [43], and cryptocurrency systems [55].

With regard to the respective object of investigation, these contributions paved the way for improvements of user interface designs [28], adequate security communication [15], as well as the development of security policies and implementation of best-practices [79]. It has been argued that security mental models contain structural and functional properties [87]. For each application, users develop a cognitive representation of its inherent components, their interconnection and correspondingly possible security threats. This representation helps them to understand where threats could emerge and how they could take effect. Mental models evolve dynamically

upon individual interaction with a given application [12].

Mental models in ML. In order to interact with an ML application, humans need a mental model of how it combines evidence for prediction [64]. This is all the more important for ML-based applications which often inherit a certain opacity. As Lage et al. [46] pointed out, the number of necessary cognitive chunks is the most important type of complexity in order to understand applications. During interaction with black-box processes, humans strive for reduced complexity which may lead to the development of inaccurate or oversimplified mental models [32, 40].

A dedicated line of research therefore elaborates on the relevance and nature of mental models in the context of explainable artificial intelligence. Mental models have been found to serve as scaffolds not only for a given ML application [82], but also for its embedding in organizational practices [89]. For data science teams, these workflows usually consist of predefined steps (Figure 1) and necessitate interpersonal collaboration [60]. Following Arrieta et al. [2], we argue that individual collaborators within these teams (e.g., ML engineers, software engineers) develop separate internal representations of a given workflow or application. The need for appropriate mental models thereby increases with the enlarged scope of ML applications [47] and involved stakeholders [49, 77].

Recent work in this line of research called for qualitative studies at the intersection of the HCI and ML communities, to better understand the cognitive expectations practitioners have on ML systems [7, 40]. Suchlike studies seem all the more relevant as various industry initiatives propagate a human-centric approach to AI, explicitly referring to mental models.² However, the current scientific discourse lacks a dedicated consideration of cognition in AML. In order to fill this gap, we present the first qualitative study to elicit mental models of adversarial aspects in ML.

3 Methodology

This section describes the design of our semi-structured interviews, the drawing task, our recruiting strategy, the participants, and the data analysis. Our methodology was designed to investigate the perception of ML security and is, to the best of our knowledge, the first mental model study of AML.

3.1 Study design and procedure

To assess participants' perceptions, we conducted semi-structured interviews enriched with drawing tasks. We draw inspiration from recent work in usable security which also investigated mental models [43, 87].

Before the interview, participants were informed about the general purpose of our study and the applied privacy measures. We further assured each participant that their answers would

²e.g., <https://pair.withgoogle.com/chapter/mental-models/>

Table 1: Participants with their random IDs. Capital letters denote that participants work in the same company. We denote the application domain and the working experience (Exp.) in years. Knowledge in ML, Security and AML is encoded as completed lectures (++) , seminar/self-study (+) or none (.) .

ID	Company		Exp.	Education			Degree	
	Application domain			ML	Sec.	AML		
1		Human resources	7	++	+		PhD	
3	A	Healthcare	0.4				PhD	
4	B	Cybersecurity	8	++	+		PhD	
6	C	Business intelligence	15	++	++	+	PhD	
7		Computer vision	12	++			BSc	
9		Computer vision	9	++			MSc	
10		Cybersecurity	no questionnaire handed in					
11		Business intelligence	1	++			PhD	
12		Retail and commerce	1.4			++	PhD	
14		AI as a service	5	++		+	PhD	
15		Computer linguistics	5	+	+		MSc	
16	C	Business intelligence	3	++	+	+	PhD	
18	A	Healthcare	1.5	++			PhD	
19	B	Cybersecurity	15	++	++	+	MSc	
20	A	Healthcare	1.2	++			MSc	

not be judged. Participants were then instructed to complete a questionnaire on demographics, organizational background and a self-reflected familiarity with field-related concepts (Appendix D) before the interview. This questionnaire was filled with or without the authors’ presence. The answers have later been used to put participants’ perceptions in context to their organizational and individual background.

The threefold structure of our interviews covered 1) a specification of a given ML project a participant was involved in, 2) the underlying ML pipeline of this project and 3) possible security threats within the project. We chose this approach as the different attack vectors form part of the ML-pipeline as shown in Section 2.1. The detailed interview guideline can be found in Appendix C. As a last step of our interviews, we confronted the participants with exemplary attacker models for some of the threats considered relevant in industrial application of ML [45]. To assess practitioners’ understandings of these threats, study participants had to elaborate on these attack vectors within their specific setup (Appendix D.2).

To assess the participants’ knowledge about (A)ML in general, participants were asked to fill an additional questionnaire after the interview (Appendix D.3). In this questionnaire, we tested general knowledge in ML and independently asked for a self-reflected familiarity rating with some of the attacks we discussed in Section 2.1. This questionnaire was handed to the participants after the interview as to avoid priming.

We conducted one pilot interview to evaluate our study design. This first participant met all criteria of our target population in terms of employment, education and prior knowledge. As his explanations and drawings matched our expectations, we only added a specific question regarding the collaborators within a given ML-based project.

The average interview lasted 40 minutes and was jointly conducted by the first two authors of this paper between April and July 2020. To minimize interviewer biases, we equally distributed the interviews, where one author was the lead interviewer and the other took notes. Due to the COVID-19 pandemic, interviews were conducted remotely and relied on a freely available digital whiteboard³.

3.2 Recruitment

Recruitment for a study on applied ML in corporate environments presents a challenge, as only a small proportion of the overall population works with ML. Furthermore, the topic touches compliance and intellectual property of participating organizations. Hence, many companies are skeptical about the exchange with third parties. Consequently, many current contributions with industrial practitioners as study participants are conducted by corporate research groups (e.g., [33, 45]).

We tried to initiate interviews with two large multinational companies. Unfortunately, both denied our request after internal risk assessments. Therefore, we focused on smaller companies where we could present our research project directly to decision-makers and convince them to participate in our study. We relied on the authors’ networks (pilot participant, *P11*) and public databases for start-ups (more details in Appendix A) to find potential participants and used direct-messaging on LinkedIn and emails to get in contact.

Recruitment of study participants happened in parallel to interview conduction. Some participants forwarded our interview request to internal colleagues, so that we talked to multiple employees of some participating companies (see Table 1). We aimed to recruit experienced and knowledgeable participants and hence our requirements were a background in ML or computer science and positions such as data scientists, software engineers, product managers, or tech leads. We did not require any prior knowledge in security. After 8 interviews, no new topics (in our case for example new pipeline elements, whether defenses were mentioned, or how attacks were depicted in drawings) emerged. The research team thus agreed after 15 interviews that saturation was reached [14], and we stopped recruiting. The participants were randomly assigned an ID (a number between 1 and 20) which was used throughout our analysis. All participants were offered an euro 20 voucher as compensation for their time.

3.3 Participants

We summarize demographic information in Table 1. One participant, *P10*, did not hand in the questionnaire and is consequently not included in the following statistics. 14 participants identified as male, one identified as female, our sample is thus skewed towards males when considering ML practitioners [38]. As previous work found security perception of

³<https://awwapp.com/>

women and men to exhibit only some differences [57], this bias is acceptable for a first exploration but should be studied in depth in future work. Our participants had an average age of 34 years (standard deviation (STD) 4.27). As intended for a first exploration of practitioners' perception of AML, our sample covered various application domains and organizational roles which we now describe in detail.

Education and prior knowledge. The majority of participants (9 of 14) has a PhD, with all participants holding some academic degree. While our sample skews towards PhDs compared to the overall population of ML practitioners [38], previous work reports no correlation between overall education and security awareness [13]. Most participants (12 of 14) reported that they had attended lectures or seminars on ML. Roughly half (6 of 14) reported to have a similar background in security. To obtain a more objective measure we conducted a test about ML knowledge and asked participants to rate their familiarity with AML attacks (details in Appendix B). While we found that all participants were indeed knowledgeable in ML, we found that few attacks were well known to them.

Employment. Regarding the size of the companies, four participants worked in companies with less than ten employees, five in companies with less than 50 and the remaining six participants in companies with less than 200 employees. The companies' application areas were as diverse as healthcare, security, human resources, and others. Most participants were working in their current positions 6 years (STD 4.9). Their roles were diverse: Most (8 of 15) were in managing positions. Three were software or ML engineers, three more researchers. One of the participants stated to be both a researcher and a founder. One participant did not report his role.

Finally, we asked participants to report which goals were part of their companies' AI/ML checklist. Almost all participants (13 of 14) reported that performance mattered in their company. Half (7 of 14) stated that privacy was important. Slightly less than half (6 of 14) focused on explainability and security. Least participants (4 of 14) listed fairness as a goal in their products. To conclude, when interpreting these numbers, one should keep in mind that not all five goals apply equally to all application domains. Furthermore, our sample is too small to derive per area or per company insights, and we thus leave a detailed analysis for future work.

3.4 Data analysis

We adopted an inductive approach, where we followed recent work in social sciences and usable security that constructed theories based on qualitative data [43, 61]. To distill observable patterns in interview transcripts and drawings, we applied two rounds of open coding, e.g. we assigned one or several codes to sentences, words, or parts of the drawings. We then performed Strauss and Corbin's descriptive axial coding to group our data into categories and selective coding to relate these categories to our research questions [75]. Throughout

the coding process, we used analytic memos to keep track of thoughts about emerging themes. The final set of codes for interview transcripts and drawings is listed in Appendix E.

As a first step, the first two authors independently conducted open coding sentence by sentence and sketch by sketch. This allowed for the generation of new codes without predefined hypotheses. Afterwards, the resulting codes were discussed and the research team agreed on adding specific codes for text snippets relating to the confusion of standard security and AML. As a second step, two coders independently coded the data again. After all iterations of coding, conflicts were resolved and the codebook was adapted accordingly.

During axial coding, the obtained codes were grouped into categories. The first two authors independently came up with proposed categories which have then been discussed within an in-person meeting. While the grouping was undisputed for some of the categories presented in Appendix E (e.g. AML attacks, pipeline elements), for others the research team decided for (e.g. confusion, relevance) or against (e.g. type of ML model applied) the inclusion of a corresponding category only after detailed discussion. In addition, dedicated codes for the perception of participants (e.g. perceives AML as a feature, not a bug or security issue) were added to the codebook. Once the research team agreed on a final codebook, all transcripts and drawings were coded again using corresponding software.⁴ In doing so, we aimed for inferring contextual statements instead of singular entities.

The codes and categories served as a baseline for selective coding. Independently, the researchers came up with observations and proposals for specific mental models. Every proposal included a definition of the observation, related codes, exemplary quotes and drawings. The first two authors then met multiple times to discuss the observations and the corresponding relations of codes and categories. The resulting code tree contains 77 interview codes in 12 groups, 44 for drawings (in 5 groups), as depicted in Appendix E.

Over all interviews, the coders agreed on 989 codes while disagreeing on 136. Analogously, there were 275 codes on drawings in total, with 42 disagreements. We further calculated Cohen's kappa [22] to measure the level of agreement among the coders. For interview transcripts, we reached $\kappa = 0.71$; for the codes assigned to drawings $\kappa = 0.85$. These values indicate a good level of coding agreement since both values are greater than 0.61 [48]. Given the semi-technical nature of our codebook, we consider these values as substantial inter-coder agreement. Irrespective of this and in line with best practices in qualitative research, we believe that it is important to elaborate how and why disagreements in coding arose and disclose the insights gained from discussions about them. Each coder brought a unique perspective on the topic that contributed to a more complete picture. Due to the diverse background of our research team in AML, usable security and

⁴Available at <https://www.taguette.org/> and <https://www.maxqda.com/>.

economic geography, most conflicts arose regarding the relevance of technical and organizational elements of transcripts and drawings. These were resolved during conceptual and on-the-spot discussions within the research team.

3.5 Expectations of mental models

Given previous work on mental models and ML, we designed our study in a way that participants would first visualize their pipeline and later add corresponding attacks and defenses. For the pipeline, we expected that participants would name basic steps or components, such as data (collection), training, and testing. In general, we assumed participants' descriptions would vary in technical detail. Regarding AML, one of our motivations to conduct this study was to learn which knowledge our participants had. As a recent phenomenon, AML might not be known at all in practice, although practitioners might be aware of attacks relevant to their specific application. In particular, we did not expect practitioners to depict attacks using a starting and target point, as done in Figure 1.

3.6 Ethical considerations

The ethical review board of our university reviewed and approved our study design. We limited the collection of personal data as much as possible and used ID's for participants throughout the analysis. Since all participants were employed at existing companies and partially shared business-critical information, we aimed to avoid company-specific disclosures in this paper. Finally, we complied with both local privacy regulations and the general data protection regulation (GDPR).

4 Empirical results

In this section, we discuss our findings from the interviews and drawings. Given the unexplored nature of mental models of AML, we focus on two main facets, and discuss additional findings that require a more in depth analysis (in the sense of future work) at the end of this section.

The first of the two main facets is the (mingled) relationship between ML security (AML) and security unrelated to ML (non-AML security). We found that our participants, while not referring to AML and non-AML security interchangeably, still exhibited an often vague boundary between the two topics. The second facet concerns the view on ML as part of a larger workflow or product in industry, as opposed to the focus on an isolated model in academia. As a description of a high level workflow requires a high level perspective, we investigate whether it is equivalent to one, which we find not to be true. Afterwards, we then discuss potential facets requiring a more in depth investigation: the application setting, prior knowledge of the participant, and the perceived relevance of AML.

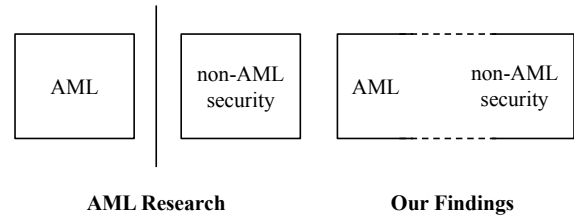


Figure 2: High-level intuition Section 4.1. While in research, non-AML security and AML are rather distinct, our participants do not always clearly distinguish the two fields.

4.1 Non-AML security and AML

Non-AML security deals with the protection against digital attacks in general. In our case, it encompasses topics like access control, cryptography, malicious code execution, etc. Non-AML security provides sound solutions by deploying defenses or implementing design choices. In AML, threats are much more connected with the functioning of ML. For many AML attacks, it is unclear which defenses work due to the ongoing arms-race. Although both topics are conceptually different, we found that our participants did not distinguish between security unrelated to ML and AML, as visualized in Figure 2. In our interviews, on the one hand, the boundary between non-AML security and AML often appeared blurry or unclear, with the corresponding concepts intertwined. On the other hand, there were crucial differences in the perception between non-AML security and AML threats. One difference is that whereas security defenses were often clearly stated as such, AML mitigations⁵ were often applied without security incentives. Finally, we find a tendency to not believe in AML threats. Many participants denied responsibility, doubted an attacker would benefit, or stated the attack does not exist in the wild. There was no such tendency in non-AML security.

4.1.1 Mingling AML and non-AML security

We first provide examples showing that non-AML security and AML were not distinguished by our participants. Afterwards, we investigate if non-AML security and AML are used interchangeably, by investigating the co-occurrence of codes.

Vagueness of the boundary between security and AML.

There are many examples for a vague boundary between non-AML security and AML. For example *P20* reasoned about evasion: “*this would require someone to exactly know how we deploy, right? and, where we deploy to, and which keys we use.*” At the beginning, the scenario seems unclear, but the reference to (cryptographic) keys or access tokens shows that the participant has moved to classical security. Analo-

⁵We are aware that AML is far from being solved, and communicated this to our participants if required. In this study, we define defenses as techniques which increase the difficulty for an attacker, like retraining or explainability.

gously, when *P18* reasoned about membership inference: “*but that could be only if you break in [...] if you login in to our computer and then do some data manipulation.*” Again, this participant was reasoning about failed access control as opposed to an AML attack via an API. Sometimes, ambiguity in naming confused our participants. For example, *P11* thought aloud: “*poisoning [...] the only way to install a backdoor into our models would be that we use python modules that are somewhat wicked or have a backdoor.*” In this case, the term ‘backdoor’ in our questionnaire caused a non-AML security mindset involving libraries in contrary to our original intention to query participants about neural network backdoors. The same reasoning can also be seen in *P11*’s drawing (compare Figure 3), where ‘backdoor’ points to python modules. Finally, *P12* stated: “*maybe the poisoning will be for the neural network. From our point of view you would have to get through the Google cloud infrastructure.*” From an AML perspective, the attack is carried out via data which is uploaded from the user. Yet, the infrastructure is perceived as an obstacle for the attack.

Correlations between non-AML security and AML attacks. In the previous paragraph, we showed that the boundaries between AML and non-AML security are blurred in our interviews. Another example is *P6* reasoning about IP loss: “*we are very much concerned I’d say the models themselves and the training data we have that is a concern if people steal that would be bad.*” In this case, it is left out how the attack is performed. Analogously, *P9* remarked: “*We could of course deploy our models on the Android phones but we don’t want anybody to steal our models.*” To investigate whether our participants are more concerned about some property or feature (data, IP, the model functionality) than about how it is stolen or harmed, we examined the co-occurrence of AML and non-AML security codes that refer to similar properties in our interviews. For example, the codes ‘model stealing’ and ‘code breach’ both describe a potential loss of the model (albeit the security version is broader). Both codes occur together six times, with ‘code breach’ being tagged one additional time. Furthermore, the code ‘model reverse engineering’, listed only two times, occurs both times with both ‘model stealing’ and ‘code breach’. However, not all cases are that clear. For example ‘membership inference’ and ‘data breach’ only occur together two times. The individual codes are more frequent, and were mentioned by three (‘membership inference’) and eleven (‘data breach’) participants. Analogously, attacks on availability (such as DDoS) in ML and non-AML security were only mentioned once together. Such availability attacks were brought up in an ML context twice, in non-AML security four times. Codes like ‘evasion’ and ‘poisoning’, in contrast, are not particularly related to any non-AML security concern. We conclude that AML and security are not interchangeable in our participants’ mental models to refer to attacks with a shared goal.

4.1.2 Differences between AML and non-AML security

In the previous subsection, we found that our participants did not distinguish non-AML security and AML. To show that this is not true in general, we now focus on the differences between the two topics. To this end, we start with the perception of defenses and then consider the overall perception of threats in AML and security. We conclude with a brief remark on the practical relevance of AML.

Defenses. Out of fifteen interviews, in thirteen some kind of defense or mitigation was mentioned; whereas all corresponding interviewees mentioned a non-AML security defense (encryption, passwords, sand-boxing, etc). An AML mitigation appeared in eight. In contrast to security defenses, however, AML defenses were often implemented as part of the pipeline, and not seen in relation to security or AML. As an example, *P9*, *P15*, and *P18* reported to have humans in the loop, however not for defensive purposes. *P10* and *P16* were aware that this makes an attack more difficult. For example, *P16* stated: “*maybe this poisoning of the data [...] is potentially more possible. There, we would have to manually check the data itself. We don’t [...] blindly trust feedback from the user.*” Analogous observations hold techniques like explainable models (3 participants apply, 1 on purpose) or retraining (2 apply, additional 2 as mitigation). For example, *P14* said: “*when we find high entropy in the confidences of the data [...] for those kind of specific ranges we send them back to the data sets to train a second version of the algorithm.*” In this case, retraining was used to improve the algorithm, not as a mitigation. We conclude that albeit no definite solution to vulnerability exists, many techniques that increase the difficulty for an attacker are implemented by our participants. At the same time, many practitioners are unaware which techniques potentially make an attack harder.

Perception of threats. There is also a huge difference in the perception of threats in non-AML security and AML. In security, threats were somewhat taken for granted. For example, *P9* was concerned about security of the server’s passwords “*because anybody can reverse-engineer or sniff it or something.*” Analogously, *P6* said to pay attention to “*the infrastructure so that means that the network the machines but also the application layer we need to look at libraries.*” On the other hand, almost a third of our participants (4 of 15) externalized responsibility for AML threats. For example, *P3* said their “*main vulnerability from that perspective would probably be more the client would be compromised.*” Analogously, *P1* remarked that ML security was a “*concern of the other teams.*” In both cases, the participants referred to another entity, and reasoned that they were not in charge to alleviate risks. Other reasons not to act include participants not having encountered an AML threat yet, and concluded AML was not relevant. More concretely, *P9* remarked: “*we also have a community feature where people can upload images. And there could be some issues where people could try*

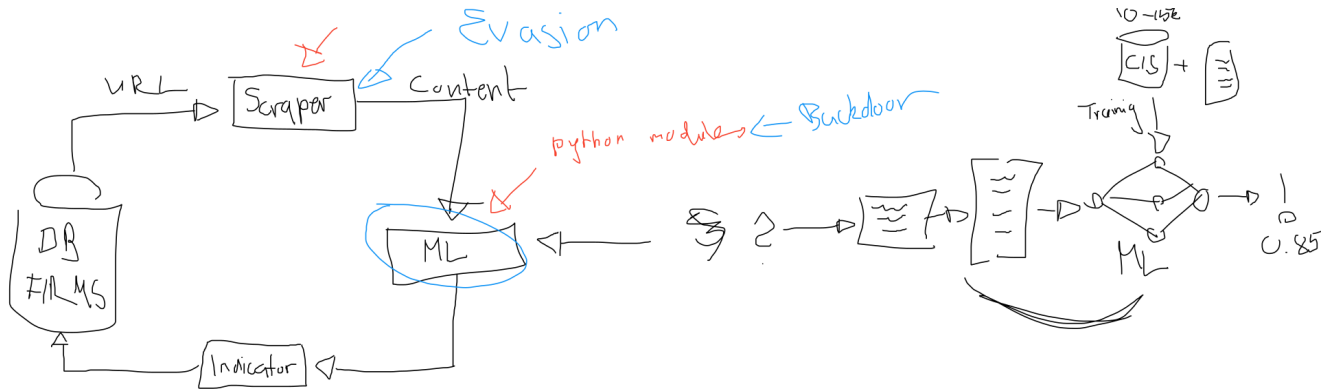


Figure 3: Drawing of P11. Red markings were added by the participant before, blue after being confronted with selected attacks.

to upload not safe or try to get around something. But we have not observed that much yet. So it's not really a concern, poisoning." Roughly half of the participants (7 of 15) reported to doubt the attackers' motivation or capabilities in the real world. For example, P1 said: "I have a hard time imagining right now in our use-cases what an attacker might gain from deploying such attacks." P20, who worked in the medical domain, stated: "I'm left thinking, like, why, what could you, achieve from that, by fooling our model. I'm not sure what the benefit is for whoever is trying to do that." Finally, many participants (9 of 15) believed that they have techniques in place which function as defenses. As an in-depth evaluation of which mitigations are effective in which setting is beyond the scope of this paper, we leave it for future work.

Practical relevance of AML. The fact that most participants did not consider AML threats relevant might be an expression of these threats being academic and not occurring in practice. Yet, our interviews showed that there are already variants of AML attacks in the wild. More concretely, P10 stated: "What we found is [...] common criminals doing semi-automated fraud using gaps in the AI or the processes, but they probably don't know what AML, like adversarial machine learning is and that they are doing that. So we have seen plenty of cases are intentional circumventions, we haven't quite seen like systematic scientific approaches to crime." Our participants lack of concern might then be an indicator that harmful AML attacks are (still) rare in practice.

4.1.3 Summary

We found that non-AML security and AML were mingled in our participants' mental models: the boundaries between the corresponding threats were often unclear. Yet, security and AML were not interchangeably used to refer to attacks with a shared goal. Furthermore, non-AML security threats were treated differently than AML threats: the latter were often considered less relevant. Whereas it remains an open question whether AML and non-AML security *should* be treated differently in practice, the fact that they are currently

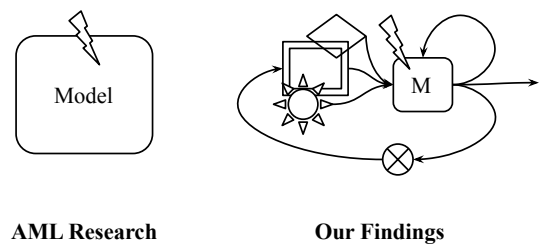


Figure 4: High-level intuition of Section 4.2. While AML research studies individual models, our participants often describe workflows with potentially several models, sometimes even the embedding system of the ML project.

poorly distinguished might due to low exposure to AML. At the same time, our interviews provided evidence for AML attacks in practice.

4.2 ML models and ML workflows

Many of our participants did not only refer to an ML model, but discussed a workflow or an entire system. This is in stark contrast to AML research, where models are often studied in isolation, possibly due to a lack of available data. This finding is visualized in Figure 4. In this subsection, we first discuss our participants view on ML models and the described systems. We then investigate whether such views are equivalent to a high level view on ML related projects, and conclude the section with a short discussion on some of our participants' struggles to assess threats at a high level.

4.2.1 Model versus system view

We first focus on the description of the ML model itself. Afterwards, we describe practitioners' views of ML models within larger systems and conclude the section with relating both findings to the technical level of abstraction.

ML model perspective. The general perception of the ML

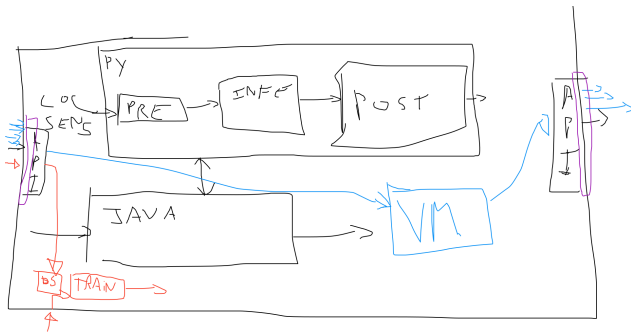


Figure 5: Drawing of *P16*. Colors were added after selected attack were presented to the participant. Red refers to evasion, purple to reverse engineering, blue to membership inference.

pipeline (Figure 1) seems to affect mainly the relevance of ML-models as such within the pipeline. More concretely, participants talked about models as pipeline components. Many (11 of 15) of our participants presented their projects in chronological order or with an implicit flow. Examples are visible in Figure 3 or Figure 6. Moreover, 6 out of 15 participants explained a pipeline not only as being composed by several steps, but remarked potentially several applications of ML within, or that several (different) pipelines exist. For example, *P14* reported that “the models are chained one after the other,” and *P7* stated that “we have both like unsupervised training and unsupervised training.” We conclude that often there is not a single model deployed, but data may be processed by several models, potentially in sequential order.

System perspective. Moreover, participants showed a strong focus on the surrounding or embedding of their ML-based project. In other words, not only the pipeline around the model was important, but also the surrounding infrastructure of the project. Out of 15 participants, 5 described their ML pipeline as a classifier as embedded into the larger project context (for example visible in Figure 3 or Figure 5). Related to this embedding, in two of the interviews, the topic of technical debt (or long-term maintenance) arose. In this context, *P6* stated: “how [...] we can also have to something that is maintainable in the long term.”

4.2.2 Technical abstraction level

The previous findings suggest a high level of technical abstraction in our interviews. While this is true on average, some (5 of 15) participants described their project minutely. For example *P12* described their application almost at the code level: “[...] we want to have for each node, that is basically the union of those two columns [...].” However, whereas the same participants also described their project as a workflow, they did not talk about the embedding of the project. On the other hand, *P18* remarked on their “supervising” (e.g., high level) perspective, yet provided no context. We conclude that our

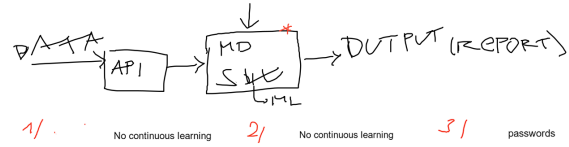


Figure 6: Drawing of *P18*. Red star indicates the most important component of the pipeline, not an attack.

sample does not allow conclusions about the level of technical abstraction and perspective on ML model, which is thus left for future work. We did find, however, that a high level perspective seemed to make threat assessment harder for at least some participants. Asked to specify a certain threat model, *P19* stated for example: “It’s like everywhere. Internal threats, external threats. Trying to mess with the communication, trying to mess if we model something.” In a similar manner, *P14* explained that an adversary could “try to put some pythons in non conforming ways to trigger networks.” Both descriptions are hard to interpret in technical terms, although both participants seemed aware of security threats in general. The same problem persists for defenses that our participants apply to encounter AML-specific security threats. *P18*, for example, first explained that “the countermeasures are all in the API.” After rechecking the documentation, the participant was able to provide further details on the applied defenses.

4.2.3 Summary

Our findings illustrate an important point which at the same time is very intuitive. Whereas most research papers focus on a single model when investigating ML security, in practice, models are trained and deployed in the context of other models or as components of larger workflows. At the same time, one pipeline may also contain several applications of ML. These views are not to be confused with the technical detail of a projects’ description. We furthermore find evidence that the right level of detail is crucial to providing useful information.

4.3 Additional facets of mental models

Eliciting mental models with only fifteen interviews seems ambitious, in particular in the context of a technique so versatile as ML. In the following, we thus discuss potential aspects of mental models that have to be studied in more depth in future work. These aspects include, but are not limited to the application setting, the effect of prior knowledge, and the perceived relevance of AML. We also found evidence of structural and functional components in our participant’s mental models. As the occurrence of these in AML mental models can be anticipated from prior work in mental models [87], we leave the corresponding discussion to Appendix F.

4.3.1 Application setting

Our sample is too small to make general statements about the application area. However, since almost a third (4 of 15) participants work in cybersecurity, we attempt to investigate whether working in security affects sensitivity to AML. Hence, we first divide the participants into security and non-security groups, starting with participants working in security-related fields. *P10*, who worked in a setting with cybersecurity reported: “*there is some standard AML attacks on ML you can use, but we design our system knowing that very well; on the other hand, we know that there is no perfect security, so, again defense is in monitoring and vigilance, but it’s not something that can be fully automated in our opinion.*” *P10* was in general very sensitive towards AML. *P4*, also from a cybersecurity setting, was less concerned about evasion: “*I can’t imagine yet how it can be applied for real life, for example [...] since we are pretty close on our development.*” Yet, *P4* also stated the need to gather more information about AML. Hence, also participants who worked in security-related areas had diverse mental models with respect to concrete attacks.

Participants from non-security fields have similarly diverse mental models. This diversity is also reflected in the drawings. *P11* (Figure 3) added some attacks (in red) before we provided explanations of evasion, backdooring and membership inference (added in blue). *P18* (Figure 6), on the other hand, did not add any threats in their drawing. Analogously, opinions also differ in the interviews; e.g., *P15* who worked in a non-security setting, was aware of security issues: “*one interesting thing of course is that the solution is in some ways constraint by adversarial security considerations so for example you cannot use natural language generation very much because of potential adversarial behavior.*” On the other hand, and confirming the drawing, *P18* reported that “*we do not really protect the machine learning part.*” Investigating the diversity of mental models induced by the application area in more depth is thus left for future work.

4.3.2 Prior knowledge

Another potential factor on a practitioner’s mental model is knowledge about or exposure to the topic at hand. However, we find no strong relation between education and capability or knowledge about AML in our sample. For example, one participant self-reported high knowledge in AML, but also stated: “*maybe the poisoning will be for the neural network.*” Here, a general attack, poisoning, is related to a specific model (neural networks). On the other end of the spectrum, *P9* did not self-report any knowledge about security or AML, but correctly remarked: “*Somebody could send us 100.000 images and collect all the results and try to build a model from that.*” We conclude that in our sample, self-reported prior knowledge is not related to AML knowledge. Yet, more work is needed to understand more in depth the complex relationship between exposure, education, and mental models of AML.

4.3.3 Perceived relevance of AML

Last but not least, we found little awareness of AML in our sample. As already discussed in section 4.1.2, this might be a consequence of little exposure to AML attacks in the wild. On the other hand, we found all levels of concern about AML in our sample. More concretely, a third of our the practitioners (5 of 15) did not mention AML at all before we explicitly asked. Another third reported that they were not very concerned about AML. For example, *P1* stated that evasion, or “*injecting malicious data to basically make the model [...] predict the wrong things*” was “*a concern that is not as high on my priority list.*” *P15*, analogously, said: “*mainly the machine learning pipeline this is the less critical security problem,*” reasoning that “*simply a performance would be unexpected.*” Yet, over a third (6 of 15) of the participants reported to feel insecure about AML when confronted with the topic. Of these six participants, two previously showed low priority on AML, and three did not mention AML at all. An example of insecurity is *P4*, who stated they needed “*some more research on it.*” Some participants, like *P19*, were concerned about specific attacks: “*I maybe need to learn more about this membership.*” In summary, some practitioners consider AML threats important, whereas some participants did not know AML well, and yet others did not consider it an important threat. From each of these three groups, there was at least one participant that felt not well informed. After the interviews (e.g., off the record) some participants stated that their awareness for AML had increased due to the interview. Many also inquired about defenses against specific threats, further confirming that they were indeed concerned about specific attacks.

5 Future work

Our findings expose the lack of knowledge about AML in practice, and thus show the need for additional research at the intersection of AML and cognitive science. In this section, we summarize these potential directions of future work. We first discuss theoretical research on mental models of AML and secondly more practical research that applies findings derived from mental models to AML.

5.1 A theory on mental models of AML

Our work is a first step to describe mental models of AML. For well-grounded mental models, more research is needed to investigate different aspects, as discussed in the previous section about the technical detail, application area and prior knowledge, for example. However, more research is also required concerning the development of mental models, and how a user based threat taxonomy (as opposed to a research based taxonomy) could look like.

Temporal involvement of mental models of AML. A better understanding about the development of individual mental

models could help to assess necessary steps to make practitioners take into account AML. In addition, research on how mental models are shared between various AI practitioners might help to implement adequate defenses within and across corporate workflows. Corresponding starting points can be found in cognitive science [59], where the convergence of mental models has been studied as a three-phase process of orientation, differentiation and integration [41].

Inherent threat taxonomies of mental models. Whereas academia has proposed clear threat models in ML security, it is unclear whether or to which degree these are also used or useful in practice. In this context, it could be interesting to consider existing taxonomies by Biggio et al. [9] and Barreno et al. [5]. These frameworks seem promising to investigate which specific structural elements practitioners consider relevant for specific attack vectors and how they perceive the causal evolution of these attacks. In line with recent work by Wang et al. [84], such user-centric attack taxonomies might help to understand practitioners' reasoning on AML.

5.2 Applying mental models to AML

Secondly, but not less important, is the question how AML research can benefit from the study of mental models and which problems could be tackled in this context. Examples include the usability of AML tools and libraries, a more realistic threat modelling in AML research as well as a general assessment of AML attacks in the wild.

Utility and usability of AML tools and libraries. We found that practitioners' mental models depend on available and provided information. Future research should therefore elaborate on the needed specificity of the available information. Furthermore, an evaluation of the available AML tools and libraries with regards to capabilities and needs of industrial practitioners might ease their usage across application domains. In line with recent work on fairness [50] and ethics [21], we consider this crucial for designing usable and accessible tools, corporate guidelines and regulations.

Practical threat modelling for AML research. As stated in Section 2, AML research has been criticized for the limited practical relevance of its threat models [27, 30]. Mental models could alleviate this issue in two ways. On the one hand, understanding which threats occur in which applications and how they are perceived helps to shift research towards designing practical and usable defenses. On the other hand, a deeper understanding of why non-AML security and AML are mingled allows us to adapt and improve current threat modelling. To this end, however, it is also important to know which threats need to be studied in the first place.

AML in the wild. Given the previous insight and evidence of semi-automated, ML-related fraud, a more detailed assessment of which attacks are conducted in the wild would be beneficial. Future work could investigate this with a focus on different groups of ML practitioners, including for example

ML engineers, auditors, and researchers, or dependant on the application. Furthermore, our work outlines that the model perspective usually taken in AML is of limited use in practice. More work is needed to study AML in the context of entire ML pipelines and end-to-end workflows.

6 Practical implications

Similar to Kumar et al. [45], we find that most of our participants lack an adequate and differentiated understanding to secure ML systems in production. Given that we found only reports of semi-automated fraud in our sample in Section 4.1.2, the absence of strong AML in practice might explain this lack of knowledge. Yet, as discussed in Section 4.3.3, 6 of 15 participants felt insecure about ML security. We thus now discuss the diverse implications of our study on how to tackle these insecurities and the overall lack of knowledge. We start with the question how to raise awareness for AML. Afterwards, discuss the implications of our findings for the embedding of AML in corporate workflows and finish with implications for regulatory frameworks of AML.

Raising awareness of and increasing confidence about AML. Although we did not ask about privacy specifically, the general data protection regulation was often mentioned by our participants. For example, *P6* stated: “*we are also subject to GDPR so we cannot just ignore the security aspects of the process.*” Like other participants (*P12*, *P18*), *P6* mentioned GDPR before we had asked about membership inference and thus privacy. Legislation might thus be a tool to increase awareness of AML. Independently, a third of our participants felt insecure about AML (Section 4.3.3). Given that several participants reported used software (*P9*, *P14*, for example “*TensorFlow*”), infrastructure (*P14*) or service provider (*P3*, *P12*, *P20*, for example “*Google*”), advertising tools to assess AML risks might be helpful for our participants. In particular as AML libraries⁶, but also overviews like the Adversarial ML Threat Matrix⁷ already exist. Our findings on the confusion between AML and non-AML security (Section 4.1.2) suggest these tools need to either enforce dedicated audits for both AML and non-AML security or combined countermeasures to address both areas jointly. Another solution to the feeling of insecurity, reported by our participants themselves (Section 4.3.3, *P19*: “*I maybe need to learn more about this membership*”), could be to provide materials for education.

Embedding AML into corporate workflows. Whereas academia generally studies AML with the perspective of an individual model, in practice, the entire ML pipeline and broader AI workflow need to be considered. As discussed in Section 4.2, in our interviews, for example *P6* and *P16* (see Figure 5) described the entire workflow of their AI application, whereas other participants focused on the ML pipelines

⁶For example the Adversarial Robustness Toolbox, CleverHans, RobustBench, or the SecML library, just to name a few.

⁷<https://github.com/mitre/advmthreatmatrix>

(for example *P18*, as visible in Figure 6). To successfully integrate AML into corporate workflows, however, more effort is needed. All actors working on an ML product need to be able to identify relevant and possible attacks and implementable defenses. Potential factors to consider here are for example different applications areas, as discussed in Section 4.3.1. Also the existing knowledge of the target audience should be considered, as the in Section 4.3.2 discussed variation of knowledge in our sample shows.

Creating appropriate regulatory and standardization frameworks for AML. Lastly, our study has implications for regulatory approaches that enable appropriate security assessments. The differences in application (Section 4.3.1) and prior knowledge (Section 4.3.2) we found imply that regulatory frameworks need to find a way to formally encompass these differences with regards to necessary security measures. The currently proposed ‘Legal Framework for AI’ by the European Commission, for example, differentiates certain types of ML applications of which some are prohibited or classified as high-risk and thus require a certain risk management. Furthermore, as discussed in Section 4.2, our results indicate that it is essential to communicate such frameworks at the right technical abstraction level to encompass both technical ML practitioners and non-technical stakeholders. Standardization efforts could incorporate this requirement by providing adequate information at multiple mental abstraction levels [17]. For example, recently proposed frameworks like the NIST Taxonomy and Terminology of AML⁸ explicitly lists references that might help practitioners develop more complex mental models. As mentioned above, a similar regulatory approach to privacy, the European general data protection regulation, had served as a scaffold for their privacy perception.

7 Limitations

We followed an inductive approach to investigate mental models through qualitative analysis. Hence, the data collected is self-reported and subjected to a coding process. We continued coding and refining codes until a good level of inter-coder agreement was reached. Nonetheless, all our findings are subject to interpretation and do not generalize beyond the sample, both of which is inherent to qualitative analyses. Finally, due to the COVID-19 pandemic, all interviews were conducted remotely and the interface limitations of the digital whiteboard might have impacted the participants’ sketches.

Given the qualitative approach and reached saturation, the small sample size of 15 is indeed acceptable [28, 87]. Due to the small sample size, however, several factors cannot be addresses in depth, as discussed in Section 4.3. Examples include, but are not limited to, the application setting and the perceived relevance. Ideally, future work provides a more in depth analysis of these topics in a larger quantitative study.

⁸<https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf>

All participants were employed at European organizations with <200 employees. This is due to the fact that while several multinational companies stated great interest in our research, they denied participation after internal risk assessments. As mental models of ML systems are always embedded in organizational practices [89], we strongly encourage future research to assess our findings within larger samples including more variety, for example academics, small and large companies. Given that previous work found differences in general security behavior depending on gender [57], and cultural background [44], we also strongly encourage a more in depth analysis of these aspects.

Furthermore, AML itself is a subject of study of which the perception evolves continuously. With an increasing awareness for security within applied machine learning, the findings presented can only be valid temporarily. Machine learning is applied in a wide range of settings. Consequently, not all attacks are relevant within each application domain. For example, a healthcare setting is subjected to other threats than a cybersecurity setting. For the sake of studying abstract facets of mental models, we did not consider the application in the present work. Yet, we would like to point out the necessity to study this aspect of AML in general.

8 Conclusion

Based on our semi-structured interviews with industrial practitioners, we take a first step towards a theory of mental models of AML. We described two facets of practitioners’ mental models and sketched more facets as an anchor for in-depth investigation by future work. These include the technical abstraction level, application setting, prior education, and the perceived relevance of AML. We provided more details on the first facet, or the blurry relationship between AML and non-AML security. These two topics were often mingled, yet not used interchangeably by our participants. The second facet can be understood as a first step to refined threat models in AML research. As apposed to a single model, our participants instead described workflows and relationships between potentially several ML models in a larger system context.

A clear understanding of the elicited mental models allows to improve information for practitioners and adjustments of corporate workflows. More concretely, our results help to raise awareness for AML, thus making practitioners feel less insecure. We further suggest that both application area and prior knowledge are considered when embedding AML into corporate workflows. Finally, regulatory frameworks might reduce uncertainty about AML and increase the awareness for possible AML threats. However, a wide range of subsequent research towards an encompassing theory of mental models in AML is still required. Last but not least, we are convinced that the AML community will benefit from further practical assessment of attacks in practice, as our work already provides evidence of semi-automated fraud in the wild.

Acknowledgements

The authors would like to thank Antoine Gautier, Michael Schilling, the anonymous reviewers and the shepherd for the insightful feedback. This work was supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0753) and by BMK, BMDW and the Province of Upper Austria within the COMET program managed by FFG in the COMET S3AI module.

References

- [1] Simon Anell, Lea Gröber, and Katharina Krombholz. End user and expert perceptions of threats and potential countermeasures. In *EuroS&P Workshops*, 2020.
- [2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 2020.
- [3] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10, 2015.
- [4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.
- [5] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *CCS*, 2006.
- [6] Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 1989.
- [7] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *FAccT*, 2020.
- [8] Battista Biggio, Iginio Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *ECML PKDD*, 2013.
- [9] Battista Biggio, Giorgio Fumera, and Fabio Roli. Pattern recognition systems under attack: Design issues and research challenges. *International Journal of Pattern Recognition and AI*, 28, 2014.
- [10] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In *ACML*, 2011.
- [11] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 2018.
- [12] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. ‘it’s reducing a human being to a percentage’ perceptions of justice in algorithmic decisions. In *CHI*, 2018.
- [13] Franziska Boenisch, Verena Batts, Nicolas Buchmann, and Maija Poikela. “i never thought about securing my machine learning systems”: A study of security and privacy awareness of machine learning practitioners. In *Mensch und Computer 2021*. 2021.
- [14] Glenn A Bowen. Naturalistic inquiry and the saturation concept: a research note. *Qualitative research*, 8, 2008.
- [15] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the gap in computer security warnings: A mental model approach. In *S&P*. Ieee, 2010.
- [16] Jakub Breier, Xiaolu Hou, Dirmanto Jap, Lei Ma, Shivam Bhasin, and Yang Liu. Practical fault attack on deep neural networks. In *CCS*, 2018.
- [17] David A Broniatowski et al. Psychological foundations of explainability and interpretability in artificial intelligence. *NIST: National Institute of Standards and Technology, US Department of Commerce*, 2021.
- [18] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, 2017.
- [19] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.
- [20] Nicholas Cheney, Martin Schrimpf, and Gabriel Kreiman. On the robustness of convolutional neural networks to internal architecture and weight perturbations. *arXiv preprint arXiv:1703.08245*, 2017.
- [21] Shruthi Sai Chivukula, Ziqing Li, Anne C Pivonka, Jingning Chen, and Colin M Gray. Surveying the landscape of ethics-focused design methods. *arXiv preprint arXiv:2102.08909*, 2021.

- [22] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20, 1960.
- [23] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *KDD*, 2004.
- [24] Hilko Donker, Palle Klante, and Peter Gorny. The design of auditory user interfaces for blind users. In *NordiCHI*, 2002.
- [25] James K Doyle and David N Ford. Mental models concepts for system dynamics research. *System dynamics review*, 14, 1998.
- [26] Gamaleldin F Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*, 2018.
- [27] Ivan Evtimov, Weidong Cui, Ece Kamar, Emre Kiciman, Tadayoshi Kohno, and Jerry Li. Security and machine learning in the real world. *arXiv preprint arXiv:2007.07205*, 2020.
- [28] Kevin Gallagher, Sameer Patil, and Nasir Memon. New me: Understanding expert and non-expert perceptions and usage of the tor anonymity network. In *SOUPS*, 2017.
- [29] Dedre Gentner and Albert L Stevens. *Mental Models*. Erlbaum, 1983.
- [30] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.
- [31] Kathrin Grosse, Thomas A Trost, Marius Mosbach, Michael Backes, and Dietrich Klakow. On the security relevance of initial weights in deep neural networks. In *ICANN*. Springer, 2020.
- [32] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. Can children understand machine learning concepts? the effect of uncovering black boxes. In *CHI*, 2019.
- [33] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *CHI*, 2019.
- [34] Yujie Ji, Xinyang Zhang, Shouling Ji, Xiapu Luo, and Ting Wang. Model-reuse attacks on deep learning systems. In *CCS*, 2018.
- [35] Yujie Ji, Xinyang Zhang, and Ting Wang. Backdoor attacks against learning systems. In *CNS*, 2017.
- [36] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *CCS*, 2019.
- [37] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. Prada: protecting against dnn model stealing attacks. In *EuroS&P*, 2019.
- [38] Kaggle. State of machine learning and data science. <https://storage.googleapis.com/kaggle-media/surveys/Kaggle's%20State%20of%20Machine%20Learning%20and%20Data%20Science%202021.pdf>, 2021.
- [39] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *SOUPS*, 2015.
- [40] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *CHI*, 2020.
- [41] Deanna M Kennedy and Sara A McComb. Merging internal and external processes: Examining the mental model convergence process through team communication. *Theoretical Issues in Ergonomics Science*, 11, 2010.
- [42] Julie Khaslavsky. Integrating culture into interface design. In *CHI*, 1998.
- [43] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel Von Zezschwitz. "if https were secure, i wouldn't need 2fa"-end user and administrator mental models of https. In *S&P*, 2019.
- [44] Hennie A Kruger, L Drevin, S Flowerday, and Tjaart Steyn. An assessment of the role of cultural factors in information security awareness. In *ISSA*, 2011.
- [45] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comisioneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *S&P Workshops*, 2020.
- [46] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *AAAI*, 2019.
- [47] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, 2016.

- [48] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, 1977.
- [49] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesting, and Kevin Baum. What do we want from explainable artificial intelligence (xai)?—a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296, 2021.
- [50] Michelle Seng Ah Lee and Jatinder Singh. The landscape and gaps in open source fairness toolkits. In *CHI*, 2021.
- [51] Hsiao-Ying Lin and Battista Biggio. Adversarial machine learning: Attacks from laboratories to the real world. *Computer*, 54, 2021.
- [52] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and SGD can reach them. *ICML*, 2019.
- [53] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *NDSS*, 2018.
- [54] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *ICCD*, 2017.
- [55] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, Edgar Weippl, and Katharina Krombholz. User mental models of cryptocurrency systems - a grounded theory approach. In *SOUPS*, 2020.
- [56] Janosch Maier, Arne Padmos, Mortaza S Bargh, and Wolfgang Würndl. Influence of mental models on the design of cyber security dashboards. In *VISIGRAPP (3: IVAPP)*, 2017.
- [57] Tanya McGill and Nik Thompson. Exploring potential gender differences in information security and privacy. *Information & Computer Security*, 2021.
- [58] Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Wenke Lee, Yuval Elovici, and Battista Biggio. The threat of offensive ai to organizations. *arXiv preprint arXiv:2106.15764*, 2021.
- [59] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management*, 36, 2010.
- [60] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. *Organization*, 1, 2022.
- [61] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong? a qualitative usability study. In *CCS*, 2017.
- [62] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *CCS*, 2018.
- [63] Chalapathy Neti, Michael H Schneider, and Eric D Young. Maximally fault tolerant neural networks. *Transactions on Neural Networks*, 3, 1992.
- [64] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *UIST*, 2018.
- [65] Seong Joon Oh, Max Augustin, Mario Fritz, and Bernt Schiele. Towards reverse-engineering black-box neural networks. In *ICLR*, 2018.
- [66] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction poisoning: Towards defenses against model stealing attacks. In *ICLR*, 2020.
- [67] Nicolas Papernot. A marauder’s map of security and privacy in machine learning: An overview of current and future research directions for making machine learning secure and private. In *Workshop on Artificial Intelligence and Security*, 2018.
- [68] Elissa M Redmiles. "should i worry?" a cross-cultural examination of account security incident response. In *S&P*, 2019.
- [69] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. Why doesn’t jane protect her privacy? In *PETS*. Springer, 2014.
- [70] Benjamin IP Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J Doug Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *IMC*, 2009.
- [71] Martina-Angela Sasse. How to t (r) ap users’ mental models. In *Human Factors in Information Technology*, volume 2. 1991.
- [72] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *S&P*, 2017.
- [73] Iliia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. *arXiv preprint arXiv:2006.03463*, 2020.

- [74] Nancy Staggers and Anthony F. Norcio. Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies*, 38, 1993.
- [75] Anselm Strauss and Juliet Corbin. *Basics of qualitative research*. Sage publications, 1990.
- [76] David Stutz, Nandhini Chandramoorthy, Matthias Hein, and Bernt Schiele. Bit error robustness for energy-efficient dnn accelerators. In *MLSys*, 2021.
- [77] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. *arXiv preprint arXiv:2101.09824*, 2021.
- [78] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.
- [79] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. "i don't own the data": End user perceptions of smart home device data practices and risks. In *SOUPS*, 2019.
- [80] Te Juin Lester Tan and Reza Shokri. Bypassing backdoor detection algorithms in deep learning. In *EuroS&P*, 2020.
- [81] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *USENIX*, 2016.
- [82] Jennifer Villareale and Jichen Zhu. Understanding mental models of ai through player-ai interaction. *arXiv preprint arXiv:2103.16168*, 2021.
- [83] Melanie Volkamer and Karen Renaud. Mental models—general introduction and review of their application to human-centred security. In *Number Theory and Cryptography*. Springer, 2013.
- [84] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *CHI*, 2019.
- [85] Rick Wash and Emilee Rader. Influencing mental models of security: a research agenda. In *New Security Paradigms Workshop*, 2011.
- [86] Tsui-Wei Weng, Pu Zhao, Sijia Liu, Pin-Yu Chen, Xue Lin, and Luca Daniel. Towards certificated model robustness against weight perturbations. In *AAAI*, 2020.
- [87] Justin Wu and Daniel Zappala. When is a tree really a truck? exploring mental models of encryption. In *SOUPS*, 2018.
- [88] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *SOUPS*, 2017.
- [89] Amy X Zhang, Michael Muller, and Dakuo Wang. How do data science workers collaborate? roles, workflows, and tools. *ACM on Human-Computer Interaction*, 4, 2020.

A Details on recruiting

We searched online databases like crunchbase⁹, AIhubs¹⁰, and lists with promising AI start-ups (for example the list by Forbes¹¹) to find potential participants.

B Participants' prior knowledge in (A)ML

To measure our participants' knowledge in ML, we constructed a questionnaire based on ML job interview questions¹²(Appendix D.3). Given that participants were not informed they had to take a test, we aimed to select a broad range of topics easy to query with multiple choice answers that were not too hard. The questionnaire had 8 questions, with the participants correctly answering on average 6.64 questions (STD 1.14). Guessing would yield an average of 2.66 correct questions. Thus, while we do not know how reliable our questionnaire estimates ML knowledge, we conclude that our participants are indeed knowledgeable in ML.

We also investigated the familiarity of our participants with AML attacks. To avoid priming, we asked participants to rate their familiarity after the interview. As sanity checks, we added two rather unknown terms, adversarial initialization [31] and neural trojans [54] (similar to backdoors). The results are depicted in Figure 7. Only one participant reported to be familiar with one attack (evasion). In general, most participants reported to have heard of most common attacks (evasion, poisoning, membership inference, and model stealing). As expected for the sanity check, adversarial initialization and neural trojans were largely unknown.

C Interview protocol

Thank you so much for taking the time to give us your perspective on security in machine learning. This study consists in III parts. Part I aims at exploring your role in ML-projects. Part II addresses the underlying machine learning pipeline. In

⁹<https://www.crunchbase.com/> for European companies operating in AI and having raised more than 1 million dollar funding

¹⁰<https://www.appliedai.de/hub/2020-ai-german-startup-landscape>

¹¹<https://www.forbes.com/sites/alanohnsman/2021/04/26/ai-50-americas-most-promising-artificial-intelligence-companies/?sh=653894c477cf>

¹²For example <https://www.springboard.com/blog/machine-learning-interview-questions/>

part III, we want to know how you perceive the security of machine learning. In part II and III, please visualize the topics (and relationships between them) that we ask you about. There are no rules, no wrong way to do it, and don't worry about spelling things perfectly. Nothing is off limits and you can use any feature of the digital whiteboard. After this last part, we will ask you about your knowledge about security of machine learning before this study.

Part I: Machine learning project

- Can you briefly describe what AI- or machine learning-based project you are currently involved in?
- Can you tell us a bit more about the goal of this project?
- Who else is involved in this project?
- What is your collaborators role in the project?

Part II: Machine learning pipeline

- What kind of pipeline do you currently apply within this machine learning based project?
- Which part of this pipeline is crucial for your business, or identical to your product?

Part III: Security within project and pipeline

- Is security something you regularly incorporate into your workflow?
- Have you encountered any issues relating to security in the projects you described?
- Where in the pipeline did these security-related issues originate?
- Can you specify the cause of the security-related issues?
- Can you specify how these security-related issues evolve in your pipeline?
- Which goal pursues an adversary with a such a threat?
- What is the security violation of the threat?
- How specific is the depicted threat?
- Are you aware of any further possible security threats in the scope of your project or pipeline?
- Which countermeasures do you implement against any of the aforementioned threats?

Thank you so much for taking the time to give us your perspective on security in machine learning.

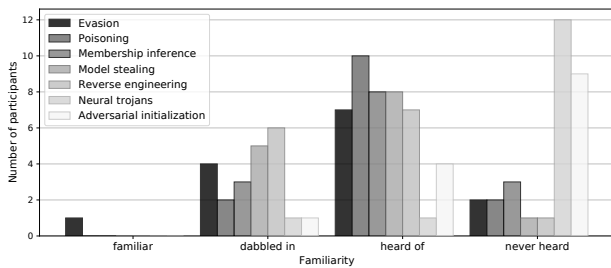


Figure 7: Self-reported familiarity of interviewed participants with different attacks on ML. Total of participants is 14, as one participant did not hand in questionnaire.

D Questionnaires

D.1 Demographics questionnaire

Thank you for participating in our research study about security in machine learning. Please take a couple of minutes to respond to the following questions.

- How old are you? _____
- What gender do you identify with?
 - male female _____
- What is your level of education? (please specify highest)
 - Highschool
 - Bachelor in _____
 - Master / Diploma in _____
 - Training / Apprenticeship in _____
 - PhD, area: _____
- What is your profession? _____
- What is your role in your team? _____
- How long have you been working in your current profession? _____
- What is the number of employees at your company/organization? _____
- What is the application domain of your product? _____
- Which of these goals are part of your organization's AI/ML-model checklist?
 - Explainability Fairness Privacy
 - Security Performance
- In which of these areas have you taken a lecture or intense course? Please add the title of the course if applicable.
 - Machine Learning _____
 - Security _____
 - Adversarial Machine Learning _____
- In which of these areas have you taken a seminar, or read up on? Please add the title of the seminar/book if applicable.
 - Machine Learning _____
 - Security _____
 - Adversarial Machine Learning _____

D.2 Attacks used in Interviews

Please read through the following selection of attack vectors and machine learning and explain whether you consider them relevant in your specific project. If yes, please add them to your sketch in a different color.

Evasion/ Adversarial Examples. This attack targets a model during deployment. The goal of the attacker is to fool the model: changing its output significantly by altering the input only slightly. An example is to change a picture containing a dog, present it to a cat-dog-classifier, and the model's output changes from dog to cat.

Poisoning. This attack targets the training or optimization phase of the model. The goal of the attacker is to either decrease accuracy significantly, or to install a backdoor. An

example is a cat-dog classifier that always classifies images containing a smiley as cat.

Privacy/ Membership Inference. This attack targets a model at test-time. The attacker's goal is to identify individual samples from or even the whole training set. An example is to measure the confidence on an input, as some algorithms tend to be more confident on data they have seen during training. Also over-fitting eases to determine what a classifier was trained on.

D.3 ML quiz

Please answer the following questions about ML. For each question, please tick **at least** one box.

Question 1. Which loss is used to train DNN?

- 0/1-loss.
- Cross-entropy loss.
- Hinge-loss.

Question 2. What is the difference between classification and regression?

- The kind of labels we fit: reals vs discrete classes.
- Regression is the name of classification in psychology / medical science.
- Regression is for discrete labels, classification for real valued ones.

Question 3. What is the difference between L_1 and L_2 regularization?

- L_1 yields sparser solutions.
- L_2 yields sparser solutions.
- none - they differ only in few practical applications.

Question 4. In the bias-variance trade-off, what does high variance imply?

- The analyzed data shows high variance.
- The clf is overly complex and potentially overfits.
- The data is likely to be classified fair (e.g., low bias).

Question 5. Why is Naive Bayes naive?

- Due to historic reasons.
- Due to the assumption that all features are independent.
- Because the application is simple and straightforward.

Question 6. What is cross-validation?

- Training on one task and then transferring the model to another task.
- Splitting the dataset and training/evaluating on different subsets.
- A method to reduce overfitting or choosing hyper-parameters.

Question 7. What are kernels in machine learning?

- Essentially similarity functions.
- A part of SVM, potentially yielding non-linear SVM.
- A specific instance of a similarity function used in SVM.

Question 8. What is pruning?

- Deletion of for example weights in a model.
- Deletion of specific points of the data.
- A technique to get a smaller from a large model with similar performance.

To conclude the study, we will ask you to rate your background knowledge on attacks *before* this study according to the following four classes:

Familiar. You are familiar with this concept, and can write down the mathematical formulation.

Dabbled in. You could explain in a five minute talk what the concept is about.

Heard of. You have heard of the concept and you could put it into context if necessary.

Never heard. You did not know about this concept before this survey.

For each concept, please tick **one** box. *The original questionnaire was formatted as table. To ease readability, we list them as questions here.*

Evasion / adversarial examples.

- familiar dabbled in heard of never heard

Poisoning / backdooring

- familiar dabbled in heard of never heard

Model stealing

- familiar dabbled in heard of never heard

Model reverse engineering

- familiar dabbled in heard of never heard

Neural trojans

- familiar dabbled in heard of never heard

Adversarial initialization

- familiar dabbled in heard of never heard

E Final set of codes

The final set of codes for the interviews is depicted in Table 2, the codes for the drawings in Table 3.

F Structural and functional components

We found structural and functional components in our participants' mental models. Structural components cover multiple, constituting entities that an individual perceives as relevant within a given application. In interaction with an ML system, functional components describe an individual's perception of the relations between the structural elements. As intended, the structure of our interview and drawing task (Appendix C) allowed to investigate these properties on the level of the ML pipeline, of the attack vectors as well as of the defenses.

F.1 ML pipeline

All participants distinguish clearly separable elements within their ML workflow. The specific composition of these steps

Table 2: Final set of codes for the interviews.

A. AML attacks A.1 poisoning A.2 evasion A.3 model stealing A.4 reverse engineering A.5 membership inference A.6 availability B. AML defenses B.1 retraining B.2 interpretability B.3 basic models B.4 ensemble B.5 human in the loop B.6 regularization B.7 own implementation B.8 on purpose C. security threats C.1 data capturing C.2 access C.3 data breach C.4 code breach C.5 libraries C.6 denial of service C.7 SDK C.8 customer	D. security defenses D.1 sandboxing D.2 access control D.3 development policy D.4 server register D.5 security testing D.6 data anonymization D.7 input data format restrictions E. pipeline elements E.1 training E.2 design E.3 model E.4 data E.5 data labelling E.6 data collection E.7 data preprocessing E.8 feature extraction E.9 testing E.10 deployment E.11 API E.12 database F. pipeline properties F.1 iterative F.2 several within project	G. organization G.1 ML role in project G.2 security role in project G.3 other role on project G.4 legal constraints G.5 technical dept of ML H. customer H.1 requirements H.2 privacy relevant data I. cloud I.1 used for security I.2 used but potential security risk I.3 not used because of security I.4 neutral J. relevance J.1 mentioning AML J.2 security low priority J.3 AML low priority J.4 encountered security issue K. confusion K.1 across ML attacks K.2 security and AML K.3 vagueness of concepts K.4 what security means	L. perception L.1 security externalized L.2 AML feature not bug L.3 doubting attacker L.4 believing defense is effective L.5 has not encountered threat L.6 attacks too specific L.7 insecurity about AML L.8 unspecific attack L.9 holistic attacker specificity L.10 pipeline specific defense L.11 importance of data L.12 high level perspective L.13 coding perspective
--	--	--	--

defines the structure of a certain ML pipeline. For two participants, this structure reflects the ML pipeline that we introduced in Figure 1. When asked to sketch the kind of pipeline applied, *P4* talked about “*data*”, “*training*”, “*testing*”, and “*visualization*”. We argue that these structural components serve as a scaffold for an individual’s mental model. Interestingly, the mental models of 12 out of 15 participants covered additional components that we did not expect prior to the study. The sketches of *P3*, *P7*, and *P11* (Figure 3), for example, contain explicit elements for data capturing. *P1*, *P9*, *P12*, as well as *P20* included dedicated elements representing a specific database to their drawing. Five participants also highlighted structural elements within the deployment environment during the interviews. *P14*, for example, specified on an API for deployment “*on several kinds of hardware architectures*”. Analogously, *P1* described an API that “*can be used to allow the user to interact with the models*” Hence, these structural elements concerning data and deployment seem to be of importance for the corresponding mental models. However, the perception of industrial practitioners does not only focus on these structural components but also covers functional aspects. *P6* for instance stated that his ML pipeline “*forks into a number of different directions and there are also interactions between the different components*”. In the corresponding sketch, multiple arrows within and across specific ML models indicate this interconnection of single components. Other drawings include this functional perspective through straight lines connecting the structural components, arrows connecting some of the structural components in a subsequent manner (e.g. *P14*), and arrows connecting all structural components in a subsequent manner (*P18* in Figure 6).

F.1.1 Attack vectors

The identified structural and functional components seem to be similarly relevant for mental models on attack vectors. For any kind of ML-specific threat, participants were able to precisely locate where they situated the corresponding, structural starting point. These have been specifically named during the interview and sketched via labelled arrows (e.g. Figure 3, *P11*), additional annotations (*P11*, *P15*), highlighted parts of potentially vulnerable pipeline components (e.g. Figure 8, *P10*) or as entire steps within a given ML workflow that have been marked as vulnerable (*P9*, *P20*). Strikingly, we saw a wide overlap in the perception of potential focal starting points for attack vectors. Study participants considered the model itself, the input of their ML pipeline, or the deployment environment to be particularly vulnerable. Figure 5 (*P16*) shows this for the latter. When confronted with poisoning and reverse engineering attacks, *P16* marked the input and output of his pipeline as possible starting points for threats (purple rectangles) and talked about how a competitor could “*screw our labeled dataset*” or a customer might “*ask a lot of questions to the API*”. However, the perception of attack vectors did also cover functional components. *P1*, for example, depicted the causal sequence of a “*data injection attack*” as three consecutive red arrows connecting different components of his ML pipeline. This is all the more relevant, as *P1* provided such a functional explanation and drawing for each of the attack vectors we presented to him. His mental models, hence, clearly seem to contain functional components. This is also the case for *P16*, who similarly provided explanations on the functional evolution of certain attacks within his

Table 3: Final set of codes for the drawings.

A. pipeline elements	B. pipeline properties	C. named explicitly	D. attacks	E. drawing
A.1 training	B.1 iterative	C.1 hardware	D.1 no attacks	E.1 boxes
A.2 design	B.2 linear	C.2 software	D.2 poisoning	E.2 symbols
A.3 model	B.3 abstracted	C.3 human	D.3 evasion	E.3 inner/outer
A.4 data	B.4 several	C.4 privacy sanitization	D.4 membership inference	E.4 flow within pipeline
A.5 data labelling	B.5 explainable	C.5 output	D.5 libraries	E.5 workflow embedding
A.6 data collection	B.6 MLaaS	C.6 classification	D.6 data collection	E.6 attacks graphical
A.7 data preprocessing		C.7 server	D.7 input/output	E.7 attacks words
A.8 feature extraction			D.8 unspecific attack	E.8 attacks causal
A.9 testing			D.9 defenses	E.9 attacks pointwise
A.10 deployment			D.10 exit points	
A.11 deployment environment			D.11 input points	

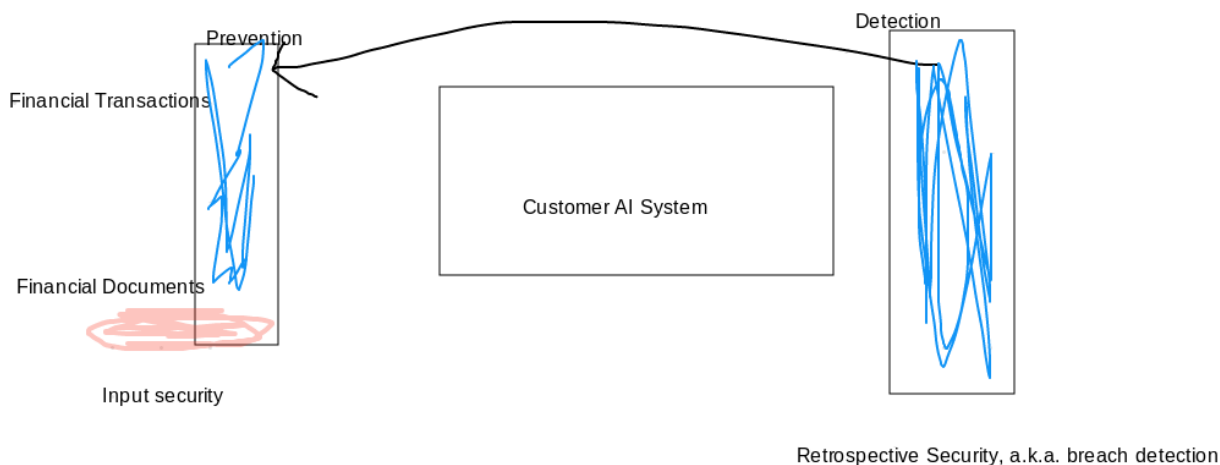


Figure 8: Drawing of *P10*. Important components of the workflow added in blue, possible starting points for attacks in red.

workflow and even added corresponding functional elements to his sketch (blue and red arrows in Figure 5).

F.1.2 Defenses

Although we found participants’ defenses explanations and sketches to be rather sparse, structural and functional properties are also relevant for the corresponding mental models. As visible in the sketch of *P18*, defenses are often thought of as structurally bound to specific components of a workflow/pipeline (Figure 6, *P18*). Data (*P14*), training (*P6*) and the models themselves (*P10*) have been specifically named as focal points for implementing defenses. In the case of defenses implemented at the model component, *P14* stated to “regularize in a way that makes it less sensitive to an adversary”. Hence, these implemented defenses are cognitively attached to the classifier as a focal pipeline component. However, security mental models also contain functional properties. In the case of human-in-the-loop-defenses, for example, *P14* stated to send certain classifications “back to the data sets to train a second version of the algorithm” if the output

confidence for certain data exhibited high entropy. This is depicted in the corresponding sketch by an arrow pointing from a rectangle with the caption “CPU” at the end of the pipeline to “raw data” (initial step of the pipeline). Similarly, *P7*, a participant working in video surveillance, explained the defense they had implemented to secure the transfer of input data (from cameras and on-site computers) into their pipeline: “This can only go out, never go in. [...] Nothing from the internet can connect to that server”. Industrial practitioners, hence, perceive defenses as containing functional components to unfold their full effect.

F.1.3 Summary

We conclude that mental models in AML contain of structural components which are cognitively put into (internal) relation. However, the specific unfolding of these internal conceptual representations seems to depend on the corresponding application and its underlying ML pipeline.

Replication: The Effect of Differential Privacy Communication on German Users' Comprehension and Data Sharing Attitudes

Patrick Kühtreiber
University of Göttingen

Viktoriya Pak
University of Göttingen

Delphine Reinhardt
University of Göttingen

Abstract

Differential privacy (DP) has become a standard for privacy-preserving data collection. However, there is little understanding of users' comprehension of this privacy technique, which could increase users' willingness to share personal data. Xiong et al.'s 2020 study tackles this problem by investigating the effect of differential privacy communication to laypeople, with an average of 466 participants per study primarily from USA and India. Since privacy decisions have been shown to depend on participants' culture in multiple past studies, we have replicated this study with German participants to compare the results with the original study and to gain further insights about differential privacy communication in a different cultural context. After having translated the original questionnaire into German, we conducted two studies with an average of 728 participants. While we could confirm that participants did not fully understand differential privacy and that a new method to communicate the effects of differential privacy is needed, participants in our study were more willing to share data than the participants from USA and India. This finding is surprising, as Germans have been shown to be more worried about their privacy than other cultures.

1 Introduction

The benefits of using personal data for machine learning are most prominent in healthcare applications [7, 9, 34]. Among ethical considerations, there are also privacy concerns [19] due to the fact that most applications require a lot of data to train the models. As data breaches appear to be ubiqui-

tous [16], many people are reluctant to share their private information [20, 37]. One of the key points of the European *General Data Protection Regulation (GDPR)* is that data subjects (i.e. individuals whose personal data are collected) must consent to the data processing [8]. It is therefore of major interest to investigate steps that allow data subjects to consent easily if their personal data are protected.

Among methods to protect privacy in such a context, DP is a promising solution to this problem. DP was introduced by Cynthia Dwork in 2006 [15] and it has since influenced many different areas of research, such as federated learning [39], data mining [18], and location-based services [2]. In principle, DP sets a statistical bound on the privacy risk of individuals who share their data. It does that by introducing carefully calibrated noise into the data, which masks the contribution of each individual data subject to a certain degree but still maintains the usability of the collected data, albeit sacrificing accuracy. The underlying promise of DP is that nothing about an individual in a dataset should be learnable that could not have been learned if the individual was not in the dataset [14].

Furthermore, the original model of DP has been extended to a more privacy-preserving model, referred to as *Local Differential Privacy (LDP)* [25]. In this model, data perturbation happens on the user's device (instead of a central entity with the original DP). As a result, the raw data do not leave the device, thus providing more privacy. However, since the noise is locally applied, the utility cannot be optimized by taking into account other users' data. In the following, we will refer to both models as (L)DP, if no distinction is necessary. Already used in practice by Google [17], Apple [43], and Microsoft [13], amongst others, (L)DP promises to be a solution to many problems faced in collecting data. However, it is not very well known outside of the technical and research communities, especially not to laypeople.

Laypeople may be reluctant to share information, though, because they fear for their privacy [20, 37]. Helping them to understand how their privacy is protected may help them to make informed decisions about sharing their data. However, only few publications [5, 11, 48] tackle this challenge. Among

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

them, the studies conducted by Xiong et al., presented in [48], investigate the effects of DP communication on the users' comprehension and their willingness to share personal data. While the authors tested many different and creative ways to explain DP, the studies have been conducted with young and educated participants who were recruited via Amazon MTurk, which has been shown to include mostly users from USA and India [12]. Nevertheless, it has been shown in [28] that cultural/age differences may impact the results. Also, replication studies have been shown to enhance the understanding of a certain subject [35] and clarify potentially false assumptions drawn from previous research [21].

To investigate these potential differences and validate the results in a different context, our contributions are as follows: We have replicated the original studies with participants from a different cultural and demographic background, directly compared self-reported and actual understanding of differential privacy, and evaluated whether personal health app usage impacts the willingness to share personal data. Tab. 1 illustrates the differences in our study compared to [48].

	Original study	Our study
Country	USA/India	Germany
Age	80% < 45y	Representative of the German population
Education	60% bachelor's degree	
#Experiments	4	2
Avg. #participants	~ 466	~ 728

Table 1: Differences from the original study.

As a result, we conducted two studies to (1) test the willingness to share low- and high-sensitivity data with a health app and its respective server depending on different text-based descriptions of (L)DP and (2) to evaluate the trust in and comprehension of these techniques. Similar to the original study, we only evaluated one description of DP or LDP respectively in the first study, while we evaluated eleven different descriptions in the second study.

The obtained key results are as follows.

1. We can confirm that the participants' attitudes are similar in both groups DP and LDP. Unlike originally expected, participants in the LDP group did not share more data with the app server than participants in the DP group, even though it is safer to do so under LDP.
2. Participants who were presented with a description that emphasizes the implications of the LDP, i.e., that privacy is protected even if the company's data base is breached, participants, indicated the largest willingness to share personal data, as in the original study.
3. The communication of (L)DP has a greater effect in our study compared to the original study. Participants whose privacy was protected via (L)DP wanted to share significantly more personal data than those in the control group where no privacy protection was communicated.

4. Overall, we experience a smaller variance in the results of the different descriptions of (L)DP as compared to [48]. Moreover, we find that there exists a correlation between participants who used health apps in their private life and their willingness to share data and their trust in the app, the server, and (L)DP.
5. As in the original study, our participants' comprehension of (L)DP was not very high; thus more effective communication methods are needed.

The remainder of this paper is structured as follows: We summarize the theoretical and technical background of (L)DP in Sec. 2 and present related work, including the original study in Sec. 3. We present our methodology in Sec. 4 and our experiments in Sec. 5 and Sec. 6. We discuss our results in Sec. 7 and make conclusions in Sec. 8.

2 Backgrounds on differential privacy

The primary assumption of (L)DP is that users send their personal data to a data curator, e.g., a company's data base. A data analyst can then analyze the data. (L)DP guarantees the users' privacy to a certain extent while keeping the data usable for the data analyst. However, one key element of (L)DP is that data analysts never see raw or perturbed data but only receive answers to queries of the noisy dataset. The thread model only considers attacks on the data curator, but not on the user's device itself.

2.1 DP vs. LDP

The global or centralized model is the original form of DP. In this model, users' raw data is sent to a trusted curator. Only then is the perturbation of the data carried out (see Fig. 1). Perturbation of the data in the global model takes place via noise that is added, e.g., from the Laplacian or the Gaussian distribution [14].

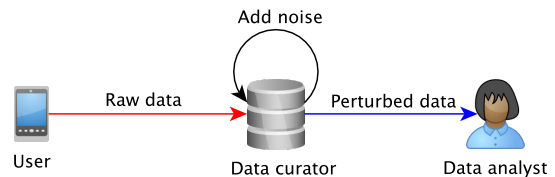


Figure 1: Differential Privacy

In the local model LDP, the data is perturbed on the device before it is being sent to the data curator. The privacy advantage in this case is that raw unperturbed data never leave the device (see Fig. 2). However, the accuracy of the data is lower, as the perturbation of data does not occur on data aggregates but on the data of single users. Perturbation is usually achieved via *randomized response (RR)* [46]. RR can be

best explained by imagining a scenario in which a participant has to answer a (sensitive) “Yes” or “No” question. However, before they answer, they first flip a coin. If it lands “heads” they answer truthfully, and if it lands “tails”, the participant flips the coin again and answers “Yes”, if it lands “heads” and “No”, if it lands “tails”. This way, there is a 25% chance of the answer being incorrect, thus providing plausible deniability to the participants and encouraging them to answer the questions truthfully (if the coin lands “heads”). Other than the previously described basic version of RR, one can also imagine biased coins or spinners representing the weights added to certain outcomes. This way, a data collector can emphasize privacy (by adding more weight to the randomized outcome) or accuracy (by increasing the weight of the true answer). Bullek et al. [5] conducted a study on biased spinners (see Sec. 3). The utility of LDP data is reduced by $O(\sqrt{N})$ compared to DP data, where N is the number of users [6]. In both cases, the data analyst receives only perturbed data.

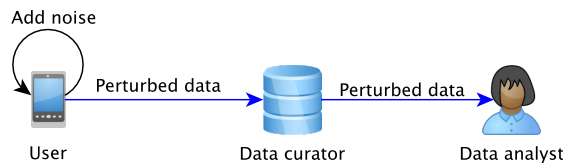


Figure 2: Local Differential Privacy

The most relevant fact for the data subject is that the privacy guarantee of LDP is higher than that of DP when considering only attacks on a company’s application server, for example, and not directly on the user’s device. This is because raw data never leaves the device and there is no centralized instance (like the trusted curator) you have to trust with your data.

3 Related work

In this section we first describe the impact of culture and privacy law on privacy attitudes, followed by relevant papers regarding usable (L)DP and the original study, which we replicate in this paper.

3.1 Cultural differences

There have been many studies investigating inter-cultural differences in regard to privacy. For example, studies have found that a country’s culture impacts its privacy regulations [31] and its citizens’ privacy regulation preferences [4]. Other studies focus on the difference in privacy attitudes in the context of digital government [10] or e-commerce adaption [32].

According to Hofstede’s cultural comparisons [1], Germany is one of the countries that avoid uncertainty, especially compared to the US or India. Also, Germany can be seen as an individualistic country, although the US scores higher in

this dimension. It has been shown that both dimensions, uncertainty avoidance and individualism, impact the risk-taking behavior of the country’s citizens regarding personal data. Citizens of collectivist countries as well as those from countries with a high uncertainty avoidance place more emphasis on privacy [44]. For example, Germans are more conservative when sharing data on online social networks [28] and trust providers of activity trackers less [22] when compared to US-Americans. Further studies have found that the medical history is seen more sensitive in the US, while income level is a little more sensitive for German participants [30, 40]. Moreover, Germans tend to feel less in control about the processing of their personal data [33]. However, none of the existing studies comparing cultures has focused on (L)DP.

3.2 Differences in privacy law

Privacy and data protection rights are perceived differently in the US and the EU. Whereas in the EU data privacy is seen as an individual right, in the US the right to privacy is not directly granted by the constitution and is context-dependent [3]. The different European privacy laws were harmonized in 2018 within the GDPR, which grants extensive data privacy rights to all EU citizens and heavily fines companies that do not comply. Since 2020 the *California Consumer Privacy Act (CCPA)* has granted people in California more extensive privacy rights as well, but its scope regarding individual privacy rights is still limited compared to the GDPR [3].

Early research shows that the existence of privacy regulations such as the GDPR can reduce data subjects’ privacy concerns [47]. However, more recent studies show that increased knowledge about these regulations does not yield the same result [36]. We can therefore assume that our sample — German citizens who are protected by the GDPR — might be more concerned about their privacy than the sample of the original study, which consisted mainly of US citizens.

3.3 Usable differential privacy

The first study concerning usable (L)DP was presented by Bullek et al. in 2017 [5]. This study focused on the participants’ understanding of RR, which is used in LDP (see Sec. 2.1). The participants were presented with three spinners that all had a different bias towards the true answer (40%, 60%, and 80%). That means, that a participant has a 40/60/80 percent chance of having to answer truthfully and a 60/40/20 percent chance that the answer is randomized (equally between “Yes” and “No”). To make this concept more accessible to laypeople, the authors designed (animated) spinners that would land on a certain field that would tell the participant how to answer the sensitive questions asked in the questionnaire. The study provided some seemingly contradictory results. As expected, participants preferred the spinner that provided the most amount of privacy; however, the sec-

Low-sensitivity	High-sensitivity
reason to use the health app	date of birth
exercise experience	family medical record
exercise time	substance use
gender	surgery record
height	diagnostic record
weight	income level
vegetarianism	current medication

Table 2: Low- and high-sensitivity questions

and most chosen one was the spinner that provided the least amount of privacy. Participants justified their choice of the least anonymous spinner by stating that it would otherwise feel like lying [5].

Another recent study in this area was conducted by Cummings et al. [11] and published while we were conducting the replication study presented in this paper. The goal was not only to evaluate the impact of DP communication on the willingness to share data but also how different DP explanations affect the users' expectations of DP. The authors synthesized 76 different DP descriptions into 6 short descriptions that all convey a certain theme (technique, trust, risk, etc.). The participants were presented with one of those descriptions and one of two relevant scenarios (disclosure of salary or medical records with DP). Being exposed to DP descriptions did raise the participants' privacy expectation; however, it did not increase their willingness to share data [11].

3.4 The original study by Xiong et al.

In the original study, Xiong et al. investigated effective communication of (L)DP and its impact on data-sharing decisions [48]. To this end, four experiments were conducted. These experiments consisted of online surveys, and their participants were recruited via Amazon MTurk.

3.4.1 Experiments 1 and 2

The participants in experiments 1 and 2 were presented with a scenario, in which they had to imagine downloading a health app that asks seven low-sensitivity and seven high-sensitivity questions (see Tab. 2). The participants did not actually have to provide these answers to the researchers, but instead had to answer how they would like their answers to be processed: 1.) not at all (opt out), 2.) only used by the app locally on the device (local only), or 3.) used by the app as well as the application server (both). To test the effect of (L)DP communication, participants in experiment 1 were randomly assigned to one of the four categories: DP, LDP, gain, and control. Participants in the DP and LDP groups were presented with a description of DP and LDP, respectively. The introduction to the questionnaire in the gain group was framed in a positive way (gain framing [45]), and the control group was presented with a neutral introduction. No descriptions of (L)DP or any

other data protection technique were presented to neither the gain nor the control group.

After confirming the effects of the gain framing, the authors repeated the experiment with different descriptions of (L)DP in experiment 2 (which was split into two separate surveys). The findings of experiments 1 and 2 suggest that (L)DP communication has little effect overall; however, there was an increase in sharing high-sensitivity questions. Contrary to the actual privacy guarantee, DP ranked higher than LDP which suggests that LDP was not well understood. In experiment 2, the authors tested further descriptions of (L)DP, which only confirmed the findings of experiment 1. Participants found DP easier to understand. However, when the description of LDP emphasized the data perturbation process, participants were more willing to share data with the app locally.

3.4.2 Experiment 3

In experiment 3, the authors examined the understanding of eleven different descriptions of (L)DP and also investigated the reasoning behind the participants' sharing decisions via open questions. The findings indicate that terms like "random" and "noise" are hard to understand. Participants were willing to share more information if the implication of the presented technique was also mentioned. As reasons to share data, participants noted that they had no privacy concerns, wanted to improve the utility of the app, or that they simply trust the presented (L)DP technique. Participants who did not want to share their data wrote that they distrusted the techniques, the requested data was too sensitive, data breaches could still occur, or that they distrusted the application or tech companies in general.

3.4.3 Experiment 4

Finally, experiment 4 investigated whether the self-reported understanding rates were accurate by asking five comprehension questions. Findings revealed that participants did not fully understand the implication of (L)DP in most cases. Only one description that emphasized the implications of LDP generated a high correct response rate for the implication-question. As a result, we used the existing studies as a basis for our work. However, our participants have a different cultural and demographic background, we have changed the number of studies, and we analyze whether personal health app usage affects the outcome. This way we increase the generalizability of the findings in [48] and are also able to compare self-reported and actual understanding of (L)DP.

4 Methodology

We started the replication study by translating the English questionnaire in [48] into German. Two of the authors translated the questions (and answers) independently of each other

and then discussed and resolved the differences. For example, some expressions like “health app” have a literal German translation that we only used when we agreed that it is more common than the English term.

Following the translation, the questionnaires were created in LimeSurvey and the participants were recruited via an ISO 29362-certified panel provider. All participants were financially rewarded if they completed the study. We set age and gender quotas to ensure a representative sample of the German population [42]. Our university does not have an official IRB process, but we adhered to ethical standards set by the German Research Foundation. All questionnaires have been approved by the university’s data protection officer.

4.1 Differences to the original study

We replicated the study conducted by Xiong et al. in order to compare the responses of different populations. However, we also made the following changes: (1) demographics as detailed in Sec. 4.1.1, (2) a reduced number of studies as detailed in Sec. 4.1.2, (3) the introduction of an additional question, and (4) correlation of the participants’ self-reported and actual understanding of (L)DP. Note that we also performed additional statistical tests in Sec. 5 and Sec. 6.

4.1.1 Demographics

The participants in [48] were recruited via Amazon MTurk. Xiong et al. did not ask where their participants were from; however, we know from other research that the majority of MTurk users are from the USA (75%) and India (16%) [12]. In comparison we focused on German participants only. Another major difference is the age and education of the participants. The original study is heavily skewed towards college educated (60% bachelors degree) younger people (80% younger than 45). Instead, we used quotas in our questionnaire to recruit participants that are representative of the German population, as illustrated in Tab. 3. We also asked the participants an additional question to see whether they are currently using a health app.

4.1.2 Study design

As depicted in Fig. 3, Xiong et al. conducted four experiments (excluding pilot studies and the division of the second experiment into two sub-experiments). As detailed in Sec. 3.4, experiments 1 and 2 used the same questionnaire with different descriptions of (L)DP and tested these descriptions on four different groups. We used the best descriptions found by the authors and used them in our experiment A, thereby compressing experiments 1 and 2 of the original study. Another difference is that Xiong et al. had already confirmed the effect of framing the questions in a positive way (gain framing [45]), which is why we used three different groups:

	Categories	Exp 1 (518)	Exp 2 (937)
Gender	Male	50.95%	53.1%
	Female	46.8%	45.8%
	Other	0.15%	0%
	No answer	2.1%	1.1%
Age	18-24	15.4%	15.3%
	25-34	30.1%	21.8%
	35-44	27.1%	24.3%
	45-54	15.6%	28.2%
	55 or older	10.1%	9.8%
	No answer	1.7%	0.6%
Education	No high school	23.2%	33.7%
	High school	39%	34.3%
	Bachelor	14.7%	12.1%
	Master	18.1%	16.3%
	PhD	1.7%	2.2%
	No answer	3.3%	1.4%
IT background	Yes	15.4%	16.1%
	No	79.5%	82.1%
	No answer	5.1%	1.8%
Health app	Yes	47.9%	54.3%
	No	49.6%	45%
	No answer	2.5%	0.7%

Table 3: Demographics

DP, LDP, and control, with control including the description of the gain framing of the original study. We examined the different descriptions of (L)DP in our experiment B, in which we not only asked for the participants’ self-reported understanding of the presented descriptions but also checked their comprehension with knowledge questions. Both of these are taken from experiments 3 and 4 of the original study and were originally separated. As a result, we can directly correlate self-reported understanding and actual comprehension of (L)DP. See Fig. 3 for our study design compared to the original study.

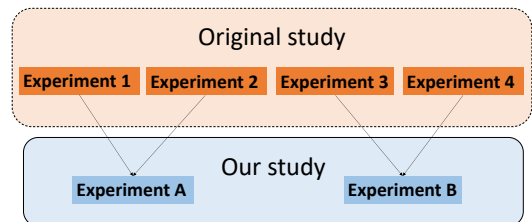


Figure 3: Study design

5 Experiment A

In this section we present the study design and the results of experiment A, before we discuss and compare the results with the ones from the original paper. The complete questionnaire for experiment A can be seen in Appendix A.

Group	Summary of description
DP	DP protects personal data via random noise added to aggregated data. Used by Harvard, US Census Bureau...
LDP	LDP protects personal data via random noise added to every answer provided by the user. Used by Apple, Google.

Table 4: Summary of (L)DP descriptions for experiment A

5.1 Study design

With changes detailed in Sec. 4.1, we conducted our first experiment. After the introduction in which participants were informed about the goal of the research, they first answered demographic questions in order to ensure the targeted quotas in terms of gender and age. The participants were then divided into three groups: DP, LDP, and control. Next, the participants were presented with the scenario in which they had to imagine themselves. In the described scenario, they had just downloaded a health app, that needed some partially sensitive information from them. All three groups were presented with the same introduction, i.e., the gain framing [48]. The DP and LDP groups were then presented with their respective descriptions of differential privacy. In Tab. 4 you can see the high-level summary of the descriptions, and the complete descriptions for experiment A are available in Appendix C. Afterwards the participants had to answer a comprehension question. If the question was not answered correctly, the description was shown again.

In the next step, we asked our participants the same questions as in [48], i.e., the participants' willingness to share potential answers to questions of the downloaded health app with the app or the app server. As presented in Tab. 2, the questions are separated into seven low- and seven high-sensitivity questions in order to evaluate the difference in the participants attitudes towards their willingness to share low- and high-sensitivity information with the health app or the app server. The participants did not answer those questions but only chose how they would like their potential answers to be processed. They could choose not to share anything (opt out), to trust their data only to the app locally, or to share them with the app and the app server. The participants could also choose not to answer. In that case they were counted in the opt out category, as in [48].

5.2 Participants

Through our certified panel provider, a total of 990 participants were recruited. This means that our three groups, DP, LDP, and control, comprised 330 participants each. We applied the same exclusion criteria to our participants as in [48]: (1) Completion time less than 120 seconds (57 DP, 46 LDP, 99 Control) and (2) wrong answers to the comprehension question (124 DP and 135 LDP). Consequently, 149 participants remained in the DP group, 138 participants in the LDP group, and 231 in the control group. The median completion

time (before exclusions) was 199.17 seconds in the DP group, 201.55 seconds in the LDP group, and 148.27 seconds in the control group.

5.3 Results

In the following, we report all significant results of our experiment that can be directly compared to the original study and additional tests. Note that the complete results are available in Appendix E.

5.3.1 Replication tests

Similar to the original study, we first performed χ^2 tests on the three relevant decisions (opt out, local only, or both) for each question type (low-sensitivity, high-sensitivity) collapsed across participants.

Question sensitivity across all participants: We observed significant differences between low and high question sensitivity across participants of all groups. Participants chose to *opt out* more often when they were asked a high-sensitivity question (29%) than when they were asked a low-sensitivity question (15%), $\chi^2_{(1)} = 217.63, p < .001$. We observed a similar attitude in the decision *local only*, with 37% for the high-sensitivity questions and 32% for the low-sensitivity questions, $\chi^2_{(1)} = 21.48, p < .001$. Consequently, the decision to share with *both* was higher for the low-sensitivity questions (53%) than for the high-sensitivity questions (34%), $\chi^2_{(1)} = 280.5, p < .001$. This means that our participants chose to share low-sensitivity questions more often (locally and with the app server) than high-sensitivity questions.

Question sensitivity among groups: Differences among groups (control vs. DP vs. LDP) could only be observed for the decisions *opt out* and *both*. The decision rate to *opt out* was significantly larger in the control group (28%) than in the DP (16%) and LDP (17%) groups, $\chi^2_{(2)} = 139.21, p < .001$. In contrast, the decision rate to share with *both* was higher for the DP (49%) and LDP (48%) groups compared to the control group (37%) $\chi^2_{(2)} = 97.95, p < .001$. Post-hoc independent sample t-tests reveal that only the differences Control vs. DP and Control vs. LDP are significant ($p < .001$ for all four tests, Bonferroni corrected). These results indicate that (L)DP communication has the effect of increased willingness to share data. However, almost no difference between DP and LDP could be observed.

Two-way interaction of sensitivity \times condition: Finally, we replicated the 2×3 cross-table *question sensitivity (low, high) \times group (control, DP, LDP)* to perform χ^2 tests on this matrix (see Fig. 4). Again, only the decisions to *opt out* ($\chi^2_{(2)} = 8.08, p = .018$) and to share with *both* ($\chi^2_{(2)} = 9.94, p = .007$) are significant. Pairwise tests reveal that only Control vs. DP and Control vs. LDP show significant differences. For the low-sensitivity questions, only the decision to *opt out* is statistically significant for the pairs Control vs.

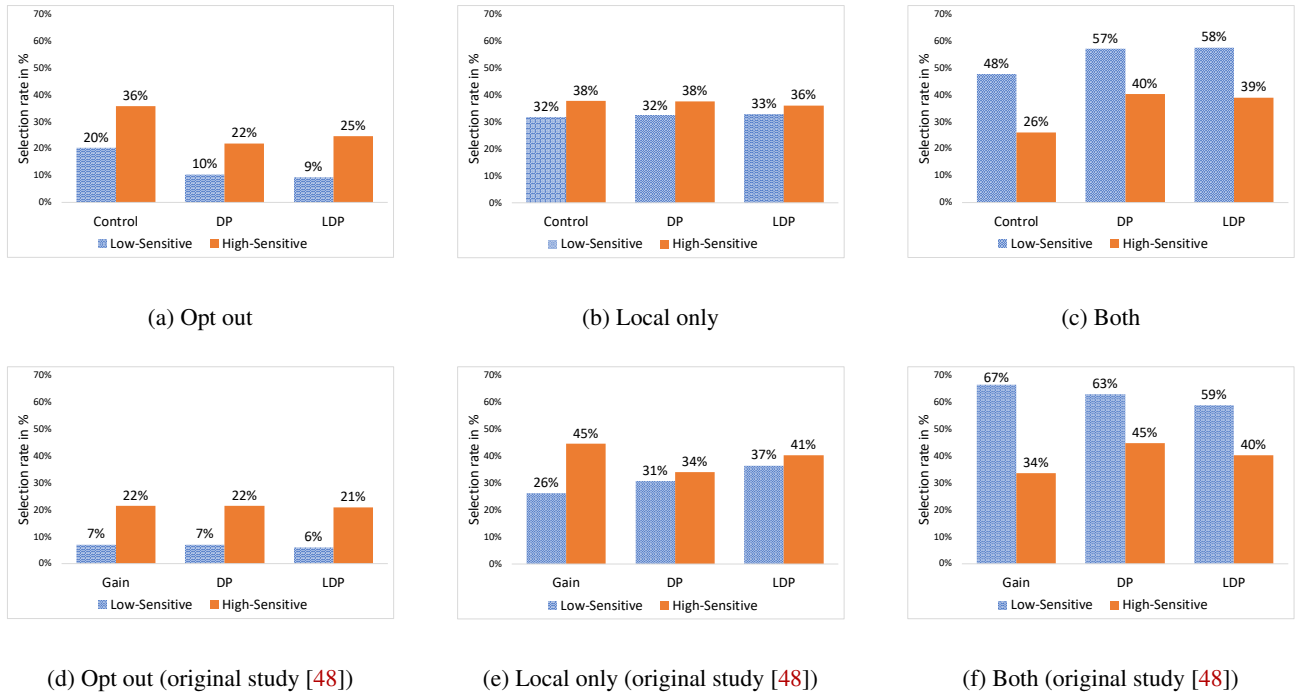


Figure 4: Selection rates across the three different conditions and both question sensitivities for experiment A (a-c) in comparison with the selection rates of the original study’s experiment 1 (d-f). As we used the gain framing in our control group, we compare the rates of our control group with the gain condition of the original study.

DP ($p = .002$) and Control vs. LDP ($p = .001$, Bonferroni corrected). For the high-sensitivity questions, there are significant differences in the two decision rates *opt out* (Control vs. DP, $p < .001$, Control vs. LDP, $p = .001$, Bonferroni corrected) and *both* (Control vs. DP, $p < .001$, Control vs. LDP, $p = .004$, Bonferroni corrected). This further confirms that (L)DP communication had a positive effect on data sharing and that participants in both (L)DP groups show little difference in their willingness to share.

5.3.2 Additional tests

In addition to the tests carried out by Xiong et al., we also tested if we could observe differences in the participants’ trust in the app, the app server, and (L)DP and their willingness to share based on their demographics. To this end, we performed Kruskal-Wallis tests and we only report the significant results. The complete statistics are available in Tab. 9 and 10 in Appendix E.

Trust: Across all three groups, only participants who were already using a health app show a significant difference in the trust in the app ($H(1) = 40.028$, $p < .001$), the server ($H(1) = 27.362$, $p < .001$), and (L)DP ($H(1) = 26.31$, $p < .001$). For example, 33% of participants who already used a health app agreed at least somewhat with the statement that they trusted that (L)DP was secure, in contrast to only 14% of those that

did not use a health app. On the other hand, 26% of those who did not use a health app distrusted the app (somewhat disagree or lower) with their private information, whereas only 7% of health app users said the same. There was no difference in trusting the app or the server among the three conditions as well as no difference in trust in (L)DP between the groups DP and LDP.

Willingness to share: There are differences in gender when participants report their willingness to share. Female participants share more with *local only* (low-sensitivity: $\chi^2_{(28)} = 20.50$, $p = .005$; high-sensitivity: $\chi^2_{(28)} = 25.76$, $p = .001$), while male participants share more with *both* (low: $\chi^2_{(28)} = 16.33$, $p = .022$; high: $\chi^2_{(28)} = 18.85$, $p = .009$). The usage of health apps stands out again, as participants who used health apps decided more often to *opt out* (low: $\chi^2_{(7)} = 15.11$, $p = .035$; high: $\chi^2_{(7)} = 19.23$, $p = .007$) and to share with *both* (low: $\chi^2_{(7)} = 15.33$, $p = .032$; high: $\chi^2_{(7)} = 32.59$, $p < .001$). Further significant results are correlations between age and the decision to *opt out* ($\chi^2_{(28)} = 56.74$, $p = .001$) and to share with *local only* ($\chi^2_{(28)} = 56.25$, $p = .001$) for the high-sensitivity questions and to share the low-sensitivity questions with *both* ($\chi^2_{(28)} = 49.84$, $p = .007$). Also, participants who reported an IT background were significantly more willing to share high-sensitivity questions with *both*. ($\chi^2_{(7)} = 18.38$, $p = .010$)

5.4 Comparison and discussion

As in the original study, there was hardly any difference in the participants' willingness to share information between the DP and the LDP group. It could be expected that people share more with *both* under the LDP condition and share more *local only* with the DP condition. However, both conditions led participants to share more with *both* and to *opt out* less in very similar rates. We could confirm that the question sensitivity is significant across all three groups.

The major difference between this experiment A and the original study's experiments 1+2 is that in our case the communication of (L)DP had a significant effect on the participants' willingness to share, especially when looking at high-sensitivity questions. However, there was hardly any difference in the *local only* decision across the three groups, which suggests an "all or nothing" mindset of our participants.

Participants showed more trust in the app, the app-server, and (L)DP when they were already using health apps, which is in line with findings in [4], and more willingness to share if they had an IT background.

6 Experiment B

Here, we present the study design and the results of experiment B before we again discuss and compare the results with the ones from the original paper. The complete questionnaire for experiment B can be seen in Appendix B.

6.1 Study design

For our second experiment, we combined the last two experiments of the original study into one LimeSurvey questionnaire (see Fig. 3). By doing so, we could directly compare the self-reported understanding of (L)DP with the comprehension questions, while they were separated in the original study. This also allowed us to test all 11 descriptions of (L)DP, which also provides additional results compared to the original study's experiment 4. We first asked for the participants' demographics and then presented one of the 11 (L)DP descriptions provided by [48]. A high-level summary of these descriptions can be seen in Tab. 5, while the complete descriptions are available in Appendix D. After the participants' introduction to (L)DP, questions regarding trust and comprehension were asked. In the following, we present a short description of the questions, while the full questionnaire is available in Appendix B.

- (Q1) Do you want to share personal data with the app server given the presented data protection technique?
- (Q2) Why? / Why not? (open question depending on the answer to the previous question)
- (Q3) The description of (L)DP was understandable. (7-point Likert scale)

(Q4) Please highlight the words you did not understand (based on a score < 4 on the previous question, participants could highlight words by clicking on them)

(Q5) Comprehension questions

- C1. Can an attacker see your data if they get access to the data base?
- C2. Can employees see your data?
- C3. Can third-party companies see your personal data?
- C4. The usability of the data is now ... when the presented data protection technique is in place (better/worse/the same)
- C5. Do the data stay useful for third-party companies?

The comprehension questions in Q5 were presented in random order. Participants also had the choice not to answer or to select that they were unsure.

Group	Summary of description
<i>LDP Flow</i>	Answers are changed before they are sent to the company. Focus on the <i>flow</i> of data.
<i>DP Flow</i>	Answers are sent to the company's data base; others only receive changed answers to queries.
<i>US Census</i>	DP introduces controlled noise into the data, personal information is protected.
<i>Google</i>	LDP guarantees users' privacy as with random coin tosses.
<i>Apple</i>	DP transforms the data before they leave the device; true data cannot be reproduced.
<i>Uber</i>	DP allows statistical analyses without revealing information about individuals.
<i>Microsoft</i>	DP allows privacy-preserving data analysis by introducing inaccuracies into the analyzes.
<i>DP Imp</i>	DP uses only a modified version of your data. Personal information is not protected if the data base is compromised. Focus on DP's <i>implication</i> on the data.
<i>LDP Imp. w/o Local</i>	DP changes your data on the app randomly before they are sent to the server. Privacy is protected if the data base is compromised. No mention of the word "local".
<i>LDP Imp</i>	LDP changes your data on the app randomly before they are sent to the server. Privacy is protected if the data base is compromised.
<i>LDP Comp</i>	LDP introduces random noise to raw data before they are sent to the server. Used by Google, Apple. Includes <i>company</i> names.

Table 5: Summary of (L)DP descriptions for experiment B

6.2 Participants

As in experiment A, we used a between-subjects factorial design for our questionnaires. Participants were divided into 11 groups. In each of these groups, a different German description of (L)DP was presented to the participants. We excluded

203 participants who did not want to answer the question of whether they want to share data (Q1), 67 participants who gave nonsensical responses to the open question why they did or did not want to share data (Q2), and 31 with a completion time of less than 60 seconds.

6.3 Results

Here, we present our results of experiment B, first starting with the replication tests and followed by our additional tests.

6.3.1 Replication tests

Willingness to share: Across all 11 groups, 53% wanted to share sensitive information (Q1) with the application server. The LDP Imp group had the largest sharing rate with 60%, and DP Flow had the lowest sharing rate of 47%, see Fig. 5.

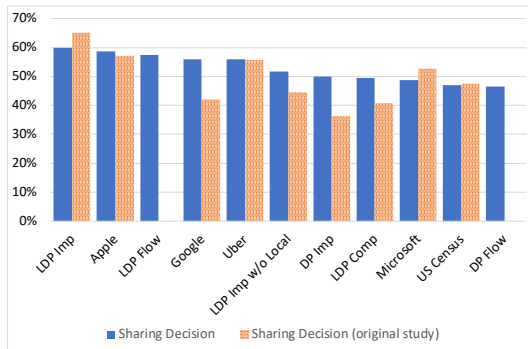


Figure 5: Sharing decision rates compared to the original study’s experiment 3 [48]

Comprehension: Across all groups, only 13% indicated an easy-to-comprehend rating of less than 4 (on a 7-point Likert scale), which means that our participants were confident in their understanding of (L)DP. Participants in the LDP Imp w/o Local group showed the highest self-reported comprehension (M=5.3, SD=1.3), followed by Apple (M=5.3, SD=1.5), DP Imp (M=5.3, SD=1.3), and Uber (M=5.2, SD=1.3). Participants in the DP Flow group report the lowest understanding (M=4.8, SD=1.4), see Fig. 6.

Participants who indicated that the description of (L)DP was not understandable (score of less than 4 in Q3) could highlight the words that were less understandable (Q4). Across all groups, the most selected words were “differential” (22), “privacy” (21), “poise” (18), and “introduces” in combination with “controlled” (9). The correct response rates of the comprehension questions (Q5) are very low throughout all groups (see Tab. 6). Participants of all groups were able to answer correctly more often than 50% on average only for the question *C3 3rd party* when they were asked about the utility

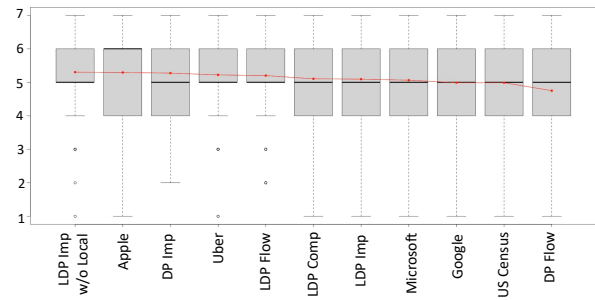


Figure 6: Self-reported easy-to-comprehend rating on a 7-point Likert scale sorted decreasing by mean (red dots).

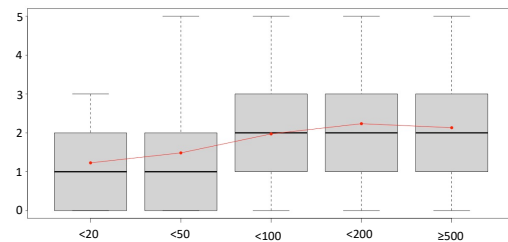


Figure 7: Time spent looking at the (L)DP descriptions in seconds vs. average number of correct responses on the comprehension questions in Q5. The red dots indicate the respective means.

of the perturbed data for third parties. Participants in the DP Imp group scored higher on the question *C1 Attacker* than the participants in other DP groups. However, this was to be expected, as the possibility of an attacker gaining access to the unperturbed data is mentioned within the DP Imp description. The only significant difference between DP vs. LDP is the correct response rate for *C1 Attacker* (30% vs. 49%, $\chi^2_{(1)} = 32.07, p < .001$, see Tab. 7).

We also tested whether participants who looked longer at the descriptions performed better at the comprehension questions. We computed Spearman’s rank correlation between the time participants were spending on the description of (L)DP and their cumulative score on the comprehension questions (Q5) and found a positive correlation ($r(935) = .237, p < .001$). Due to technical reasons, the figure we compare regarding the time spent reading the description also includes the participants’ answer to the questions whether they would like to share data (Q1) and their reasoning (Q2). The visualization of the differences in the average correct response rate based on the time spent looking at the (L)DP description can be seen in Fig. 7. There, we can see that the correct response rate peaks for participants who spent between 100 and 200 seconds reading the description and there is no improvement

when participants took longer than 200 seconds. The median time looking at the description was 57 seconds across all participants.

Sharing behavior: We used inductive coding to analyze the valid answers to the open question why participants decided (not) to share data (Q2) based on the established codes of the original study [48]. Two authors coded the answers independently and discussed the differences afterwards. If a participant's answer fell into two code categories, both were counted.

Why do participants want to share data?

Trust in DP and LDP techniques. 46% of responses fell into this category. Answers include “seems secure and recommended by experts”, “statistical analysis without identification” (Uber), “randomized data gives a sense of security” (LDP Imp w/o Local) but also wrong assumptions such as “seems to be encrypted” (Apple).

Utility considerations. This category encompasses 31% of valid responses. Examples are “more data equals better recommendations” (Apple), “probably important for using the app” (DP Flow), and “brings advantages and seems secure” (LDP Imp).

Little privacy concern for asked or any information, learned helpless, and no fear of loss. This category holds 30% of responses. It is noteworthy that most answers in this category fall into the category *no fear of loss*: “nothing to hide” (LDP Imp), “most information is online anyway” (DP Imp), and “the requested data is not that important” (Microsoft).

Why do participants not want to share data?

Too sensitive to share. The majority of responses (51%) fall in this category. Participants wrote: “personal data should stay personal” (Microsoft), “data not relevant for health app” (Microsoft), and “no advantage for me” (US Census). Another common theme in this category are participants who are skeptical about sharing their income level.

Distrust differential privacy techniques. 31% of answers revealed little trust in general or more explicitly in (L)DP: “the term ‘noise’ is not explained well enough” (US Census), and “does not sound trustworthy” (Uber)

Risks of data leak, breach, or hack. Similar to the previous category, 12.99% of answers indicated that data breaches are always possible, no matter what the security promises: “no data is secure” (Uber), and “even the best software has holes in it” (LDP Imp w/o Local).

Distrust the app or tech companies. 18% of responses explicitly stated distrust of apps or tech companies in general: “as it is a private company I distrust these promises of data security” (LDP Imp w/o Local), and “my data is none of the app's business” (LDP Imp)

Some participants, however, have opposite opinions. For example, we have two participants who each stated that the mention of Google in the description (LDP Comp) influenced their decision whether they wanted to share their data (Q1). One participant did not wish to share, stating that it “does

not seem secure especially since Google is involved”, while another participant decided to share because “it is used by Google and Apple and therefore must be secure”.

6.3.2 Additional tests

Again, we did some different additional tests to investigate potential differences in the participants' demographics.

Comprehension: We performed Kruskal-Wallis tests that revealed significant differences in IT background. Participants who indicated that they had an IT background found the description of (L)DP significantly easier to understand ($H(1) = 7.92, p = .005$) and answered correctly significantly more often to *C5 Utility 3rd party* ($H(1) = 4.652, p = .031$).

Willingness to share: Participants using apps to monitor their health were more willing to share information (Q1) than others ($H(1) = 37.47, p < .001$). Other demographics do not significantly impact the results.

Self-reported understanding vs. comprehension: We computed Spearman's rank correlation to investigate the relationship between self-reported understanding of the description (Q2) and the scores on the comprehension questions (Q5) (see Tab. 7). There was a significant positive correlation for *C1 Attacker* ($r(935) = .110, p = .001$), *C2 Employee* ($r(935) = .120, p < .001$), *C3 3rd party* ($r(935) = .168, p < .001$), and *C5 Utility 3rd party* ($r(935) = .152, p < .001$). Participants in LDP groups were on average significantly better at answering *C1 Attacker* correctly than the participants in the DP groups ($H(1) = 32.04, p < .001$).

6.4 Comparison and discussion

Compared to the original study, we obtained a similar sharing rate for the sensitive information (see Fig. 5). Across all conditions, 53% wanted to share sensitive information, opposed to 47.8% in the original study [48]. Also, we can report the largest sharing rate of 60% in the LDP Imp group, just as in the original study where the sharing rate of this condition was 65%. The major difference in this area is that no sharing rate is below 46% in our case, whereas there were some groups in the original study that had a sharing rate below that. An interesting similarity lies in the overall difficult-to-comprehend rate (A score less than 4 for Q3) of 13.4% compared to 13.3% in [48]. There is a difference in the lowest difficult-to-comprehend rate of a group: 0% in the original study's DP Imp group and 10% for our Apple group. However, our highest difficult-to-comprehend rate (DP Flow, 17%) is much lower than the one of the original study (DP w/o Names 30%). Overall, the differences in the difficult-to-comprehend rating and the sharing decision among groups are not as large as they were in the original study.

The participants' comments regarding the reasoning behind their decision to share or not to share data are very similar to the original study's. Two participants noted that they would

	DP						LDP				
	Apple	DP Flow	DP Imp	Microsoft	Uber	US Census	Google	LDP Comp	LDP Flow	LDP Imp	LDP Imp w/o Local
C1 Attacker	20%	24%	49%	29%	24%	27%	37%	45%	54%	55%	47%
C2 Employee	59%	34%	30%	33%	28%	29%	38%	32%	42%	48%	38%
C3 3 rd party	68%	53%	48%	45%	49%	49%	47%	47%	62%	65%	53%
C4 Usability	6%	23%	7%	21%	6%	11%	11%	7%	18%	14%	9%
C5 Useful 3 rd party	41%	49%	42%	48%	47%	43%	46%	47%	44%	35%	35%

Table 6: Correct response rates

		C1 Attacker	C2 Employee	C3 3 rd party	C4 Usability	C5 Useful 3 rd party
easy-to-comprehend	<i>r</i>	.11	.12	.16	-.03	.15
	<i>p</i>	<.001	<.001	<.001	.369	<.001
DP vs. LDP	χ^2	32.07	1.24	.30	.14	1.66
	<i>p</i>	<.001	.265	.583	.705	.197

Table 7: Correlations of easy-to-comprehend and difference DP/LDP regarding the comprehension questions in experiment B

like to see the data before it leaves the device in order to understand the data perturbation. There was no mention of this in the original study. An alarmingly large portion of answers fall into the privacy fatigue [26] category, with assumptions that their personal data is “never secure” and “out there anyway”. Also, the comments regarding the participants’ unwillingness to share their income level is in line with previous research about German data sharing preferences [23].

As in experiment A (see Sec. 5.4), personal health app usage has a significant impact on our participants’ answers. This time, it shows a significant difference in the decision to share. Unsurprisingly, participants with an IT background showed a significantly higher score on the self-reported comprehension of the description of (L)DP. The self-reported understanding of the (L)DP description correlated positively with almost all comprehension questions; however, the correlation coefficients are below 0.2 which suggests a weak association and thereby a limited effect size [38]. The only significant difference between participants who were assigned to DP and those that were assigned to LDP descriptions lies in the comprehension question *C1 Attacker*, where most participants in the DP groups falsely believed that an attacker does not have access to the real answers if the company’s data base is breached. As this is one of the key differences between DP and LDP, it shows once again that the difference is not clearly communicated and understood. One exception to this is the participants in the DP Imp group, where this scenario of a data base breach is explicitly mentioned. Still, even in this group most participants answered the question wrong. However, we found that participants who spent more time reading the

(L)DP descriptions performed better on the comprehension questions.

7 Discussion

While our study partially confirms the findings of the original study by Xiong et al. [48], we also provide additional insights about (L)DP communication in a different culture (Germany), different demographics, and the impact of personal health app usage. Overall, participants who were told that their data would be protected by (L)DP decided to share more high-sensitivity data than those in the control group, which indicates that (L)DP communication had a positive effect on their data sharing attitudes. Similar to the original study, the participants’ responses did not significantly differ between the LDP and the DP groups. This suggests that at least LDP was not completely understood. This also confirms the previously mentioned findings from Cummings et al. [11] that users misunderstand various descriptions of DP (see Sec. 3). Although self-reported understanding of the (L)DP descriptions was relatively high, the subsequent comprehension questions reveal that participants overestimated their understanding. Although participants with higher self-reported comprehension answered correctly more often on most comprehension questions, only few of them provided exclusively correct answers. As participants who spent more time reading the descriptions provided more correct answers, we can speculate that reading the description thoroughly improves the comprehension of (L)DP. However, it is also likely that the descriptions were not worded in a clear way. Due to the fact that users generally prefer not to read privacy statements [41], it is reasonable to assume that they do not want to read lengthy (L)DP descriptions either. As a result, alternative solutions based on more visual (L)DP communication like those proposed for privacy policies [27] should be investigated in the future. We observe the same pattern in the open answers about the participants’ willingness to share their data as in [48]. However, some participants noted that they would need an example of “how noise changes the data”, “what a hacker would have access to”, or “of what use the small inaccuracies are”. These statements indicate that users do not want only a vague privacy guarantee, which is probably too technical for laypeople to understand fully. They would rather see the actual perturbation of their data or at least a clearer

and more understandable presentation of (L)DP. Moreover, we could observe a pattern that the personal usage of health apps increases trust and the willingness to share data.

Besides the expected differences in attitudes due to cultural and regulatory differences, summarized in Sec. 3, it is also important to take the timeframe of the respective studies into account. Xiong et al. performed their study before March 2020, i.e., a time before worldwide lockdowns forced people and companies into digitalization. As our study was conducted during the summer of 2021, it is possible that our sample was more familiar with and presumably more trusting of digital technologies and less concerned about associated privacy risks.

8 Conclusion

We have replicated a study on the effect of DP communication on the willingness to share data and on the understanding of and trust in the privacy-preserving technique. Despite our different sample comprising German participants representative of the population, our results are similar to the original study in that participants' answers were not significantly different between LDP and DP models. However, the effect of DP communication could clearly be observed since the participants were significantly willing to share more data when (L)DP was applied. As a result, they trust the technology to protect their privacy. The big caveat is that even though self-reported understanding was high, follow-up comprehension questions revealed that participants did not fully understand the concept of (L)DP. Arguably, visual or otherwise more understandable differential privacy communication would help users' comprehension [24, 29].

References

- [1] Hofstede Insights. Online: <https://www.hofstede-insights.com/country-comparison/germany,india,the-usa/> (accessed in 02/2022).
- [2] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-Indistinguishability: Differential Privacy for Location-Based Systems. In *Proceedings of the 20th ACM SIGSAC Conference on Computer & communications security (CCS)*, 2013.
- [3] Catherine Barrett. Are The EU GDPR And The California CCPA Becoming the de facto Global Standards for Data Privacy and Protection? *Scitech Lawyer*, 2019.
- [4] Steven Bellman, Eric J Johnson, Stephen J Kobrin, and Gerald L Lohse. International Differences in Information Privacy Concerns: A Global Survey of Consumers. *The Information Society*, 2004.
- [5] Brooke Bullek, Stephanie Garboski, Darakhshan J Mir, and Evan M Peck. Towards Understanding Differential Privacy: When do People Trust Randomized Response Technique? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2017.
- [6] TH Hubert Chan, Elaine Shi, and Dawn Song. Optimal Lower Bound for Differentially Private Multi-Party Aggregation. In *Proceedings of the 20th European Symposium on Algorithms (ESA)*, 2012.
- [7] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease Prediction by Machine Learning over Big Data from Healthcare Communities. *IEEE Access*, 2017.
- [8] European Commission. General Data Protection Regulation (GDPR), 2016.
- [9] Panos Constantinides and David A Fitzmaurice. Artificial Intelligence in Cardiology: Applications, Benefits and Challenges. *Br J Cardiol*, 2018.
- [10] Rowena Cullen. Culture, Identity and Information Privacy in the Age of Digital Government. *Online Information Review*, 2009.
- [11] Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. "I need a better description": An Investigation Into User Expectations For Differential Privacy. In *Proceedings of the 28th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.
- [12] Djellel Eddine Difallah, Elena Filatova, and Panagiotis G. Ipeirotis. Demographics and Dynamics of Mechanical Turk Workers. *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*, 2018.
- [13] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (ICONIP)*, 2017.
- [14] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming (ICALP)*, 2006.
- [15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the 3rd Theory of Cryptography Conference (TCC)*, 2006.
- [16] Benjamin Edwards, Steven Hofmeyr, and Stephanie Forrest. Hype and Heavy Tails: A Closer Look at Data Breaches. *Journal of Cybersecurity*, 2016.

- [17] Úlfar Erlingsson, Vasył Pihur, and Aleksandra Korolova. Rappor: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 21st ACM SIGSAC Conference on computer and communications security (CCS)*, 2014.
- [18] Arik Friedman and Assaf Schuster. Data Mining with Differential Privacy. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.
- [19] Kenneth Goodman, Diana Zandi, Andreas Reis, and Effy Vayena. Balancing Risks and Benefits of Artificial Intelligence in the Health Sector. *Bulletin of the World Health Organization*, 2020.
- [20] Nadine Guhr, Oliver Werth, Philip Blacha, and Michael Breitner. Privacy Concerns in the Smart Home Context. *SN Applied Sciences*, 2020.
- [21] Anne-Wil Harzing. Why Replication Studies are Essential: Learning from Failure and Success. *Cross Cultural & Strategic Management (CCSM)*, 2016.
- [22] Aylin Ilhan and Maria Henkel. 10,000 Steps a Day for Health? User-based Evaluation of Wearable Activity Trackers. In *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS)*, 2018.
- [23] Maria Karampela, Sofia Ouhbi, and Minna Isomursu. Exploring Users' Willingness to Share Their Health and Personal Data Under the Prism of the New GDPR: Implications in Healthcare. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.
- [24] Farzaneh Karegar and Simone Fischer-Hübner. Vision: A Noisy Picture or a Picker Wheel to Spin? Exploring Suitable Metaphors for Differentially Private Data Analyses. In *European Symposium on Usable Security 2021*, 2021.
- [25] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What Can We Learn Privately? *SIAM Journal on Computing (SICOMP)*, 2011.
- [26] M. J. Keith, C. Maynes, P. B. Lowry, and J. Babb. Privacy Fatigue: The Effect of Privacy Control Complexity on Consumer Electronic Information Disclosure. In *Proceedings of the 35th International Conference on Information Systems, (ICIS)*, 2014.
- [27] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A "Nutrition Label" for Privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS)*, 2009.
- [28] Hanna Krasnova, Natasha F Veltri, and Oliver Günther. Self-Disclosure and Privacy Calculus on Social Networking Sites: The Role of Culture. *Business & Information Systems Engineering (BISE)*, 2012.
- [29] Patrick Kühnreiter and Delphine Reinhardt. Usable Differential Privacy for the Internet-of-Things. In *Proceedings of the 19th IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2021.
- [30] Ereni Markos, George R Milne, and James W Peltier. Information Sensitivity and Willingness to Provide Continua: A Comparative Privacy Study of the United States and Brazil. *Journal of Public Policy & Marketing (JPP&M)*, 2017.
- [31] Sandra J Milberg, H Jeff Smith, and Sandra J Burke. Information Privacy: Corporate Management and National Regulation. *Organization science*, 2000.
- [32] Zareef A Mohammed and Gurvirender P Tejay. Examining Privacy Concerns and Ecommerce Adoption in Developing Countries: The Impact of Culture in Shaping Individuals' Perceptions Towards Technology. *Computers & Security*, 2017.
- [33] Patrick Murmann, Matthias Beckerle, Simone Fischer-Hübner, and Delphine Reinhardt. Reconciling the What, When and How of Privacy Notifications in Fitness Tracking Scenarios. *Pervasive and Mobile Computing (PMC)*, 2021.
- [34] Kee Yuan Ngiam and Wei Khor. Big Data and Machine Learning Algorithms for Health-Care Delivery. *The Lancet Oncology*, 2019.
- [35] Caroline L Park. What is the Value of Replicating Other Studies? *Research Evaluation*, 2004.
- [36] Christine Prince, Nessrine Omrani, Adnane Maalaoui, Marina Dabic, and Sascha Kraus. Are We Living in Surveillance Societies and Is Privacy an Illusion? An Empirical Study on Privacy Literacy and Privacy Concerns. *IEEE Transactions on Engineering Management*, 2021.
- [37] Ismini Psychoula, Deepika Singh, Liming Chen, Feng Chen, Andreas Holzinger, and Huansheng Ning. Users' Privacy Concerns in IoT Based Applications. In *Proceedings of the 4th IEEE International Conference on Internet of People (IoP)*, 2018.
- [38] Louis M Rea and Richard A Parker. *Designing and Conducting Survey Research: A Comprehensive Guide*. 2014.

- [39] Nuria Rodríguez-Barroso, Goran Stipcich, Daniel Jiménez-López, José Antonio Ruiz-Millán, Eugenio Martínez-Cámara, Gerardo González-Seco, M Victoria Luzón, Miguel Angel Veganzones, and Francisco Herrera. Federated Learning and Differential Privacy: Software tools analysis, the Sherpa. ai FL framework and methodological guidelines for preserving data privacy. *Information Fusion*, 2020.
- [40] Eva-Maria Schomakers, Chantal Lidynia, Dirk Müllmann, and Martina Ziefle. Internet Users’ Perceptions of Information Sensitivity—Insights from Germany. *International Journal of Information Management (IJIM)*, 2019.
- [41] Tomáš Sigmund. Attention Paid to Privacy Policy Statements. *Information*, 2021.
- [42] Statistisches Bundesamt (Destatis). 12111-0004: Bevölkerung (Zensus): Deutschland, Stichtag, Geschlecht, Altersgruppen, 2021.
- [43] Differential Privacy Team. Learning with Privacy at Scale. Online: <https://machinelearning.apple.com/research/learning-with-privacy-at-scale> (accessed in 12/2021).
- [44] Sabine Trepte, Leonard Reinecke, Nicole B Ellison, Oliver Quiring, Mike Z Yao, and Marc Ziegele. A Cross-Cultural Perspective on the Privacy Calculus. *Social Media+ Society*, 2017.
- [45] Amos Tversky and Daniel Kahneman. Rational Choice and the Framing of Decisions. In *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*. 1989.
- [46] Stanley L Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association (JASA)*, 1965.
- [47] Jochen Wirtz, May O Lwin, and Jerome D Williams. Causes and Consequences of Consumer Online Privacy Concern. *International Journal of service industry management*, 2007.
- [48] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. Towards Effective Differential Privacy Communication for Users’ Data Sharing Decision and Comprehension. In *Proceedings of the 41st IEEE Symposium on Security and Privacy (S&P)*, 2020.

APPENDIX

A Questionnaire for Experiment A

The questionnaire is taken from Xiong et al.’s original study [48] and has been translated to German. We provide the back-translated English version of our questionnaire which is almost verbatim to the original study. The PDF version of the German questionnaire is available online¹.

Introduction

In the digital age, everyone faces the question whether to share personal data in exchange for goods, services, or other advantages. The goal of this study is to understand what kinds of information you wish to share with a health app and how these data should be used.

Demographics

The demographics were checked first in order to fulfil the quotas of the questionnaire. Participants had the possibility to answer “No answer” to all questions.

What is your age group?

Please indicate your gender.

What is your highest school-leaving qualification?

Do you have an IT background?

Do you use apps or devices to monitor your health data?

Precondition

Please assume the following for this questionnaire:

- 1. You have just downloaded the health app Orange Health and you start using it immediately*
- 2. To ensure suitable advice and recommendations regarding your health, the app asks for certain information, for example, your age and gender in regard to daily calorie intake.*
- 3. At the same time, the app server requests permission to access and collect the information in order to provide you with a better user experience. For example, the information you share will be used to train machine learning algorithms that will subsequently will be used to provide more exact recommendations for all users.*

Differential privacy communication

Here, the participants in the DP and LDP groups were shown the descriptions for DP and LDP respectively (see Sec. C). Afterwards, the following comprehension question was presented.

Please indicate which of the following descriptions of (local) differential privacy is correct:

¹<https://owncloud.gwdg.de/index.php/s/kDAUTawPdsJxAWp>

- A data protection technique that adds random noise to the collected data of user groups (e.g. average age) in order to protect the user’s privacy just as if the user had not taken part in the data collection.
- A data protection technique, which adds random noise to every user response in order to protect the user’s privacy just as if the user would not take part in the data collection.
- DP/LDP has not been used yet in any organization or company.
- I prefer not to answer.

Participants were shown the respective description again if they answered incorrectly.

Questions of the Orange Health app

Participants first were presented with an explanation of how to answer the questions. Again, this is a direct translation and adaption of the original explanation from Xiong et al. [48].

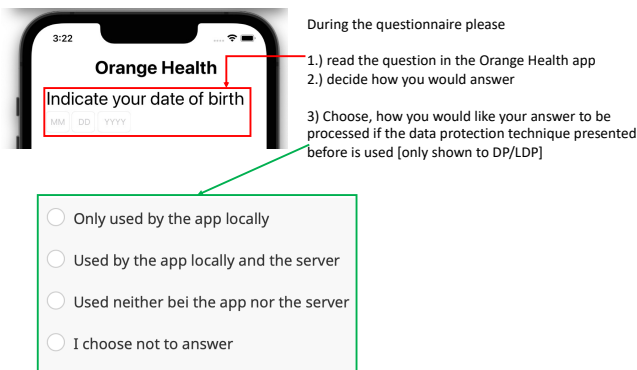


Figure 8: Back-translated explanation for the participants. Point 3 has only been provided to the groups DP and LDP.

Participants were provided with 14 screenshots of the questions in Tab. 2 similar to the one in Fig. 8 in random order and could choose the following answers.

- Only used by the app locally
- Used by the app locally and the server
- Neither used by the app nor the server
- I prefer not to answer

Trust questions

Participants could answer the following questions on a 7-point Likert scale ranging from “Strongly disagree” to “Strongly agree”.

1. I trust the Orange Health app to protect my personal information
2. I trust the app server to protect my personal information
3. I trust (local) differential privacy to protect my personal information

The third question was only asked to participants in the DP and LDP groups.

B Questionnaire for experiment B

The PDF version of the German questionnaire is also available online².

Introduction

The goal of this study is to evaluate your willingness to share personal information when a data protection technique is used. The goal is also to understand why you made this respective decision. Furthermore, we want to evaluate your comprehension of this data protection technique.

Demographics

We asked the same demographic questions as in the first questionnaire A

Precondition

We presented the same precondition as in the first questionnaire A

Differential privacy

To respect your personal information and to guarantee a better user experience, the data that are shared with the Orange Health app are collected using a data protection technique. This data protection technique is presented in the following. Please read the description carefully.

The participants were randomly assigned to one of the eleven descriptions in D.

Trust in (L)DP

Under the condition that the above described data protection technique is in use: Would you share your personal data (e.g. date of birth, family medical record, income level, substance use, medical record, previous surgeries, current medication) with the app server?

- Yes / No / No answer

²<https://owncloud.gwdg.de/index.php/s/s5hQeVmLNyy2kve>

If the participant answered yes:

Please explain briefly why you would like to share your personal data if the described data protection technique is in use?

- Open question

If the participant answered no:

Please explain briefly why you would not like to share your personal data if the described data protection technique is in use?

- Open question

Self-reported understanding of (L)DP

Participants could answer the following questions on a 7-point Likert scale ranging from “Strongly disagree” to “Strongly agree”.

Please indicate your agreement with the following statements: The previous description of the data protection technique was easy to understand.

If participants provided a score of 3 (“mildly disagree”) or less they were presented with the description again to highlight words they did not understand.

You have indicated that the description of the data protection technique was not easy to understand. Please indicate the words you find hard to understand by clicking on them to highlight them.

Comprehension questions

- C1 *Suppose you have answered truthfully to the questions in the Orange Health app and your answers have been collected with the presented data protection technique. If an attacker gets access to the data base of the Orange Health company, will he then be able to see your true answers?*
- C2 *Suppose you have answered truthfully to the questions in the Orange Health app and your answers have been collected with the presented data protection technique. Are employees within the Orange Health company able to see your true responses?*
- C3 *Suppose you have answered truthfully to the questions in the Orange Health app and your answers have been collected with the presented data protection technique. Are third parties with whom the Orange Health company shares data able to see your true answers?*
- C4 *With the changes imposed through the data protection technique, the accuracy of the aggregated data the Orange Health company receives is ... compared to the actual results without the data protection technique.*

- C5 *Suppose you have shared data such as your family medical record with the health app. Do the results, which have been collected using the data protection technique to protect your privacy, stay useful for third party companies with whom the health app company shares data?*

Participants could answer *Yes / No / Unsure / No answer* for all answers except C4, which had the options *better / worse / unchanged / unsure*.

C Descriptions of (L)DP for experiment A

Again, here and in Appendix D we provide the back-translated English versions of our German (L)DP descriptions which are almost the same as the ones provided by Xiong et al. [48].

DP

Data shared with the app will be processed using differential privacy (DP) to protect your personal data and to ensure the best user experience. DP protects the users’ privacy by adding random noise to the aggregated data, such as average age, so that the probability of deducing an individual person’s information is low. DP is used in academia as well as in the corporate world, including Harvard University, the US Census Bureau and corporations such as LinkedIn and Uber.

LDP

Data shared with the health app will be collected using local differential privacy (LDP) to protect your personal information and to ensure the best user experience. LDP protects the users’ privacy by adding random noise to every answer provided by a user. As a result, the probability of deducing a user characteristic is roughly as high as if the user had not taken part in data collection. LDP is used by companies such as Apple and Google.

D Descriptions of (L)DP for experiment B

LDP Flow

When local differential privacy (LDP) is used, the app changes the answers before they are sent from the user’s device to the company. The company sees and stores only the changed version of each user’s information and is unsure of the users’ true answers. If changed answers from a large number of users are analyzed, however, the company can still gather useful results in aggregated form about the user population, although the accuracy is reduced compared to unchanged data.

DP Flow

When differential privacy (DP) is used the app sends the user’s answers to the company. These answers are stored in

the company's data base. If the company wants to use these data either internally or with third parties, the company sends queries to the data base, uses DP techniques to change the results of the queries and uses only these changed results. The changed results only provide limited information concerning a specific user. If, however, the answers of a large number of users are analyzed, the company can still obtain useful results in aggregated form about the whole user population, even if the accuracy is lower compared to unchanged data.

US Census

Differential privacy has been developed by researchers at Microsoft and is used by many leading technology companies. There are many variants of differential privacy. The one used here introduces controlled noise into the data, so that the accuracy remains at higher levels. This method to protect privacy has been developed to maintain the data's usability and also to completely protect the personal information of each affected person.

Google

Building upon the concept of randomized response, local differential privacy (LDP) makes it possible to generate statistics about user behavior while guaranteeing the users' privacy. LDP builds upon this concept by allowing the app to send reports that are factually indistinguishable from random coin tosses and do not contain any unique user names. By aggregating reports, common statistics that are the same for many users can be derived.

Apple

Differential privacy transforms the information that is shared with the company before it leaves the device, so that the company can never reproduce the true data. The basic idea of differential privacy is to introduce statistical noise that hides the users' personal data before they are sent to the company. When a lot of people send the same kinds of data the introduced noise will cancel out on average and the company is able to gather useful information thanks to the huge amount of data.

Uber

Differential privacy is a formal definition of privacy and is accepted on a broad scale by industry experts because it provides robust privacy protection. In short, differential privacy allows general statistical analyses without revealing information about an individual within the data. That is why differential privacy provides an additional safety barrier against recognition attacks as well as attacks with auxiliary data.

Microsoft

Differential privacy is a technique that enables researchers and analysts to obtain useful analyses of data bases containing personal information. At the same time, it provides a strong protection for individual privacy. This seemingly contradictory result is reached by inserting relatively moderate inaccuracies into the analyses. These inaccuracies are large enough to protect the privacy but small enough so that the analyses remain useful for researchers and analysts.

LDP Imp. w/o Local

Data that is shared with the app will be processed with the help of the differential privacy (DP) technique to respect your personal information and to ensure the best user experience. The app will change the data on your app randomly before they are sent to the app-server. As the app-server now only stores the changed version of your personal information, your privacy is protected even if the data base of the app-server will be compromised.

LDP Imp.

Data that is shared with the app is processed with the help of the local differential privacy (LDP) technique to respect your personal information and to ensure the best user experience. The app changes the data on your app randomly before they are sent to the app server. As the app server now only stores the changed version of your personal information, your privacy is protected even if the data base of the app server is compromised.

DP Imp.

Data that is shared with the app is processed with the help of the differential privacy (DP) technique to respect your personal information and to ensure the best user experience. The health app company stores your data but only uses the modified total statistics, so that your personal information cannot be learned. Your personal information can be leaked, however, if the data base of the company is compromised.

LDP Comp

Data that is shared with the app is processed with the help of the local differential privacy (LDP) technique to respect your personal information and to ensure the best user experience. LDP protects your privacy by introducing random noise to the raw data BEFORE they are sent to the company (the raw data never leaves your device). LDP is used by companies such as Google and Apple.

E Statistics

Question	Local Only		Both		Opt out	
	χ^2	<i>p</i>	χ^2	<i>p</i>	χ^2	<i>p</i>
Sensitivity	21.48	<.001	280.5	<.001	217.63	<.001
Condition	.110	.947	97.95	<.001	139.21	<.001
Con vs. DP	N/A		<.001		<.001	
Con vs. LDP			<.001		<.001	
DP vs. LDP			.917		.729	
QS * Condition	.65	.722	9.94	.007	8.08	.018
Low vs. High QS						
Control	N/A		<.001		<.001	
DP			<.001		<.001	
LDP			<.001		<.001	
low-sensitivity						
Con vs. DP	N/A		.033		.002	
Con vs. LDP			.028		.001	
DP vs. LDP			.912		.718	
high-sensitivity						
Con vs. DP	N/A		<.001		<.001	
Con vs. LDP			.001		.004	
DP vs. LDP			.790		.465	

Table 8: Statistics for experiment A

		Trust in		
		App	Server	(L)DP
Age	$H_{(4)}$	3.297	4.705	2.007
	<i>p</i>	.509	.319	.734
Gender	$H_{(1)}$.845	1.14	.818
	<i>p</i>	.358	.286	.366
Education	$H_{(4)}$	4.131	4.628	5.912
	<i>p</i>	.389	.328	.206
IT BG	$H_{(1)}$	1.41	7.43	.848
	<i>p</i>	.842	.115	.357
Health App	$H_{(1)}$	40.028	27.362	26.31
	<i>p</i>	<.001	<.001	<.001

Table 9: Kruskal-Wallis tests on correlations between demographics and trust in experiment A

		Opt out		Local only		Both	
		Low	High	Low	High	Low	High
Age	$\chi^2_{(28)}$	38.66	56.74	36.4	56.25	49.84	37.77
	<i>p</i>	.087	.001	.133	.001	.007	.103
Gender	$\chi^2_{(7)}$	5.86	10.41	20.50	25.76	16.33	18.85
	<i>p</i>	.556	.167	.005	.001	.022	.009
Education	$\chi^2_{(28)}$	29.32	38.9	35.39	24.39	32.38	33.36
	<i>p</i>	.396	.083	.159	.661	.259	.223
IT BG	$\chi^2_{(7)}$	16.38	8.03	8.89	7.64	13.6	18.36
	<i>p</i>	.022	.330	.261	.365	.059	.010
Health App	$\chi^2_{(7)}$	15.11	19.23	6.794	4.55	15.33	32.59
	<i>p</i>	.035	.007	.451	.715	.032	<.001

Table 10: Correlation between demographics and willingness to share for experiment A

		Share	easy-to-comprehend
		Age	$H_{(4)}$ 6.488
	<i>p</i>	.166	.958
Gender	$H_{(1)}$	1.871	2.852
	<i>p</i>	.171	.091
Education	$H_{(4)}$	1.017	4.833
	<i>p</i>	.907	.305
IT BG	$H_{(1)}$.526	7.918
	<i>p</i>	.468	.005
Health app	$H_{(1)}$	37.465	1.937
	<i>p</i>	<.001	.164

Table 11: Kruskal Wallis tests for impact on demographics on willingness to share and self-reported easy-to-comprehend rate for experiment B

		C1	C2	C3	C4	C5
		Age	$H_{(4)}$.701	2.282	5.667	9.239
	<i>p</i>	.951	.684	.225	.055	.145
Gender	$H_{(1)}$.27	.139	2.38	.028	3.294
	<i>p</i>	.603	.710	.123	.867	.070
Education	$H_{(4)}$	6.578	6.89	5.928	6.978	6.819
	<i>p</i>	.160	.142	.205	.137	.146
IT BG	$H_{(1)}$.347	.001	.722	1.705	4.652
	<i>p</i>	.556	.981	.396	.192	.031
Health app	$H_{(1)}$	2.895	.176	.005	2.14	.389
	<i>p</i>	.089	.675	.944	.143	.533

Table 12: Kruskal Wallis tests for impact on demographics on the comprehension questions for experiment B

Comparing User Perceptions of Anti-Stalkerware Apps with the Technical Reality

Matthias Fassel
*CISPA Helmholtz Center
for Information Security*

Simon Anell
*CISPA Helmholtz Center
for Information Security*

Sabine Houy
Umeå University

Martina Lindorfer
TU Wien

Katharina Krombholz
*CISPA Helmholtz Center
for Information Security*

Abstract

Every year an increasing number of users face stalkerware on their phones [84]. Many of them are victims of intimate partner surveillance (IPS) who are unsure how to identify or remove stalkerware from their phones [49]. An intuitive approach would be to choose anti-stalkerware from the app store. However, a mismatch between user expectations and the technical capabilities can produce an illusion of security and risk compensation behavior (i.e., the Peltzmann effect).

We compare users' perceptions of anti-stalkerware with the technical reality. First, we applied thematic analysis to app reviews to analyze user perceptions. Then, we performed a cognitive walkthrough of two prominent anti-stalkerware apps available on the Google Play Store and reverse-engineered them to understand their detection features.

Our results suggest that users base their trust on the look and feel of the app, the number and type of alerts, and the apps' affordances. We also found that app capabilities do not correspond to the users' perceptions and expectations, impacting their practical effectiveness. We discuss different stakeholders' options to remedy these challenges and better align user perceptions with the technical reality.

1 Introduction

About one in five adults and even more young adults engage in snooping attacks on others' phones [54]. *Intimate partner surveillance* (IPS) is a specific subset of these attacks [13, 88]. Tool-based IPS often involves a type of spyware, called stalkerware (or surveillanceware), to collect live location data, contacts, call history, and text messages [15, 80].

According to the Coalition Against Stalkerware [84], 67,500 mobile users were confronted with stalkerware in 2019, a 67% increase compared to the year before. Randall et al. [76] estimated that at least 5,758 people in the US were targeted by overt stalkerware from March to May 2020. Two of the 22 apps they studied were available in the Google Play Store, the remainder were only available from third parties. In October 2020, Google banned surveillance apps from their store [37] and now only allows surveillance in parental control and enterprise management apps if they do not hide or obfuscate their surveillance practices. Hence, stalkerware often rebrands itself as parental control apps or moves to third-party websites. Most stalkerware occurrences in *clinical computer security* [43] consultations comprise such "dual-use" apps [15].

An analysis of online domestic abuse forums and an assessment of the stalkerware application (app) industry identified that IPS survivors are unsure how to recognize and remove stalkerware [49, 66]. Installing anti-stalkerware apps from the Google Play Store is one possible approach. Users may choose from various apps, ranging from traditional anti-virus companies offering general mobile security solutions to specialized apps detecting stalkerware and other spyware. Prices vary widely, some are as cheap as €5 (or \$), but in-app purchase prices up to and beyond €100 (or \$) are not uncommon. However, these apps come with severe limitations on Android since they often operate with simple name-based blocklists, which stalkerware can circumvent easily [10]. More worryingly, there have also been instances of fake anti-virus apps in the Google Play Store with limited to no functionality at all [22, 45, 63, 97]. Thus, the marketed promise of identifying stalkerware is at odds with many of these apps' abilities, constituting an expectation-ability gap. This problem affects users' ability to make informed decisions. Survivors should be made aware of these problems to allow them to question their reliance on them.

We conduct an exploratory case study with two anti-stalkerware apps to understand this mismatch between expectations and abilities. We focus on the following research questions: (RQ1) *What are the differences between users'*

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022,
August 7–9, 2022, Boston, MA, United States.

security perceptions and the anti-stalkerware apps' abilities?; and (RQ2) *How could research and design begin to remedy this mismatch and foster users' anti-stalkerware decisions?* We apply thematic analysis to app-store reviews to study perceptions of these apps. We also perform a cognitive walkthrough of the respective apps and then reverse engineer them to understand how their detection mechanisms work. Hence, we elicit expectation-reality mismatches by combining qualitative user research with a reverse-engineering approach. Based on app reviews, we identified five user approaches to building confidence in their anti-stalkerware choice, all of them intuitive to apply and with some degree of legitimacy. However, contrasting these approaches with the cognitive walkthrough and reverse engineering results demonstrates that they fail to inform users about apps' abilities to mitigate violence, abuse, and harassment. Our work helps improve the current state of anti-stalkerware by suggesting design directions, proposing toolkit-supported user decisions, and discussing systemic, platform-level approaches to combating intimate partner surveillance.

2 Background and Related Work

This section describes background information and prior work on intimate partner surveillance and our methodology.

2.1 Intimate Partner Surveillance

Insiders, i.e., persons who are familiar to the victims, are a threat to smartphone users that security experts underestimated in the past [60]. Insiders' access to victims' devices varies significantly. However, according to one study in the US, 31% of participants looked through others' smartphones without their permission [54]. Surveillance among intimate partners, a specific insider attack, is usually technically unsophisticated and relies on UI-bound attacks or ready-made apps [27]. Bellini et al. [13] and Tseng et al. [88] analyzed stories on online forums about sexual infidelity. Abusers justify their surveillance with their suspicion of sexual infidelity. They want to collect evidence, understand behavior, and control behavior [13]. Bellini et al. [13] identified a four-stage abuse cycle: setting the abusers' expectations, attitude change, escalation, and reflection. Tseng et al. [88] categorized IPS attacks based on physical and non-physical access requirements. They found that online communities are a good source of IPS threat intelligence because their users collaborate to create new IPS attacks.

Chatterjee et al. [15] identified apps that are dangerous in the IPS context. They found explicit spyware apps and more subtle dual-use apps with legitimate use-cases (e.g., FindMyFriend). Often, anti-spyware does not identify the latter as a threat. Parental control apps, a classic example of dual-use, also suffer from other privacy issues, e.g., collecting sensitive data and distributing it to third parties without consent [24]. To understand the "creepware" ecosystem, Roundy et al. [80]

developed the *creeprank* algorithm based on guilt by association. As a result, hundreds of apps were removed from official app stores and presumably moved to third-party repositories.

Based on survivors' stories, Matthews et al. [57] identified different phases of separation and technology use. Survivors' safety in the "life apart" phase depends on identifying stalkerware. Havron et al. [43] and Freed et al. [26] created a computer security clinic for IPS survivors who readily accepted support in this format. However, since anti-stalkerware apps have a low barrier for entry, survivors presumably also use them as part of their protection ensemble. Lee et al. [48] extended the theory of planned behavior to understand factors leading to anti-spyware software adoption.

2.2 Users' Security Behavior

Due to a lack of structured security education, users learn their security behaviors haphazardly from various sources. Media, negative experiences, family, peers, workplace, IT professionals, and service providers are common advice sources [79]. However, all these sources focus on different aspects of threats [72]. Hence, no single source is sufficient. Giving security advice to individuals in situations of abuse is especially sensitive: affirmative steps to prevent attackers' data access suggest a lack of trust and may worsen abuse situations [50]. Emms et al. [23] suggested approaches to improve survivors' ability to avoid traces in ongoing abuse situations. Anti-stalkerware apps not specifically adapted for use in abuse situations may only be safe to use in the life apart phase. IPS survivors seek help and support in online forums from other survivors [49]. However, forum users often lack appropriate technical know-how, making it hard to recommend safe and effective anti-stalkerware apps. Reviews influence online consumer decisions in general. The quality of the review contents and the ranking affects consumer decisions more than the number of reviews and the sources' credibility [25]. Reviews can also influence security decisions, e.g., some users check app-store reviews before their update decision [87]. Most people also learn security lessons from family members' and friends' stories [73]. However, the stories' contents, the location, and the storyteller influence lessons' effectiveness. Social influence from peers affects security features' adoption, depending on the features' visibility to others [18]. Luca et al. [19] identified peer pressure from friends as the main factor for secure instant messenger adoption. Personal negative experiences also influence future security decisions. Vaniea et al. [89] found that users avoid updating after bad update experiences.

However, advice is not the only source of behavior – software prompts and automated security decisions also impact users' security behavior [78]. Mathiasen et al. [55] found that behaving securely does not necessarily result in a secure experience. According to them, careful design focused on creating secure experiences can increase security feature adoption. Distler et al. [21] found that including security-

related information in an e-voting process improved users' secure experience. They discuss how quick and smooth security mechanisms may impede users' secure experience despite improved usability—an idea they extend on in a framework of security-enhancing friction [20]. Users' mental models of potential attackers impact their adopted protection behavior [92] since each class of attacks calls for different protection mechanisms. Zou et al. [99] studied users' reasons for adopting and abandoning security and privacy behaviors. They found low adoption for recurring interaction practices and higher privacy practice adoption rates among low-income participants. Users abandoned security and privacy practices when they found them impractical, no longer saw their value, or perceived diminished risk. Similarly, users turn off protective measures such as firewalls when they find them complicated [75].

2.3 Review Mining and Analysis

App-store reviews inform users about the apps' quality, but also developers about bugs and feature requests, as well as researchers to gain detailed insights about apps. In light of the sheer number, informality, and shortness of these reviews, researchers either mine reviews to get a broad overview or use thematic analysis to examine a subsample in rich detail.

The software engineering community explored automated ways to mine user reviews for actionable development feedback. Prior work discussed several different automatic approaches to identify informative complaints in app reviews [16, 30, 53, 69]. Khalid et al. [47] used manual qualitative analysis to identify complaints about iOS apps.

Others have focused on automatically retrieving feature requests from reviews [46, 53] using natural language processing, sentiment analysis, and LDA models. Automatic analysis of app reviews can also inform developers about usability and user experience issues [12, 44, 58, 64]. Gu et al. [38] and Guzman et al. [39] applied sentiment analysis to understand how users feel about apps and individual features.

Researchers have also used reviews to study security- and privacy-related aspects of apps. Ha et al. [40] manually coded reviews to look for security and privacy complaints and found that about 1% of them concerned app permissions. Nguyen et al. [61] analyzed reviews for security- and privacy-related reports and traced 61% of security and privacy updates to corresponding user reviews. Voskobochnikov et al. [91] analyzed cryptocurrency wallets' reviews to understand security- and privacy-relevant UX issues. They identified a subsample of relevant reviews using machine learning and natural language processing and then applied thematic analysis. Gosh et al. [32, 33] qualitatively analyzed reviews of parental control apps to understand how children responded to them. They used a keyword search to filter children's reviews and applied thematic analysis. Children found the apps overly restrictive and privacy-invasive. They criticized their parents' reliance on these apps as a bad parenting technique.

2.4 Spyware Detection

In general, there are two basic approaches to detecting and analyzing malware, including stalkerware: static and dynamic analysis [5]. Static analysis is the understanding of a program at the syntactic source code or binary level [31]. Dynamic analysis focuses on an app's run-time behavior, including system calls and network traffic. For this purpose, researchers execute and observe apps in controlled environments [52].

Knowing the reliability of on-device anti-malware scanners (commonly referred to as anti-virus) is crucial for end users' safety. These scanners base their detection mechanisms on either static or dynamic analysis. However, compared to security solutions on desktop operating systems, mobile security apps have limited visibility into other apps due to extensive sandboxing, rendering behavioral heuristics unfeasible [17, 51, 70, 71]. Security solutions thus have to rely on signatures based on code-level characteristics or use machine learning [9, 51]. Related work has investigated in-depth how easy it is to evade those signatures [11, 41, 71, 77, 98]. Yet, no study so far compared the robustness of detection mechanisms to the trust users put into these security solutions.

3 Methodology

We explore the gap between users' expectations of the apps' functionality and the apps' technical abilities. Understanding this mismatch helps to improve users' protection against stalkerware. First, we apply thematic analysis [14] to app-store reviews of the two case-study apps to understand users' security perceptions and expectations. Based on the resulting themes, we perform cognitive walkthroughs of the apps and analyze them to understand how they detect stalkerware.

3.1 Selection of Anti-Stalkerware Apps

Spyware poses an increased danger to Android users compared to iPhone users [42, 66]. Apple's iOS claims tighter security controls [7] and does not allow apps with "functionality it does not actually offer (e.g., iOS-based virus and malware scanners)" [8]. Hence, we focus on Android apps.

To cover a variety of app abilities and user expectations in our qualitative analysis, we base our selection on Chatterjee et al.'s anti-spyware list [15]. From the most-downloaded anti-stalkerware apps, we chose two to perform static analysis on: Mobile Security, Antivirus & Cleaner by Lookout¹ (100M+ installs) [86]. From the long-tail, we read app-store pages and chose a data-rich example suitable for further qualitative analysis: Anti Spy Mobile PRO² (100k+ installs) [85].

Fraudulent reviews and manipulated ratings plague free apps [74, 95, 96]. Therefore, we prefer to analyze reviews of paid apps. Lookout Mobile Security is free to download on the Google Play Store and uses an in-app subscription model.

¹Version: 10.33-6652654, Downloaded: June 2020

²Version: 1.9.10.51, Downloaded: June 2020

We can not differentiate between subscribed and unsubscribed users' reviews. Hence, we also analyzed reviews from unsubscribed users. Lookout Mobile Security is more extensive and complex than Anti Spy Mobile PRO. Lookout Mobile Security markets itself as a fully-fledged security solution, with anti-spyware as only one of its features. In contrast, Anti Spy Mobile is available as a free or paid version (€ 4.90 or \$ 3.99) on the Google Play Store. The only difference is that the paid version has automatic daily background scans. We only analyzed the paid version's reviews.

The focus on these two apps affects the results twofold: First, their features are not representative of all security apps marketed as anti-stalkerware. Second, Lookout Mobile is pre-installed for some users, so the lack of choice may impact users' reviews. Hence, reviewers' sentiments from these two apps are not generalizable to all security apps that market themselves as anti-stalkerware.

3.2 Analysis of App-Store Reviews

To understand how users perceive our case study's anti-stalkerware apps and engender trust in them, we applied thematic analysis [14] to a sample of their app-store reviews.

We fetched all reviews from the Google Play Store.³ We randomly sampled 200 comments from each app in German and English, languages all involved researchers understand well. To ensure the reviews had enough content, we only considered comments with at least ten words. We analyzed a total of 400 reviews for Lookout. Anti Spy Mobile PRO, had less than 200 reviews fulfilling our criteria, so we analyzed a total of 13 German and 102 English reviews for this app.

At the start of the thematic analysis, one researcher read all reviews and created an initial codebook. With it, both researchers coded the entire review sample. During the coding procedure, both researchers kept notes on potential themes in the data. This resulted in an inter-coder agreement of Krippendorff's alpha $\alpha = 0.86$, which suggests *excellent* agreement. Afterward, the researchers discussed all mismatches and the themes they identified. Vague reviews with multiple valid interpretations caused most of the disagreements. Resolving conflicts increased Krippendorff's alpha to $\alpha = 0.98$. Table 1 in the Appendix presents the initial codebook.

The discussions led both researchers to agree on a focus on safety and security perceptions. We repeated the above procedure and constructed an additional codebook. Krippendorff's alpha was $\alpha = 0.78$ after the initial round of coding, suggesting *substantial* inter-coder agreement. Discussing all mismatches increased Krippendorff's alpha to $\alpha = 0.96$. At the start of the discussion, the researchers added a "time of experience" code and applied it whenever appropriate. Table 2 in the Appendix presents the revised codebook. Afterward, both researchers discussed the identified themes and the presentation of the results.

³Anonymized JavaScript code: <https://pastebin.com/bRZ1v0XS>

3.3 Anti-Stalkerwares' Technical Abilities

After identifying security perceptions and expectations, we used *theoretical sampling* to understand these apps' technical abilities. Thus, we collected data about the user interface and the apps' internal detection mechanisms.

We conducted cognitive walkthroughs for both apps to improve our understanding of the reviews focusing on user experience. Based on the previously discovered themes, we focused on the following: (1) method of invoking scans (manual, scheduled, event-triggered), (2) type and amount of information in reports, (3) false positives in a general use scenario, (4) visible user interactions under regular usage. We screenshot these parts of the case study apps and deductively code them with the codebook from the review analysis.

Additionally, we reverse-engineered the case study apps to understand how they detect stalkerware. In both cases, we started with *static analysis*, i.e., decompiling and inspecting their source code. We used *dynamic analysis* to verify the results and to understand run-time behavior. This allowed us to observe and inspect the output of the apps' scanning and evaluation functions for potentially harmful behavior.

3.4 User Perceptions vs. App Capabilities

Finally, we juxtapose the trustworthiness and security perceptions with theoretical samples from each case-study app to point out mismatches between perceptions and technical reality. As far as possible, we embed the perceptions and theoretical samples into related work to provide an additional broader context. We evaluate the benefits and drawbacks of users' strategies for choosing anti-stalkerware.

3.5 Ethical and Legal Considerations

Using public data for research without explicit consent is an ethical challenge, especially concerning intimate partner abuse. Even though users can remove their public reviews, we handle all data with care to minimize potential harm. We omit usernames and rephrase quotes if they contain hints of abusive behavior, rendering identification difficult.

Reverse engineering is a legal grey area. In the US, good-faith security research is exempt from copyright law and the DMCA [65]. In the EU, decompilation is explicitly allowed to ensure interoperability with other software [90]. EU copyright law only protects the concrete expression of the source code, not the underlying ideas and principles. We carefully reviewed our results to avoid publishing information that could be considered a concrete expression.

We want to minimize potential harm from publishing results of our technical analysis. After a careful review, we identified three types of potentially harmful information: (1) well-known stalkerware that apps do not identify correctly, (2) flawed general approaches to detecting stalkerware, and (3) specific implementation details about threat classification. We informed the app providers about well-known stalkerware

their app did not identify before publication. The general flaws we identified are well-known; existing spyware and state-of-the-art anti-spyware already take them into account. Hence, publishing these general flaws does not introduce new harm. Specific implementation details on how apps classify threats are out of scope for this work. Since stalkerware could use these findings to evade detection, we refrain from publishing them. Our institution's ethical review board (ERB) approved this study.

4 Users' Perceptions of Anti-Stalkerware

To understand how users perceive the security of anti-stalkerware apps, we analyzed the app-store reviews of the two apps in our case study. We included a total of 518 reviews in our study and performed thematic analysis to find higher-level themes and patterns in the data. In the following, we report the results from this analysis, i.e., our findings on users' approaches to engendering trust in anti-stalkerware apps, general observations, and contradicting user expectations.

We identified five approaches users apply to convince others of anti-stalkerware apps' usefulness and trustworthiness.

Potentially harmful incidents. First-hand experience of an apps' protection is a popular way for users to establish trust. This approach to establishing trust covers a variety of different features. Amongst others, we have found praise for adware detection, e.g., *"has already found and removed adware three times."* (R326), spyware detection, e.g., *"Someone had put a tracking app on my phone [...] I had it figured out in about 10 minutes!"* (R425), and theft prevention, e.g., *"It [...] has saved me from losing my phone not once but twice to thiefs."* (R132). Interestingly, reviewers did not seem concerned about apps' potential shortcomings in other areas. One great first-hand experience may suffice to convince users of an app's general effectiveness.

However, we also observed this effect the other way around. As soon as users have negative experiences with core features, they lose confidence. In one case, the reviewer knew that an ex-partner spied on them, but the anti-stalkerware did not detect any malicious app: *"Never purchase this! My ex is still reads my messages - it's a disgrace"* (R477). Similarly, this reviewer's trust vanished as soon as they realized they could not locate their stolen phone: *"The whole reason I have this app is in case I lose my phone."* (R069).

While effective security apps must protect users in cases of attacks, a single thwarted attack is not a good indicator of a security app's effectiveness.

Reassuring user experience. Security apps' user experience influences the users' opinions about these apps. Frequent reminders of threats, updates, or scheduled scans keep users informed about the app's activity. Generally, attacks on users' security will be rare. So that these reminders of the ongoing protection effort can add a feeling of security for users:

"Get notified my phone is secure. That makes me feel better." (R165).

Other users may see these reminders as a disruption of their regular phone use, e.g., *"the notification is permanently visible in the status bar. This is unsettling and annoying."* (R202).

For security use-cases, where apps might only rarely need to intervene, reassuring user experience is necessary to communicate that the app is still there and doing its job. However, reassuring user experience is independent of actual security. Hence, app developers may misuse this concept.

Building trust over time. Frequently, the history of app use influenced trust. Similar to human relationships, using the app over an extended period reassured users and increased their trust in the security app. We found three types of time references: establishing authority by stating experience, insufficient evidence of protection over time, and satisfaction with the absence of incidents.

In case of establishing authority, reviewers usually said they had used the app for years before telling us their verdict, e.g., *"Works as advertised have used it for years"* (R173). Some reviewers expected security apps to demonstrate their effectiveness. R476 assumed the app was a scam because they could not determine what it does: *"I cannot tell that this does anything for my phone so I think this is a rip off"*. However, other reviewers were happy and felt safer when the security app did not find anything: *"Haven't found anything yet but thats a good thing!! Feeling alot more safe."* (R475)

These contradicting positions are interesting since they demonstrate two fundamental ways users think about apps' security. In the first one, users demand evidence of functionality, even if there is nothing wrong with their smartphone. The other approach assumes the security app's effectiveness without evidence. Even though both reviewers used the same app, they ended up with different trust assessments.

Testing app's abilities. Numerous users did not wait for incidents in their day-to-day life to establish trust. They decided to test the apps' abilities. They compared the abilities of different anti-stalkerware apps, e.g., *"This app missed two spyware apps that the others detected."* (R470). Some knew they had spyware installed and checked if a particular anti-stalkerware could remove it: *"Can't find the spyware that is obviously installed on my phone."* (R512) R291 reported using an EICAR test file to check if the security app would detect it: *"Garbage. Eicar test antivirus not detected"* (R291). In this case, the reviewer successfully tested the 'lost phone' feature: *"Locating/Alarm etc always worked when tested"* (R344).

In general, testing security features is a solid way to build trust. However, comprehensively testing apps' malware detection abilities is hard. Other security features, such as the 'lost phone' feature are easier to test than malware detection's effectiveness. Hence, reviewers could have a misleading impression of their app's abilities even after testing them.

Third-party recommendations. Reviews rarely referred to third-party resources to justify their trust in anti-stalkerware apps. In one case, a friend in IT security recommended an app: “My friend who is in IT security suggested this app to me” (R131) In another case, a reviewer referred to a study: “saw a study that showed this had best spyware detection rate (but also false positives)” (R423).

Users who got anti-stalkerware recommendations from third parties have delegated trust establishment. For them, the user experience of a security app is not as crucial as for other users – they are already confident in its security.

4.1 Observations

During our analysis, we also observed other noteworthy trends among the reviews: emotional language, assemblages of security tools, and cases of tracking family members.

We found that reviewers often used emotionally loaded language. Positive reviews, such as R145, described the protection app as a sort of guardian angel: “It’s a guardian keeping an eye on my stuff”. The name of one of the apps in our case study, i.e., Lookout, might explain why reviewers make this connection. Negative reviews often used strong language when talking about the apps’ shortcomings. Such as R114, who complained about the app’s malware detection ability: “Pathetic virus support”, or R014, who just wanted to remove the app altogether: “take this Crappy off [my phone]”. However, since app-store reviews are voluntary, these observations could be due to self-selection bias, i.e., users who feel betrayed or well protected by the app submit more reviews.

Some reviewers did not evaluate the app independently from others. Instead, they considered how the app fits into their assemblage of security tools, e.g., “Nice addition to any security set up.” (R402) or “Lookout (Basic license) is good pair with Avast Mobile Security and CCleaner.” (R098). In such cases, users focus less on a specific tool’s efficacy but rather on the feature set of the entire assemblage. However, some of these tools expect to be standalone tools, which may impact the resulting user experience.

One reviewer explicitly described their use-case for the app as tracking family members. “We did not change anything but whenever I try locating my son there is an error.” (R155) We assume that parents such as these have only the best intentions for their children’s safety. However, Gosh et al. [32] found that affected children perceive their parents’ surveillance as overly restrictive and privacy-invasive. Our case also illustrates how users employ security apps to subvert their intended use-case.

4.2 Contradicting User Expectations

We found two approaches to trusting the apps in our case study: (1) trust, based on absent negative experiences with the app, and (2) no trust without proof that the app works as intended. Using the first approach increases trust in the security app the longer it runs without incidents. Users employing the

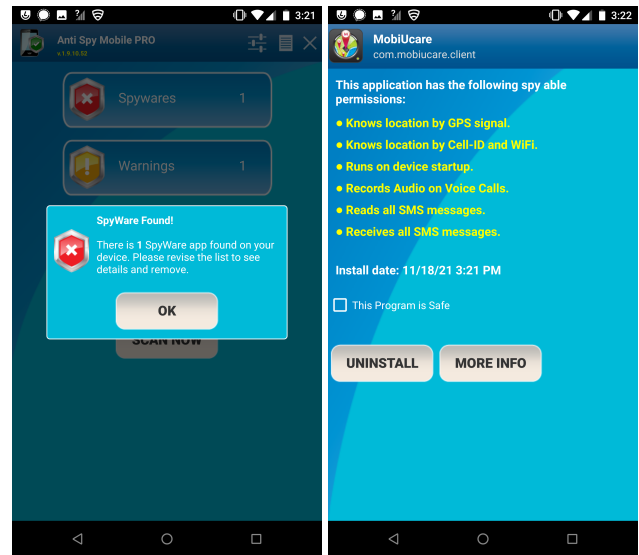


Figure 1: Anti Spy Mobile PRO’s response to a well-known spyware app.

second approach either wait until the app detects an issue or challenge the app to trigger an alert. R260 is exemplary for the first approach: “I have had this app on all my devices over the years and no problems of any kind” R215 is an example of the second approach: “I’ve not had any positive hits from this yet, so it’s difficult to say how good or bad the app is.”

The app’s user interaction impacted users’ trust in two contradictory ways: Some users thought the app was not doing anything when they could not observe any user interaction with it, i.e., they felt reassured by visible UI elements. Others interpreted the missing user interaction as a security indicator, expecting the app to respond only to security issues. R121 feels reassured when Lookout communicates that it is working: “it lets me know they are working by updating me at various time intervals and pops up on your screen when you are not thinking about them” (R121) R250 would feel more protected if Anti Spy Mobile were to indicate its ongoing operation: “there should be an anti-spy guard for the icon on the home screen. That would enhance users to feel protected and safer” (R250) In contrast, R065 is happy that the app stays silent and in the background: “it silently keeps my phone in check from the behind the curtain” (R065)

5 UI Walkthrough of Anti-Stalkerware

During our thematic analysis of the app-store reviews, we identified two approaches to how users establish trust with anti-stalkerware apps based on their user interface: (1) Incidents with potential harm and experiencing how the app handles the situation builds users’ trust; (2) Apart from potentially harmful incidents, users appreciate anti-stalkerware’s reassuring security experience during everyday use.

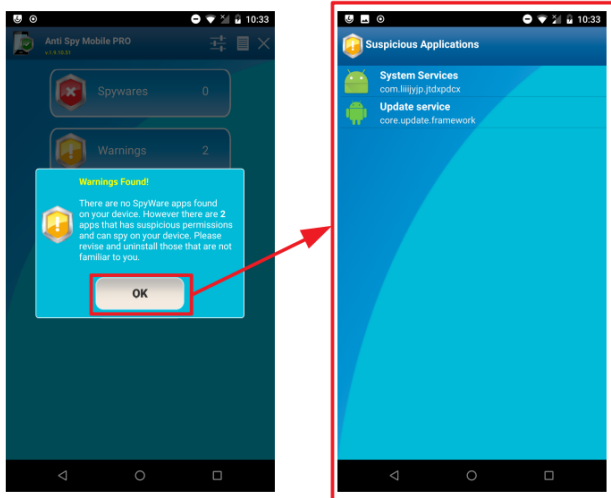


Figure 2: Anti Spy Mobile PRO’s response to spyware apps that are not on its list of well-known spyware.

This section reports the results of a cognitive walk-through [82, 93] focused on these two trust establishment approaches. For the purpose of this walkthrough, we assumed that malicious parties may have had direct access to the phone before, but they no longer do at this point. When malicious parties still have direct access, removing electronic traces of anti-stalkerware usage afterwards is necessary to keep its users safe [23]. We simulated harmful incidents by installing several spyware apps on a smartphone that we reserved for this purpose. In the resulting user interactions, we document and inspect all the parts of the UI flow and answer guiding questions about the effect on users’ trust. We simulated the day-to-day experience by using the smartphone with the installed case-study apps for 48 hours as our regular phone. We browse the web, download data, and install apps. We document and inspect user interaction and answer guiding questions about the effect on users’ trust.

5.1 Potentially Harmful Incidents

Anti Spy Mobile PRO. Opening the app shows three different classifications of apps (as buttons): (1) *Spywares* for well-known blocklisted spyware apps; (2) *Warnings* for all suspicious apps not on the blocklist; (3) *All Applications* for all other apps.

Anti Spy Mobile automatically starts a scan when users open the app for the first time. Users may trigger a scan manually with the *Scan now* button or enable automatic daily scanning in the preferences (which is the default setting). After each scan, a dialog box presents the number of identified well-known spyware apps. If it did not find any, it presents the number of suspicious apps instead. Confirming the dialog box brings users to review the apps in question (as seen in Figure 1 and 2).

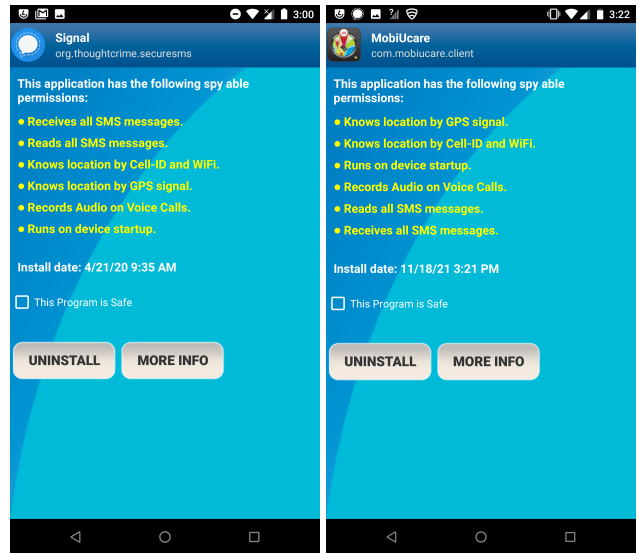


Figure 3: Additional information provided by Anti Spy Mobile PRO on a suspicious app on the left and a well-known spyware app on the right – both apps request the same “spy able” permissions.

To test Anti Spy Mobile’s reaction to a well-known spyware app, we installed MobiUcare (Phone Locator) on our test phone. Figure 1 shows the resulting “*SpyWare found*” dialog. After confirmation, Anti Spy Mobile shows the name, privacy-infringing permissions, and installations date of the detected spyware app. The “*More Info*” button would usually lead to the corresponding listing in the Google Play Store. However, this results in an error message since this app’s removal from the store.

We installed two more spyware apps: mSpy Cellphone Tracker and SpyFone. The FTC banned the latter in September of 2021 [29]. Figure 2 shows that it does not consider them well-known spyware. Instead, it informs users about suspicious apps on their phones. The text describes the classification based on requested permissions and suggests how to deal with these apps: “*you should take a close look at them and uninstall them if you are not familiar with their existence*”.

Selecting suspicious apps reveals more detailed information about them (Figure 3), such as their name, suspicious permissions, and time of installation. This view offers users three responses. First, users may want more information about the app in question. However, the corresponding button leads to the Google Play Store website, which may not provide users with sufficient threat information. With MobiUcare, the button generates an error since the app is no longer on the app store. Second, users can uninstall the app directly with a button click. However, if it concerns admin apps, this results in an error message: “*Uninstalling MobiUcare unsuccessful*”. The app does not provide any guidance in this case and acts as if the user never pressed the button in the first place. Third,

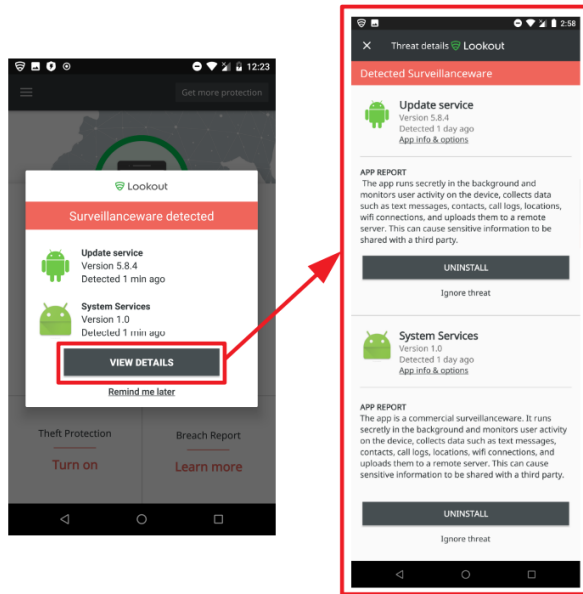


Figure 4: Lookout Mobile Security’s scan results identifying the spyware apps as surveillanceware.

if users do not want to take further action, they mark the app in question as “safe”. Then Anti Spy Mobile will stop notifying them about the app. Figure 3 shows that the threat response interface is independent of the identified threat. Anti Spy Mobile treats apps with merely suspicious permissions (the Signal messenger in this case) in the same way as apps on its list of well-known spyware.

Lookout Mobile Security. Lookout Mobile automatically scans all installed apps after installation. Users can start a scan manually at any time (see Figure 5).

To test Lookout’s response to stalkerware, we installed MobiUcare, mSpy Cellphone Tracker, and SpyFone. Lookout Mobile correctly identified all three and classified them as *Surveillanceware*. In Figure 4 a pop-up window shows all identified apps with the option to either view details or set a reminder. The remind later option does not require users to specify a time that works better for them. Such commitment devices can increase security compliance [28].

In the detailed overview, Lookout shows a classification (e.g., Surveillanceware), logo, name, version, time of detection, and an app report for each identified threat. Reports comprise three parts: a statement if the app is a commercial surveillanceware (if applicable), a list of human-readable permissions, and a generic explanation about third parties monitoring user activity without consent. The only context-dependent information seems to be Lookout’s analysis if the app in question is commercial surveillanceware.

Lookout affords users three responses for detected threats. First, users may click on App Info & options, leading them to the system’s overview of the app in question. Second, a highlighted uninstall button. While Lookout does not explicitly

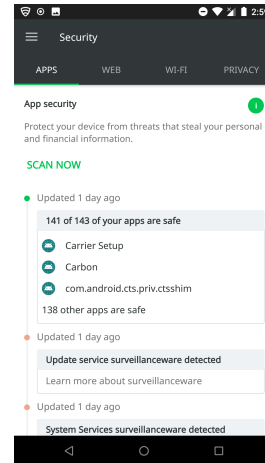


Figure 5: Dashboard of Lookout Mobile Security with scan history and re-scan option.

suggest an appropriate response to the threat, the highlighted button strongly suggests uninstalling. Lastly, it offers the option to ignore threats. Lookout does not provide users an explicit discussion of these options, not even when it identifies commercial surveillanceware.

Additionally, users have access to the scan history (see Figure 5). Upon detection of surveillanceware, this view offers users to “learn more about surveillanceware”, leading them to the built-in threat encyclopedia. The encyclopedia provides a general overview of surveillanceware abilities and only mentions a vague threat model, i.e., “Surveillanceware apps are typically installed directly by someone with physical access to the target device”. The encyclopedia avoids discussing appropriate user responses.

5.2 Reassuring Everyday Experience

Anti Spy Mobile PRO. Apart from manual scans in the app itself, Anti Spy Mobile barely interacts with users. The paid version automatically scans all apps and notifies users about the results once per day (see Figure 6). This notification does not warn about suspicious apps. Anti Spy Mobile does not intervene during day-to-day activities, such as browsing the web, downloading files, or installing apps (from the Google Play Store or third-party repositories).

Lookout Mobile Security. In general, Lookout Mobile focuses on reassuring user interaction. A sticky icon in the status bar and a permanent notification (shown in Figure 7) informs users that Lookout is active and that “everything is OK”. Another aspect of Lookout’s user interaction is its reactivity to the users’ actions. It warns users about malicious files or apps immediately after downloading or installing them, respectively. Additionally, Lookout has a setting to notify users about a WiFi network’s safety at connection time. Immediate responses improve users’ mental models when the notification links causes and effects [83].

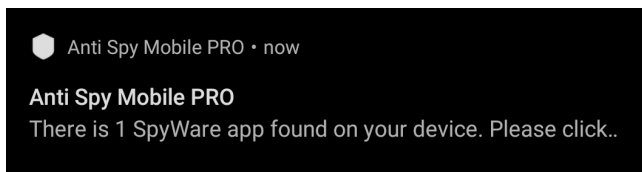


Figure 6: Anti Spy Mobile PRO’s daily scan notification.

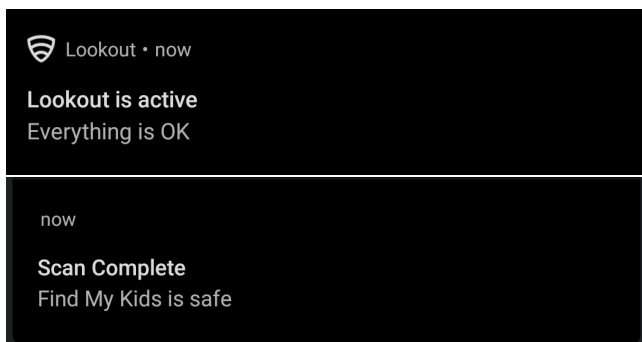


Figure 7: Lookout Mobile Security’s reassuring notifications.

Enabling Lookout’s VPN-based safe browsing feature did not affect the surfing experience. By default, Lookout analyzes downloaded files for threats (according to the description in the settings). Downloading regular files did not create a response from Lookout. However, it reacted when it detected spyware in a downloaded .apk file (Android Package, i.e., the Android app distribution format). Installing apps always created a response, regardless of the origin. Interestingly, Lookout considers the app Find My Kids safe (see Figure 7), while Anti Spy Mobile considers it well-known spyware.

6 Anti-Stalkerware under the Hood

Our thematic analysis identified two trust establishment approaches that users apply to anti-stalkerware. First, they build trust over time after seeing which threats the app caught and which it did not catch in time. Second, reviewers actively challenged the anti-stalkerware’s abilities by installing known spyware on their phones. Both approaches are based on users’ partially correct understanding of how to evaluate detection mechanisms.

To take a closer look at the detection mechanisms of our case-study apps and to understand how they determine which installed apps are threats, we performed static code analysis and dynamic run-time analysis. We follow established best practices (as outlined by OWASP [62]) for mobile app testing and rely on selected open-source tools. Android apps are typically written in Java, compiled to Dalvik bytecode, and then packaged as .apk files (essentially a zipped archive) [34]. A common first step is to transform this bytecode back into Java source code for easier comprehension. To do so, we use the Dalvik-to-Java decompiler jadx [3]. To monitor the run-time behavior of the case-study apps, we installed them on a

Nexus 5 phone and instrumented them with Frida [2]. This tool allows reverse engineers to inject and execute JavaScript in the analyzed app. We use this feature to inspect the app’s classes, methods, and data fields guided by the results of the static analysis. We further use the web proxy Fiddler [1] to intercept and inspect network traffic to the apps’ backend server, if any.

Anti Spy Mobile PRO. We started by locating the main activity of the app, representing the UI shown to users when they first open an app. The class `AntiSpyActivity.java` represents this activity and loads the start screen defined in XML format (`/resources/res/layout/start.xml`). This screen contains the *Scan Now* button, which triggers the scanner activity (`ScannerService.java`). This activity implements the core functionality of Anti Spy Mobile PRO: it calls the Android `PackageManager` [36] to get the package names of all apps installed on the device and iterates over it.

The app distinguishes between two relevant types of installed apps: *SpyWare Applications* and *Suspicious Applications*. It identifies the first category by matching apps’ package names against a list of well-known spyware apps. This blocklist of package names is embedded in the app as an XML file (`blackListPackagesDefs` in `resources/res/values/arrays.xml`). For the second category, Anti Spy Mobile PRO retrieves the apps’ requested permissions to check for “spy able” permissions related to location, microphone, and SMS access. If the sum of these weighted permissions exceeds a certain threshold, it flags an app as suspicious.

The XML file that contains the blocklist also contains an allowlist of package names (`whiteListPackagesDefs`) of apps that presumably would trigger false positives based on their permissions. This list contains for example different browsers, but interestingly also security solutions such as Lookout Mobile Security. In its current version, the blocklist contains 494 entries, while the allowlist contains 146 entries, with 30 of these package names matching apps available on the Google Play Store, respectively.

We reverse-engineered the free version (Anti Spy Mobile Basic) and confirmed that the only difference is the option to schedule automatic background scans.

We further executed Anti Spy Mobile PRO to confirm our findings from the static code analysis and inspect its behavior during the actual scanning process. During this experiment, the app classified neither of the two spyware apps `mSpy` and `SpyFone` as `SpyWare` because its blocklist does not include them. However, it classified them as suspicious based on their permissions.

Lookout Mobile Security. This app is more complex than Anti Spy Mobile PRO, both in terms of code and UI. In this case, we started by looking for the *Scan Now* button in the dashboard UI (see Figure 5). This button triggers a SQL query for the already stored results of the previous scans. We then looked at the code populating this database, which is split

across a number of different classes. We found that Lookout Mobile Security also collects information about each installed app from the Android PackageManager [36]. In addition, for apps classified as malicious, it also stores an assessment including the classification category, assessment ID, severity of the threat, and the response type.

The actual scanning mechanism is implemented both as a local and a cloud scan. In the case of a local scan, it checks for assessment in the `Policy.FLX`. This policy is distributed via over-the-air (OTA) updates, i.e., updates automatically pushed to the app without any active user interaction. For cloud scans, the app creates a request to `https://appintel.mobilethreat.net` with hashed information about the app under assessment.

Monitoring the network traffic of the app using Fiddler, we observed that during the first scan it received data from `https://ota.lookout.com`. We identified this as the source of the OTA policies, but could not identify its format. Thus, using Frida, we injected JavaScript into the process to inspect the list of assessments read from this policy file. Most of the assessments seem to be in the form of signature-based detection methods, i.e., as a blocklist. Lookout detected both spyware apps (mSpy and SpyFone) as surveillanceware based on this blocklist.

Comparison of detection mechanisms. Both Anti Spy Mobile PRO and Lookout Mobile Security detect mSpy and SpyFone, the spyware apps. However, the first app merely classifies the two spyware apps as suspicious, while the second one accurately recognizes both as surveillanceware.

Anti Spy Mobile PRO mainly works with a block- and allowlist of package names. However, package names are weak identifiers of Android apps. The Google Play Store uses it to uniquely identify apps and recommends following Java package naming convention, i.e., to “use Internet domain ownership as the basis for package names (in reverse to avoid conflicts with other developers” [35]). Still, developers can choose arbitrary or conflicting package names for their apps, particularly when they are distributed via third-party repositories. Malware authors have been known to use the tactic of imitating package names of benign apps, or randomly generating package names to evade detection [52]. The package names of mSpy (`core.update.framework`) and SpyFone (`com.rzjzmlrm.vhqpmgzo`) seem to follow this pattern. Technically, stalkerware distributors could even automatically generate new package names for each customer.

Furthermore, these lists are part of the resources embedded in the .apk file, and the app does not implement any functionality to update this file. Thus, any changes in the blocklist need to be pushed as part of app updates through the Google Play Store—which users may or may not install [59, 87]. The update history indeed includes *[UPDATE] Spyware definitions update*, but updates have been sparse since 2018 [6].

In addition to the detection based on the package name, Anti Spy Mobile PRO also flags apps as suspicious if they re-

quest permissions that could be used for spying. Nevertheless, Anti Spy Mobile PRO does not provide more information about these apps than the requested permissions to the users and does not describe or explain what these apps do.

Lookout Mobile Security, on the other hand, dynamically fetches signature-based blocklists from the server and checks for newer versions during each launch. However, in this case, the scan is a “black box”: we have no insights about the type of scans performed on Lookout’s servers and the features they base their detection on.

7 Discussion

We compare our thematic analysis results, i.e., users’ strategies for establishing trust in their installed anti-stalkerware, with our user interface walkthrough and reverse engineering results – highlighting the expectation-ability gap. Then we discuss different stakeholders’ options to reduce this gap and improve users’ anti-stalkerware decisions in the future.

7.1 Contrasting Users’ Expectations with Actual Protection Capabilities

Potentially harmful incidents. One of the ways reviewers decided to trust anti-stalkerware apps depends on their incident response. This approach relies on apps’ ability to detect incidents. Users’ trust depends on the information and user agency that apps provide. Our walkthrough revealed that Anti Spy Mobile PRO’s suspicious apps produced easily identifiable false positives – potentially decreasing users’ trust. Also, we found inconsistent results: Anti Spy Mobile considered *Find my Kids* well-known spyware, while Lookout Mobile considered it safe. This mismatch highlights the need for context-sensitive classification, especially for dual-use apps. Neither app did a great job informing users about specific threats and providing context-appropriate user agency options. For example, Anti Spy Mobile PRO offers the same information and response options, whether it concerns well-known spyware or merely suspicious apps. Reverse engineering the apps showed that Anti Spy Mobile PRO uses a package name list of well-known spyware apps and a list of well-known benign apps. Updating these lists requires an app updating the app. Lookout Mobile checks apps against local OTA policies, regularly updated from Lookout’s servers. Anti Spy Mobile PRO further uses a permission-based approach to identify suspicious apps not on the list of well-known apps, resulting in easily identifiable false positives. Hence, relying on potentially harmful incidents as a strategy to establish trust with anti-stalkerware apps comes with risks. It relies on users’ ability to recognize harmful incidents to understand if the app should have detected and prevented them. Waiting for such moments is risky. Ideally, users trust their anti-stalkerware app before they face attacks. Lastly, awarding trust in this way may deceive users. One instance where the app protected them may lead users to overgeneralize the assumed protection.

Reassuring user experience. The analyzed reviews contained praise for reassuring user interaction in benign everyday scenarios. In addition to the regular alerts in case of threats, Lookout Mobile incorporates user interface elements that communicate the current positive security status, e.g., “everything is OK”. Showing users the security mechanisms during threats as well as in benign situations helps build users’ mental models [83]. Distler et al.’s study [21] suggests that visualizing security mechanisms improves user experience. Notably, in our case study, Lookout Mobile always seemed confident in its safety assessments. In contrast, Anti Spy Mobile depends on permissions-based classification — leading to false positives. In addition, Lookout Mobile was very reactive, immediately notifying users about their actions’ safety consequences. The timing of privacy and security notices may affect users’ decisions in general [4]. Observing links between cause and effect forms users’ mental models, making this immediacy between action and response beneficial [83]. However, moderately delayed privacy feedback may be a compromise to minimize interruption [67]. Reassuring user experiences have benefits in benign situations. They improve users’ mental models and appear to improve user experience overall. The immediate response to potential threats may improve users’ mental models by linking cause and effect. The certainty of anti-stalkerware’s verdicts, warranted or not, may heighten users’ trust. Ultimately, reassuring user experiences do not make apps more secure. Hence, users who rely on this trust establishment approach are prone to deception.

Assumptions about apps’ detection capabilities. Reviews contained two approaches based on assumptions of the anti-stalkerware’s detection abilities. First, reviewers evaluated the app’s abilities over time, building trust similar to a personal relationship. Second, reviewers explicitly tested and challenged the app’s detection ability with selected spyware or test viruses. Both approaches are flawed. Using the first approach, users assume they can detect a threat when the app can not. Since they may not recognize when the app fails to detect threats, they may only be aware of incidents where the app protects them. Using the second approach, users generalize their test results from a single test to the apps’ abilities to detect other malicious software, which might seriously mislead users. Even worse, since they tested the apps’ ability personally, they put significant trust in their assessment.

Reliance on third-party evaluations. Some reviewers exclusively relied on third-party evaluations of anti-stalkerware apps. Depending on the third party may be the safest choice to establish trust. However, it also comes with drawbacks. First and foremost, trust in the third party is required — moving the issue of trust establishment from the app to the third party. Then, the third party has to have reviewed the users’ chosen app. The effectiveness of this approach relies on reputable third parties. Ideally, trusted third parties are well-known for providing fair assessments. However, social effects may

impact the choice of trusted third parties. Users rely on tech-savvy family members and friends even when they can not provide fair assessments. In any case, users can not influence and may not even know which aspects third parties consider for their reviews (e.g., usability, user agency, detection rate).

Relying on third-party reviews, users do not experience how the app reacts in case of an incident, which may affect their comfort, comprehension, and ultimately their safety.

7.2 Implications and Future Work

The thematic analysis results suggest that judging anti-stalkerware apps’ efficacy is hard for users. In the current circumstances, their safest option is to rely on IPV-specific evaluation results of certified antivirus testing labs. In the future, we should try to support and improve users’ existing evaluation approaches and give them more agency to safely build trust in anti-stalkerware apps. However, adapting apps and operating systems to make intimate partner surveillance difficult and less surreptitious would likely limit the proliferation of stalkerware and other abuse-enabling apps more effectively.

Reassuring experiences are useful (if done correctly) but cannot be trusted. One of the themes in our thematic analysis was that users felt reassured and well protected based on UI elements. The UI walkthrough confirmed that one of the apps relied on positive messaging to communicate to users about its work. Mathiasen et al. [55, 56] refer to this as *secure experiences*, which are not necessarily the same as security. According to them, users will base their security decisions on previous secure experiences. Spero et al. [83] argue that user interfaces that hide security mechanisms hinder users from building detailed mental models of security. Hence, security mechanisms should present users with model-building information, whether they face security risks or not. As an example, Distler et al. [21] found that visualizing security mechanisms in an e-voting apps led to an increase in perceived security. While these kinds of reassuring and secure experiences may be understudied, they appear to provide several benefits: (1) they communicate to users that a security system is working, even when no security risk calls for action; (2) they may improve users mental model of security; and (3) they help improve users’ security decisions later on. However, these kinds of secure experiences become a problem if they oversell the actual security, regardless of the intention. Therefore, simple reassurances that everything is safe may not be the best approach to building secure experiences. The anti-stalkerware apps in our case study probably use reassuring experiences to justify their existence to users. Without them, it may appear like anti-stalkerware apps do nothing of value, even when they work well. In summary, reassuring user experiences may improve users’ mental models and security decisions, but users cannot rely on them alone to establish trust in security mechanisms.

Demonstrate stalkerware detection to users. In our thematic analysis, we found reviewers used several different (flawed) tactics to evaluate the detection efficacy of anti-stalkerware apps. Also, we found that the anti-stalkerware’s response to stalkerware (user experience, information, and agency) affects users’ trust. Hence, it would make sense to encourage and improve this kind of evaluation behavior. We suggest offering a toolkit for users to install on their phones. This toolkit should be able to install (and remove) a wide variety of stalkerware and dual-use software and track the anti-stalkerware’s response. Such a toolkit would affect users in three ways: (1) all users would have the ability to safely and soundly evaluate their chosen tool’s detection mechanism, (2) users could safely experience their tools response to malicious software, and (3) it would reduce the need to trust third-party reviews of anti-stalkerware apps. Similar to this approach, Parson et al. [66] suggest that a government body should track and evaluate anti-virus engines and publish public reviews. However, in contrast to our suggestion, users would then not experience their chosen app’s response to threats.

Provide context-specific advice and give users agency. Detection ability is an important but not the only factor for users’ safety. The type and amount of information apps present to users influence their response. Additionally, users’ agency to respond to detected threats is crucial. Both information and agency need to be context-sensitive to the users’ circumstances and the specific detected threats. For example, for IPS survivors safe responses to detected surveillance threats may be different before and after they have left their partner. This could include additional context-specific response options, e.g., generating fake location data or partially removing permissions without alerting the stalker. Without context-sensitive advice and user options, even an anti-stalkerware app with great detection ability may endanger users.

Leverage operating system’s power to limit abuse. Improving anti-stalkerware apps and users’ protection abilities is an individualistic approach to combating IPS. However, a systemic approach may be more effective in reducing IPS. Considering potential abuse in the design stage for operating systems, apps, and accessories may help fight IPS on a system level. Defensive design is a widely adopted approach across many disciplines. However, it focuses on unintentional errors in programming code and resulting apps. Other general approaches take intentional abuse into account at every step of the design process to mitigate interpersonal harm [68, 94]. Levy and Schneier [50] offered design considerations to ameliorate intimate privacy risks. Slupska and Tanczer [81] suggested an approach to threat model intimate partner violence in the design process. Interestingly, the two most common smartphone platforms, iOS and Android, are not equally susceptible to stalkerware targeted at consumer audiences [42, 66]. Parsons et al. report on the stalkerware industry [66] and the limited options to install these stalker-

ware apps on iOS without jailbreaking. Consequently, most commercial stalkerware for iOS devices rely on the target’s iCloud account. Reputable companies do not want to publicly support dedicated stalkerware, so these apps are not published in app stores—or are quickly removed. This may result in a proliferation of other abuse-enabling dual-use apps (such as parental control apps) and their legitimate use-cases make them harder to police. Since legitimate use-cases are here to stay, it is necessary to adapt the design of these apps and the operating systems to limit misuse. The authors report recommendations applicable to platform providers that may curb stalkerware. They call for prominent, ongoing, and meaningful consent notices. These make it harder to install stalkerware surreptitiously on others’ smartphones. Additionally, they call for on-device platform heuristics that detect misuse of ostensible dual-use software. Platforms have the power to disable abuse-enabling apps entirely – which may protect users unable to manage apps on their device. Platform providers have significant power over the kind of software they allow to run and which kind of app activities they make visible to users. Using this power would be an effective measure against the current stalkerware ecosystem.

8 Conclusion

Choosing effective anti-stalkerware solutions is a struggle. This case study evaluated two anti-stalkerware apps from multiple perspectives to understand users’ selection and trust strategies. We identified five approaches that users apply: two based on user interaction, two based on the assumed detection abilities, and one on trusted third parties. All approaches are intuitive to apply and have some degree of legitimacy. However, the cognitive walkthroughs and reverse engineering approaches revealed severe drawbacks. We found that users’ strategies do not inform them sufficiently about these apps and their abilities to mitigate violence, abuse, and harassment.

Our work helps improve current anti-stalkerware by suggesting design directions that increase users’ trust and safety. These design directions focus on reassuring user experience, context-sensitive advice, and risk-appropriate user agency. Also, we suggest a user-deployable, toolkit-supported approach to evaluate anti-stalkerware’s detection abilities and user experience. Such a toolkit-based approach builds on and encourages existing user behavior while improving its efficacy and safety. Lastly, while our study focuses on individualistic responses to anti-stalkerware, we emphasize the need for a systemic, platform-level approach to effectively combat intimate partner surveillance.

Acknowledgements

We thank the reviewers for their feedback on improving our paper. In particular, we thank our anonymous shepherd for their responsive, helpful, and kind guidance. The first author

conducted their work as part of the Saarbrücken Graduate School of Computer Science, Saarland University.

This research has received funding from the Vienna Science and Technology Fund (WWTF) through project ICT19-056, as well as SBA Research. SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.

References

- [1] Fiddler | Web Debugging Proxy and Troubleshooting Solutions. <https://www.telerik.com/fiddler>. (Accessed on June 8th, 2022).
- [2] Frida • A world-class dynamic instrumentation framework. <https://frida.re>. (Accessed on June 8th, 2022).
- [3] Jadx. <https://github.com/skylot/jadx>. (Accessed on June 8th, 2022).
- [4] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys*, 50(3):1–41, October 2017.
- [5] Perna Agrawal and Bhushan Trivedi. A Survey on Android Malware and their Detection Techniques. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–6, Coimbatore, India, 2019. IEEE.
- [6] AppBrain. Anti spy mobile PRO: Changelog. <https://www.appbrain.com/app/anti-spy-mobile-pro/com.antispycell>, 2022. (Accessed on June 8th, 2022).
- [7] Apple. App security overview. <https://support.apple.com/guide/security/app-security-overview-sec35dd877d0/web>, 2021. (Accessed on June 8th, 2022).
- [8] Apple. App store review guidelines. <https://developer.apple.com/app-store/review/guidelines/>, 2021. (Accessed on June 8th, 2022).
- [9] Daniel Arp, Michael Spreitzenbarth, Malte Hübner, Hugo Gascon, and Konrad Rieck. Drebin: Effective and Explainable Detection of Android Malware in Your Pocket. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, 2014. Internet Society.
- [10] AV-Comparatives. Android Test 2019 - 250 Apps. <https://www.av-comparatives.org/tests/android-test-2019-250-apps/>, January 2019. (Accessed on June 8th, 2022).
- [11] Alessandro Bacci, Alberto Bartoli, Fabio Martinelli, Eric Medvet, Francesco Mercaldo, and Corrado Aaron Visaggio. Impact of Code Obfuscation on Android Malware Detection based on Static and Dynamic Analysis. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy - ICISSP*, pages 379–385, Funchal, Madeira, Portugal, 2018. SciTePress.
- [12] Elsa Bakiu and Emitza Guzman. Which Feature is Unusable? Detecting Usability and User Experience Issues from User Reviews. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, pages 182–187, Lisbon, Portugal, 2017. IEEE.
- [13] Rosanna Bellini, Emily Tseng, Nora McDonald, Rachel Greenstadt, Damon McCoy, Thomas Ristenpart, and Nicola Dell. So-Called Privacy Breeds Evil": Narrative Justifications for Intimate Partner Surveillance in Online Forums. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), December 2020.
- [14] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
- [15] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The Spyware Used in Intimate Partner Violence. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 441–458, San Francisco, CA, USA, May 2018. IEEE.
- [16] Ning Chen, Jialiu Lin, Steven C. H. Hoi, Xiaokui Xiao, and Boshen Zhang. AR-Miner: Mining Informative Reviews for Developers from Mobile App Marketplace. In *Proceedings of the 36th International Conference on Software Engineering, ICSE 2014*, pages 767–778, Hyderabad, India, 2014. ACM.
- [17] Jerry Cheng, Starsky H.Y. Wong, Hao Yang, and Songwu Lu. SmartSiren: Virus Detection and Alert for Smartphones. In *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services, MobiSys '07*, pages 258–271, San Juan, Puerto Rico, 2007. ACM.
- [18] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. The Role of Social Influence in Security Feature Adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15*, pages 1416–1426, Vancouver, BC, Canada, February 2015. ACM.

- [19] Alexander De Luca, Sauvik Das, Martin Ortlieb, Iulia Ion, and Ben Laurie. Expert and Non-Expert Attitudes towards (Secure) Instant Messaging. In *Proceedings of the Twelfth USENIX Conference on Usable Privacy and Security*, SOUPS '16, pages 147–157, Denver, CO, USA, 2016. USENIX Association.
- [20] Verena Distler, Gabriele Lenzini, Carine Lallemand, and Vincent Koenig. The Framework of Security-Enhancing Friction: How UX Can Help Users Behave More Securely. In *New Security Paradigms Workshop 2020*, NSPW '20, pages 45–58, Online, USA, October 2020. ACM.
- [21] Verena Distler, Marie-Laure Zollinger, Carine Lallemand, Peter B. Roenne, Peter Y. A. Ryan, and Vincent Koenig. Security - Visible, Yet Unseen? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2019)*, CHI '19, pages 1–13, Glasgow, Scotland, UK, 2019. ACM.
- [22] Paul Ducklin. The google play “Super antivirus” that’s not so super at all... <https://nakedsecurity.sophos.com/2018/01/19/the-google-play-super-antivirus-thats-not-so-super-at-all-report/>, January 2018. (Accessed on June 8th, 2022).
- [23] Martin Emms, Budi Arief, and Aad van Moorsel. Electronic Footprints in the Sand: Technologies for Assisting Domestic Violence Survivors. In Bart Preneel and Demosthenes Ikonomou, editors, *Privacy Technologies and Policy*, pages 203–214, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.
- [24] Álvaro Feal, Paolo Calciati, Narseo Vallina-Rodriguez, Carmela Troncoso, and Alessandra Gorla. Angel or Devil? A Privacy Study of Mobile Parental Control Apps. In *Proceedings on Privacy Enhancing Technologies*, volume 2020, pages 314–335, April 2020.
- [25] Raffaele Filieri. What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research*, 68(6):1261—1270, 2015.
- [26] Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ristenpart, and Nicola Dell. “Is My Phone Hacked?” Analyzing Clinical Computer Security Interventions with Survivors of Intimate Partner Violence. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):202:1–202:24, 2019.
- [27] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. “A Stalker’s Paradise”: How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, Montreal, QC, Canada, 2018. ACM.
- [28] Alisa Frik, Nathan Malkin, Marian Harbach, Eyal Peer, and Serge Egelman. A Promise Is A Promise: The Effect of Commitment Devices on Computer Security Intentions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, Scotland, UK, 2019. ACM.
- [29] FTC. FTC bans SpyFone and CEO from surveillance business and orders company to delete all secretly stolen data. <https://www.ftc.gov/news-events/pres-s-releases/2021/09/ftc-bans-spyfone-and-ceo-from-surveillance-business>, 2021. (Accessed on June 8th, 2022).
- [30] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. Why People Hate Your App: Making Sense of User Feedback in a Mobile App Store. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1276–1284, Chicago, IL, USA, 2013. ACM.
- [31] Asit Kumar Gahalaut and Padmavati Khandnor. Reverse engineering: An essence for software re-engineering and program analysis. *International Journal of Engineering Science and Technology*, 2(06):2296—2303, 2010.
- [32] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr, and Pamela J. Wisniewski. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, Montreal, QC, Canada, April 2018. ACM.
- [33] Arup Kumar Ghosh and Pamela Wisniewski. Understanding User Reviews of Adolescent Mobile Safety Apps: A Thematic Analysis. In *Proceedings of the 19th International Conference on Supporting Group Work*, GROUP '16, pages 417–420, Sanibel Island, FL, USA, 2016. ACM.
- [34] Google. Android developers: Configure your build. <https://developer.android.com/studio/build>. (Accessed on June 8th, 2022).
- [35] Google. Android developers: <manifest>. <https://developer.android.com/guide/topics/manifest/manifest-element.html#package>. (Accessed on June 8th, 2022).
- [36] Google. Android developers: PackageManager. <https://developer.android.com/reference/android/content/pm/PackageManager>. (Accessed on June 8th, 2022).

- [37] Google. Developer Program Policy: September 16, 2020 announcement - Play Console Help. <https://support.google.com/googleplay/android-developer/answer/10065487>, September 2020. (Accessed on June 8th, 2022).
- [38] Xiaodong Gu and Sunghun Kim. "What Parts of Your Apps are Loved by Users?" (T). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 760–770, Lincoln, NE, USA, November 2015. IEEE.
- [39] Emitza Guzman and Walid Maalej. How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. In *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, pages 153–162, Karlskrona, Sweden, 2014. IEEE.
- [40] Elizabeth Ha and David Wagner. Do Android users write about electric sheep? Examining consumer reviews in Google Play. In *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, pages 149–157, Las Vegas, NV, USA, 2013. IEEE.
- [41] Mahmoud Hammad, Joshua Garcia, and Sam Malek. A Large-Scale Empirical Study on the Effects of Code Obfuscations on Android Apps and Anti-Malware Products. In *Proceedings of the 40th International Conference on Software Engineering, ICSE '18*, pages 421–431, Gothenburg, Sweden, 2018. ACM.
- [42] Diarmaid Harkin and Adám Molnár. Operating-System Design and Its Implications for Victims of Family Violence: The Comparative Threat of Smart Phone Spyware for Android Versus iPhone Users. *Violence Against Women*, 27(6-7):851–875, May 2021.
- [43] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical Computer Security for Victims of Intimate Partner Violence. In *Proceedings of the 28th USENIX Conference on Security Symposium, SEC '19*, pages 105–122, Santa Clara, CA, USA, 2019. USENIX Association.
- [44] Steffen Hedegaard and Jakob Grue Simonsen. Extracting Usability and User Experience Information from Online User Reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13*, pages 2089–2098, Paris, France, 2013. ACM.
- [45] Alex Hern. 'Fake' Android antivirus app developer says Virus Shield was a 'foolish mistake'. <http://www.theguardian.com/technology/2014/apr/10/fake-android-antivirus-app-developer-virus-shield>, April 2014. (Accessed on June 8th, 2022).
- [46] Claudia Iacob and Rachel Harrison. Retrieving and analyzing mobile apps feature requests from online reviews. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 41–44, San Francisco, CA, USA, 2013. IEEE.
- [47] Hammad Khalid. On identifying user complaints of iOS apps. In *2013 35th International Conference on Software Engineering (ICSE)*, pages 1474–1476, San Francisco, CA, USA, 2013. IEEE.
- [48] Younghwa Lee and Kenneth A Kozar. An empirical investigation of anti-spyware software adoption: A multi-theoretical perspective. *Information & Management*, 45(2):109–119, 2008.
- [49] Roxanne Leitão. Technology-Facilitated Intimate Partner Abuse: A qualitative analysis of data from online domestic abuse forums. *Human-Computer Interaction*, 36(3):203–242, 2021.
- [50] Karen Levy and Bruce Schneier. Privacy threats in intimate relationships. *Journal of Cybersecurity*, 6(1), May 2020.
- [51] Martina Lindorfer, Matthias Neugschwandtner, and Christian Platzer. MARVIN: Efficient and Comprehensive Mobile App Classification through Static and Dynamic Analysis. In *2015 IEEE 39th Annual Computer Software and Applications Conference, COMP-SAC*, pages 422–433, Taichung, Taiwan, 2015. IEEE.
- [52] Martina Lindorfer, Matthias Neugschwandtner, Lukas Weichselbaum, Yanick Fratantonio, Victor van der Veen, and Christian Platzer. ANDRUBIS - 1,000,000 Apps Later: A View on Current Android Malware Behaviors. In *2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, pages 3–17, Wroclaw, Poland, 2014. IEEE.
- [53] Walid Maalej and Hadeer Nabil. Bug report, feature request, or simply praise? On automatically classifying app reviews. In *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, pages 116–125, Ottawa, ON, Canada, August 2015. IEEE.
- [54] Diogo Marques, Ildar Muslukhov, Tiago Guerreiro, Konstantin Beznosov, and Luís Carriço. Snooping on Mobile Phones: Prevalence and Trends. In *Proceedings of the Twelfth USENIX Conference on Usable Privacy and Security, SOUPS '16*, pages 159–174, Denver, CO, USA, 2016. USENIX Association.
- [55] Niels Raabjerg Mathiasen and Susanne Bødker. Threats or Threads: From Usable Security to Secure Experience?

- In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, NordiCHI '08, pages 283–289, Lund, Sweden, 2008. ACM.
- [56] Niels Raabjerg Mathiasen and Susanne Bødker. Experiencing Security in Interaction Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2325–2334, Vancouver, BC, Canada, 2011. ACM.
- [57] Tara Matthews, Kathleen O’Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. Stories from Survivors: Privacy & Security Practices when Coping with Intimate Partner Abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2189–2201, Denver, CO, USA, 2017. ACM.
- [58] Stuart McIlroy, Nasir Ali, Hammad Khalid, and Ahmed E. Hassan. Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. *Empirical Software Engineering*, 21(3):1067–1106, June 2016.
- [59] Andreas Möller, Stefan Diewald, Luis Roalter, Florian Michahelles, and Matthias Kranz. Update Behavior in App Markets and Security Implications: A Case Study in Google Play. In *Research in the LARGE: Proceedings of the 3rd International Workshop. Held in Conjunction with Mobile HCI*, pages 3–6, 2012.
- [60] Ildar Muslukhov, Yazan Boshmaf, Cynthia Kuo, Jonathan Lester, and Konstantin Beznosov. Know Your Enemy: The Risk of Unauthorized Access in Smartphones by Insiders. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '13, pages 271–280, Munich, Germany, 2013. ACM.
- [61] Duc Cuong Nguyen, Erik Derr, Michael Backes, and Sven Bugiel. Short Text, Large Effect: Measuring the Impact of User Reviews on Android App Security & Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 555–569, San Francisco, CA, USA, 2019. IEEE.
- [62] OWASP. Mobile security testing guide (MSTG). <https://mobile-security.gitbook.io/>, 2021. (Accessed on June 8th, 2022).
- [63] Danny Palmer. Can you trust your Android antivirus software? Malicious fake protection apps flood Google Play Store. <https://www.zdnet.com/article/can-you-trust-your-mobile-antivirus-software-malicious-fake-protection-apps-flood-google-play-store/>, June 2017. (Accessed on June 8th, 2022).
- [64] Sebastiano Panichella, Andrea Di Sorbo, Emitza Guzman, Corrado A. Visaggio, Gerardo Canfora, and Harald C. Gall. How can i improve my app? Classifying user reviews for software maintenance and evolution. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 281–290, Bremen, Germany, September 2015. IEEE.
- [65] Sunoo Park and Kendra Albert. A Researcher’s Guide to Some Legal Risks of Security Research. https://clinic.cyber.harvard.edu/files/2020/10/Security_Researchers_Guide-2.pdf, 2020. (Accessed on June 8th, 2022).
- [66] Christopher Parsons, Adam Molnar, Jakub Dalek, Miles Kenyon, Bennett Haselton, Cynthia Khoo, and Ronald Deibert. The Predator in Your Pocket: A Multidisciplinary Assessment of the Stalkerware Application Industry. <https://citizenlab.ca/docs/stalkerware-holistic.pdf>, 2019. (Accessed on June 8th, 2022).
- [67] Sameer Patil, Roberto Hoyle, Roman Schlegel, Apu Kapadia, and Adam J. Lee. Interrupt Now or Inform Later? Comparing Immediate and Delayed Privacy Feedback. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1415–1418, Seoul, Republic of Korea, April 2015. ACM.
- [68] Eva PenzeyMoog. *Design for Safety*. A Book Apart, August 2021.
- [69] Minh Vu Phong, Tam The Nguyen, Hung Viet Pham, and Tung Thanh Nguyen. Mining User Opinions in Mobile App Reviews: A Keyword-Based Approach (T). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 749–759, Lincoln, NE, USA, 2015. IEEE.
- [70] Georgios Portokalidis, Philip Homburg, Kostas Anagnostakis, and Herbert Bos. Paranoid Android: Versatile Protection for Smartphones. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, Austin, TX, USA, 2010. ACM.
- [71] Mila Dalla Preda and Federico Maggi. Testing android malware detectors against code obfuscation: A systematization of knowledge and unified methodology. *Journal of Computer Virology and Hacking Techniques*, 13(3):209–232, August 2017.
- [72] Emilee Rader and Rick Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, September 2015.
- [73] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as Informal Lessons about Security. In *Proceedings of*

the Eighth Symposium on Usable Privacy and Security, SOUPS '12, Washington D.C., USA, 2012. ACM.

- [74] Mizanur Rahman, Nestor Hernandez, Ruben Recabarren, Syed Ishtiaque Ahmed, and Bogdan Carbunar. The Art and Craft of Fraudulent App Promotion in Google Play. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19*, pages 2437–2454, London, United Kingdom, 2019. ACM.
- [75] Fahimeh Raja, Kirstie Hawkey, Pooya Jaferian, Konstantin Beznosov, and Kellogg S. Booth. It's Too Complicated, so i Turned It off! Expectations, Perceptions, and Misconceptions of Personal Firewalls. In *Proceedings of the 3rd ACM Workshop on Assurable and Usable Security Configuration, SafeConfig '10*, pages 53–62, Chicago, IL, USA, 2010. ACM.
- [76] Audrey Randall, Enze Liu, Gautam Akiwate, Ramakrishna Padmanabhan, Geoffrey M. Voelker, Stefan Savage, and Aaron Schulman. Trufflehunter: Cache Snooping Rare Domains at Large Public DNS Resolvers. In *Proceedings of the ACM Internet Measurement Conference, IMC '20*, pages 50–64, Virtual Event, USA, 2020. ACM.
- [77] Vaibhav Rastogi, Yan Chen, and Xuxian Jiang. Droid-Chameleon: Evaluating Android Anti-Malware against Transformation Attacks. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security, ASIA CCS '13*, pages 329–334, Hangzhou, China, 2013. ACM.
- [78] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How I Learned to be Secure: A Census-Representative Survey of Security Advice Sources and Behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16*, pages 666–677, Vienna, Austria, 2016. ACM.
- [79] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. I Think They're Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288, San Jose, CA, USA, 2016. IEEE.
- [80] Kevin A. Roundy, Paula Barmaimon Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, and Acar Tamersoy. The Many Kinds of Creepware Used for Interpersonal Attacks. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 626–643, San Francisco, CA, USA, 2020. IEEE.
- [81] Julia Slupska and Leonie Maria Tanczer. Threat Modeling Intimate Partner Violence: Tech Abuse as a Cybersecurity Challenge in the Internet of Things. In Jane Bailey, Asher Flynn, and Nicola Henry, editors, *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 663–688. Emerald Publishing Limited, June 2021.
- [82] Rick Spencer. The Streamlined Cognitive Walkthrough Method, Working around Social Constraints Encountered in a Software Development Company. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '00*, pages 353–359, The Hague, The Netherlands, 2000. ACM.
- [83] Eric Spero and Robert Biddle. Out of Sight, Out of Mind: UI Design and the Inhibition of Mental Models of Security. In *New Security Paradigms Workshop 2020, NSPW '20*, pages 127–143, Online, USA, 2020. ACM.
- [84] Coalition Against Stalkerware. The State of Stalkerware in 2019. https://media.kasperskycontenthub.com/wp-content/uploads/sites/100/2020/03/18084439/Kaspersky_The-State-of-Stalkerware-in-2019_Updated.pdf, April 2020. (Accessed on June 8th, 2022).
- [85] Google Play Store. Anti spy mobile PRO. <https://play.google.com/store/apps/details?id=com.antispycell>, 2021. (Accessed on June 8th, 2022).
- [86] Google Play Store. Mobile security - lookout. <https://play.google.com/store/apps/details?id=com.lookout>, 2021. (Accessed on June 8th, 2022).
- [87] Yuan Tian, Bin Liu, Weisi Dai, Blase Ur, Patrick Tague, and Lorrie Faith Cranor. Supporting Privacy-Conscious App Update Decisions with User Reviews. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices, SPSM '15*, pages 51–61, Denver, CO, USA, 2015. ACM.
- [88] Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infidelity Forums. In *Proceedings of the 29th USENIX Conference on Security Symposium*, pages 1893–1909. USENIX Association, 2020.
- [89] Kami E. Vaniea, Emilee Rader, and Rick Wash. Betrayed by Updates: How Negative Experiences Affect Future Security. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14*, pages 2671–2674, Toronto, ON, Canada, 2014. ACM.
- [90] Arne Vidstrom. The legal boundaries of reverse engineering in the EU. <https://vidstromlabs.com/blog/the-legal-boundaries-of-reverse-engineering-in-the-eu/>, May 2019. (Accessed on June 8th, 2022).

- [91] Artemij Voskoboynikov, Oliver Wiese, Masoud Mehrabi Koushki, Volker Roth, and Konstantin Beznosov. The U in Crypto Stands for Usable: An Empirical Study of User Experience with Mobile Cryptocurrency Wallets. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [92] Rick Wash. Folk Models of Home Computer Security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 1–16, Redmond, WA, USA, 2010. ACM.
- [93] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. The Cognitive Walkthrough Method: A Practitioner’s Guide. In *Usability Inspection Methods*, pages 105–140. John Wiley & Sons, Inc., 1994.
- [94] Karl Wieggers. Designing around bad actors and dangerous actions. <https://uxdesign.cc/designing-around-bad-actors-and-dangerous-actions-8fc7984c510d>, February 2021. (Accessed on June 8th, 2022).
- [95] Zhen Xie and Sencun Zhu. AppWatcher: Unveiling the Underground Market of Trading Mobile App Reviews. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, pages 1–11, New York, NY, USA, 2015. ACM.
- [96] Zhen Xie, Sencun Zhu, Qing Li, and Wenjing Wang. You Can Promote, but You Can’t Hide: Large-Scale Abused App Detection in Mobile App Stores. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, ACSAC '16, pages 374–385, Los Angeles, CA, USA, 2016. ACM.
- [97] Jinjian Zhai, Humayun Ajmal, and Jimmy Su. Preying on Insecurity: Placebo Applications With No Functionality on Google Play and Amazon.com. <https://www.fireeye.com/blog/threat-research/2014/06/preying-on-insecurity-placebo-applications-with-no-functionality-on-google-play-and-amazon-com.html>, June 2014. (Accessed on June 8th, 2022).
- [98] Min Zheng, Patrick P. C. Lee, and John C. S. Lui. ADAM: An automatic and extensible platform to stress test android anti-virus systems. In Ulrich Flegel, Evangelos Markatos, and William Robertson, editors, *Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 82–101, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- [99] Yixin Zou, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. Examining the Adoption and Abandonment of Security, Privacy, and Identity Theft Protection Practices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Honolulu, HI, USA, 2020. ACM.

A Codebook for the Thematic Analysis

Table 1 shows the initial codebook. Table 2 shows the codebook we used to focus on the users’ perception of the case-study apps’ safety and security.

Table 1: Initial codebook that included users' general perceptions about the apps.

CODES	DESCRIPTION	ANTISPY	LOOKOUT	TOTAL
Effect +	Review reports an event that demonstrated the app's efficacy	119	41	160
		110	39	149
Experience +	Review focuses on the app's great user experience	155	19	174
		161	12	173
Performance +	Review highlights the technical performance of the app (e.g., quick scans or low battery drain)	40	12	52
		37	14	51
Usability +	Review reports that the app is easy to understand and/or use	20	11	31
		20	10	30
Payment +	Positive experience with payment for the app itself or the subscription	18	3	21
		19	3	22
Response +	Positive experience with responsive app developers or support team	8	10	18
		7	12	19
Privacy +	Reviewer praises the app for its privacy-preserving approach	5	4	9
		4	5	9
Effect -	Review reports an event that demonstrated the app's inadequacy	28	27	55
		27	23	50
Experience -	Review focuses on the app's bad user experience	1	1	2
		2	0	2
Performance -	Review highlights the bad technical performance of the app (e.g., battery drain, slow scans, or bugs)	94	18	112
		101	18	119
Usability -	Review reports that the app is hard to understand and/or use	47	15	62
		42	18	60
Payment -	Negative experience with payment for the app itself or the subscription	48	20	68
		44	21	65
Response -	Negative experience with unresponsive app developers or support team	22	2	24
		21	2	23
Privacy -	Reviewer perceives the app as privacy-infringing	5	4	9
		4	3	7

Table 2: The codebook for the second coding iteration that focused on the users' perception of the app's effectiveness, i.e. the *effect* code in the previous codebook.

CODES	DESCRIPTION	ANTISPY	LOOKOUT	TOTAL
Real Life Safe	Experience report of an event where app protected reviewer from harm	52	27	79
		50	26	76
Test passed	Reviewer tested the app's detection capabilities and was satisfied by the results	5	3	8
		6	3	13
Secure Feeling	Experience of using the app gave reviewer a feeling of security	67	9	76
		70	9	79
Notifications	Prompt notifications about security incidents gave reviewers a secure feeling	17	2	19
		19	2	21
Real Life Fail	Experience report of an event where app failed to protect reviewer from harm	13	8	21
		12	8	20
Test Fail	Reviewer tested the app's detection capabilities and was not satisfied by the results	13	8	21
		15	9	24
Insecure Feeling	Experience of using the app did not reassure reviewer about its security	8	10	18
		9	10	19
Likes Feature	Reviewer praise a specific feature of the app	10	3	13
		10	3	13
Misses Feature	Reviewer complains about a feature they had before or would like to have	27	5	32
		24	5	29
Update	Review concerned changes to the app by a software update	13	2	15
		14	2	16
Time of Experience	Reviewers reference their long usage experience with the app to communicate their trust in the app's capabilities	19	3	22
		19	4	23

Users' Perceptions of Chrome's Compromised Credential Notification

Yue Huang

University of British Columbia

Borke Obada-Obieh

University of British Columbia

Konstantin Beznosov

University of British Columbia

Abstract

This paper reports the challenges that users experienced and their concerns regarding the Chrome compromised credentials notification. We adopted a two-step approach to uncover the issues of the notification, including qualitatively analyzing users' online comments and conducting semi-structured interviews with participants who had received the notification. We found that users' issues with the notification are associated with five core aspects of the notification: the authenticity of the notification, data breach incidents, Google's knowledge of users' compromised credentials, multiple accounts being associated with one notification, and actions recommended by the notification. We also identified the detailed challenges and concerns users had regarding each aspect of the notification. Based on the results, we offer suggestions to improve the design of browser-based compromised credential notifications to support users in better protecting their online accounts.

1 Introduction

The widespread availability of usernames and passwords exposed by data breaches remains a big threat to users and organizations. According to the Verizon 2021 data breach investigations report [9], credentials are the primary means by which an attacker hacks into an organization, with 61% of breaches attributed to leveraged credentials. By using the breached credentials, an adversary can try to log into other systems based on the assumption that users often reuse their credentials across multiple systems [18, 23, 88]. Credential stuffing, as this is known, is dangerous to both users and organizations.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022, August 7–9, 2022, Boston, MA, United States.

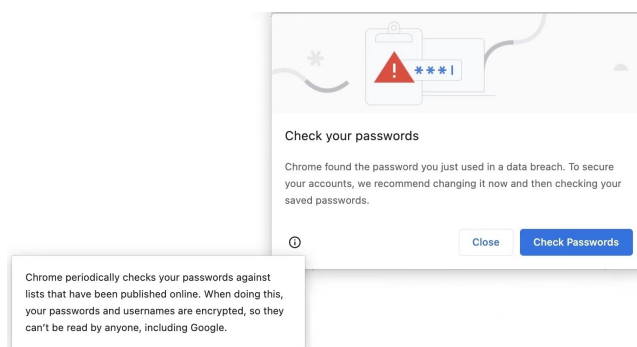


Figure 1: Chrome's Pop-up Compromised Credential Notification

For instance, in 2020, the credentials (i.e., username-password pairs) of over 530,000 Zoom teleconferencing accounts were found for sale on the dark web [108]. The credential information was not from any breach at Zoom itself; it was obtained through credential stuffing. This incident led many companies worldwide, including Google, SpaceX, and NASA [71], to ban the use of Zoom [108] and other video conferencing apps. If their accounts are hijacked, users can lose access to important information and documents and even suffer from fraudulent transactions, unauthorized fund transfers, other financial losses as well as impersonation [61].

In response, service providers and product developers started alerting users when their credentials appear in breaches. Compromised credential checking [118] has been adopted in browsers [94, 103], password managers (PMs) [101], browser extensions [20], and mobile devices (e.g., iPads [56] and smartphones [83]) to notify users when their passwords and/or usernames appear in the leaked data sets. For instance, Have I Been Pwned (HIBP) [59] is a website that allows users to check whether their personal data (e.g., phone number) has been compromised by data breaches. Browsers such as Firefox [107] and Microsoft Edge [103] are making use of HIBP to warn their users about leaked pass-

words. Google uses a similar approach to alert Chrome users if any username-password pairs saved in their Google account have been breached [47]. Specifically, whenever a user signs in to or registers on a site, a pop-up *notification* is triggered if the credentials used have been found in a data breach [43] (see Figure 1).

The notification about compromised credentials is different from warnings about an invalid TLS certificate, phishing, or other security issues. For instance, a phishing warning is often presented when a web page is considered suspicious [122]. In other words, no harm has been done yet (e.g., users have not been tricked into providing personal information) when the phishing warning pops up. In contrast, the notification of compromised credential alerts users that their credentials have already been leaked. The notification nudges users to take action to reduce the risk of account hijacking.

Prior studies focused on security *warnings* about phishing, malware, and invalid certificates. Researchers discovered that most people do not pay attention to the warnings [99], do not read the warning text [106] or do not fully understand it [12, 14], are unaware of the risks behind the warning [26], and simply fail to act on the warnings [40]. Design guidelines [10, 31, 53] and mechanisms (e.g., polymorphic warnings [16]) have been implemented to help users better understand the warnings [14] and respond to them [12, 40].

Users' perceptions about the browser-based compromised credential notification have received little attention. The most relevant work was conducted by Redmiles [95], who studied participants' responses to suspicious login incidents on their Facebook accounts. The results suggest that users often seek out additional information to understand the incident, that their threat models affect their understanding of the incident, and that their response behaviors are informed by their understanding of the incident. Other studies report that users' awareness of credentials compromises was so low that they might not take effective action (e.g., reset passwords) [12] or might not act until long after they receive a password breach email (i.e., a mean time of 26.3 days) [58]. However, no study has yet been conducted to specifically investigate users' perceptions of the browser-based compromised credential notification.

As compromised credential checking by web browsers is gaining popularity, there is a need to understand end users' perceptions. Differing from the notification of breached credentials of a certain account (e.g., Facebook accounts [95]), compromised credential notifications from browsers alert users concerning all credential information for an account that was potentially exposed in credential breaches. Millions of users have received such notifications [22], yet end users' perceptions, especially the issues and concerns they may have, have not been studied. An investigation of the challenges users are facing can inform the future design of such notifications to improve the user experience and help to better protect their accounts. Since Chrome has the greatest market share among

web browsers [1], our study focused on the perceptions of Chrome users who had received a Chrome compromised credential notification (referred to in this paper as "3CN").

We conducted our investigation through analysis of online comments and interviews with participants. By analyzing users' online comments, we discovered various challenges they experienced and concerns they had regarding the 3CN. We later explored the reasoning behind the identified issues through semi-structured interviews with participants who had received at least one 3CN.

Our work makes the following contributions. First, to the best of our knowledge, our work is the first to investigate the challenges and concerns of users in relation to browser-based compromised credential notification. Second, we discovered that users' issues with the 3CN were associated with five core aspects of the notification. We also reported the detailed challenges and concerns users had regarding each core aspect of the notification. Last, we made design suggestions about better ways to communicate risks to users, to improve users' risk comprehension, to address users' concerns, and to motivate users to take action to protect their online accounts.

2 Background and Related Work

2.1 Google Password Checkup

Google's Password Checkup allows users to check the security of the passwords that they have saved in Chrome's password manager. This feature was originally released as a Chrome extension in 2019 [94] and was integrated into the browser in October 2019. As of February 2022, it is turned on by default in Chrome, but it can be turned off manually [47].

There are two ways for users to learn about their exposed passwords and usernames. In the first case, by turning on the Chrome setting "Warn you if your passwords are exposed in a data breach," users will get a pop-up notification on the website where they try to log in or register with exposed credentials (see Figure 1) [46]. The content of the 3CN has been updated several times with minor changes [54, 86] to convey the same takeaway message – the user's credentials have been found publicly online, and the user is advised to change the compromised passwords. From the moment the notification pops up, users have two options: click on "Close" to shut the notification or click on "Check Passwords" to be directed to <chrome://settings/passwords> to see the general information about their saved accounts. By clicking on "Check Passwords," users are directed to see all the detected issues with their saved credentials, including "Compromised passwords," "Weak passwords," and "Reused passwords," if there are any. For each account listed on the page, users can see the account's username, check the current password for the account, edit the saved credentials of the account, or remove the saved account (see Figure 2a). If users wish to change the password of an account, they are directed to the website to

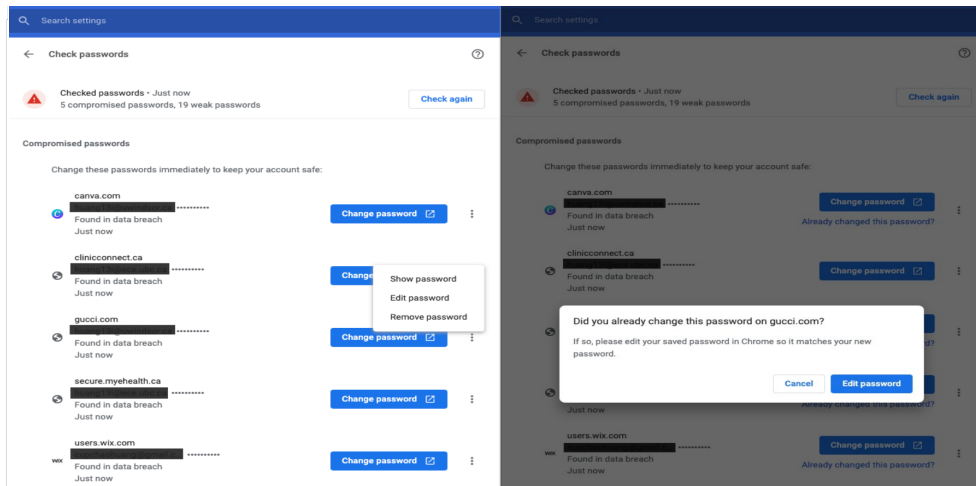


Figure 2: Screenshots of the researcher’s Google account passwords displayed when they click “Check Passwords” on the 3CN warning. Users can show, edit, or remove passwords by clicking the three-dot menu (a) or can update their saved passwords in Chrome (b).

make the change there. After clicking on the “Change password” button (see Figure 2b), a note of “Already changed this password?” is shown under the button. Users are then directed to update their saved password to match their new password on Chrome. The second way to check password security is to manually go through several steps in the browser (i.e., Open Chrome → Settings → Passwords → Check Passwords) to get to the same page to learn about the issues that Chrome password manager has identified [47].

Chrome password manager never learns the plaintext of user credentials during password checking. By using multiple rounds of hashing, k-anonymity, and private set intersection with blinding [45, 111], Google can tell whether a user’s credentials are compromised without knowing their unsafe username-password pair exposed by the data breach [46]. Specifically, Chrome first encrypts users’ credentials and sends the encrypted credentials to Google servers to compare against an encrypted list of known leaked credentials. If the Google servers detect a match between the encrypted credentials, Chrome displays the 3CN that suggests the user change their password [45]. The detailed protocol of Google’s Password Checkup is described in [111], and a simplified illustration of the protocol can be found in [45].

2.2 Risk Communication and Warnings

The main goal of risk communication is to inform individuals of risks so that they can make informed decisions [77]. Experts usually design the communication and deliver it to individuals. The communication can take the form of warnings, notices, status indicators, and polices [35]. It has been found that the mental models of technical experts and users are not always the same [73]. Therefore, one cannot assume that the

experts recognize what users need to know [32]. Guidelines have been proposed to improve the design of risk communication [24, 81], such as dispelling misconceptions [13].

As one type of risk communication, security warnings have received considerable attention. Much work has been done to evaluate the various types of security warnings, including browser warnings in general [6, 10] and warnings about phishing [29, 90], malware [7], invalid certificates [5, 31], and PDF downloads [7]. For instance, Akhawe and Felt [6] conducted a field study to investigate people’s perceptions about Google Chrome’s and Mozilla Firefox’s malware and phishing warnings. They found that the warnings were effective in practice and suggested communicating security information to users.

Many issues regarding the security warnings were identified. Studies have shown that most people do not pay attention to the computer warnings [99], often do not read the warning messages [16, 106], or do not fully understand the warning [12, 26] because of the technical words used [14, 36]. Users become habituated to security warnings [63, 66], and they end up not heeding them [40], even when the situation is hazardous or sensitive (e.g., online banking) [99].

Methods and guidelines have been proposed to motivate users to act on security warnings. For instance, varying the appearance of warnings (i.e., polymorphic warnings [16]) can help capture users’ attention and convince them to take action to mitigate a hazard [31]. Showing the warnings less frequently has been shown to reduce the habituation effect [66, 121]. Attractors (e.g., icons, images, and colors) can be effective in attracting users’ attention [15, 120]. Guidelines about how to design warnings also have been discussed [29, 31]. Suggested by Harbach et al. [53], several steps should be taken to reduce the text’s difficulty as perceived by the user, such as keeping headlines simple, using

as few technical words as possible, and using short sentences.

2.3 Password Breaches

Researchers have explored users' responses after password breaches. Shay et al. [100] investigated users' perceptions about account hijacking. They found that users believed they share responsibility for keeping the accounts secure. Redmiles [95] explored how users respond to a suspicious login incident on their Facebook account. The results showed that participants may reach out for support to understand the incident. Participants' responses included on-platform behaviors (e.g., changing passwords) and off-platform behaviors (e.g., adjusting the security setting). Bhagavatula et al. [11] examined whether and how constructively users changed their passwords after a breach announcement and found that even though the participants were likely to be affected, that few users took action. Huh et al. [58] evaluated users' reactions upon receiving a LinkedIn password reset email and discovered that only 46% participants reset their passwords.

2.4 Password Reuse

People often reuse their passwords across accounts. One common strategy for users to cope with a large number of accounts is to reuse passwords across different accounts [23, 104]. People report that the more accounts they have, the more they reuse passwords across accounts [23, 85]. Researchers have also investigated how people reuse their passwords. Users' choice of passwords depends on whether they use the accounts frequently or perceive a greater need for account security. Some people reuse passwords that they have to enter frequently [116], and other people tend to reuse passwords on infrequently used accounts because those accounts were considered to have "less need for security" [104]. Furthermore, other studies [37, 85] suggested that people tend to reuse passwords more on low-importance accounts and avoid reusing passwords for high-importance accounts.

2.5 Password Managers

Password managers (PMs) can help users centrally store, organize, and auto-fill passwords for local applications and online services. There are three primary categories of password manager implementations: built into the browser (e.g., Firefox Monitor [107]), standalone password managers (e.g., 1Password [101] and LastPass [68]), and password management within operating systems (e.g., Keychain Access on Mac [119]).

Studies have been done to explore people's perceptions about PMs. Researchers have investigated the factors that influence people's intention to adopt PMs [62, 104, 105], users' PM use [78, 89, 102], and perceived issues with PMs [48, 64]. For instance, Karole et al. [64] conducted a comparative

usability study of three PMs and found that users' comfort level with giving control to password managers influences their perceptions of the PMs.

To the best of our knowledge, our work is the first to study users' challenges with a browser-based compromised credential notification. We discovered five core aspects of the notification with which users had issues. We further identified the detailed challenges users experienced and concerns they had regarding each aspect. Our qualitative analysis of online comments and interviews allowed us to investigate not only *what* problems users faced with notifications, but also *why* these were problems. We believe these insights can improve notification design and better secure users' online accounts.

3 Method

We used a two-step approach to investigate the issues with 3CN. We first gathered and analyzed reviews, feedback, comments, and support requests posted on online platforms about 3CN. This approach allowed us to uncover a wide range of issues and concerns users had regarding the notification. Unless otherwise noted, we refer in this paper to all these types of collected data as "*comments*." To better understand users' reasoning for their concerns and challenges, we then conducted interviews with Chrome users who had received a 3CN. In this section, we describe our online comments collection, interview process, data analysis, and our method's limitations.

3.1 Data Collection

3.1.1 User Comments

We collected comments because they are considered promising and helpful data for studying users. Such comments contain a wealth of information about users' opinions, challenges, and experiences with systems and services [57, 60]. The abundance of online comments can be reliable and relevant indicators of the quality of the services and products from users' perspectives [75]. Analyzing user reviews has been frequently used by developers and researchers to understand and evaluate issues with many products, including mobile applications [70, 114], e-commerce services [74, 124], and websites [57, 117].

We gathered users' comments from various online platforms. The platforms included the Google Chrome Help Center [41], Reddit [3], news websites (e.g., The Verge [51]), IT support sites (e.g., WeLiveSecurity [34]), and Q & A websites (e.g., Quora [2]).¹ As our focus was on the issues and concerns users had regarding the 3CN, Chrome Help Center support requests [41] were the primary source for gathering users' comments. Specifically, we employed a keyword researching method [69] to search the Chrome Help Center

¹See the list of online platforms at https://github.com/AUXResearcher/SOUPS102/blob/main/Online_sources.pdf.

using several keywords or phrases, such as “password notification,” “compromised credentials,” and “password pop-up alert.” We also used the Google search engine to search the web for the same keywords, filtering the returned pages for those that were indeed about the 3CN and contained users’ comments. We excluded pages without user comments (e.g., news websites [87]), pages about Chrome’s phishing warning [42], and page about other subjects unrelated to the 3CN. We then manually checked the comments posted on each web page to ensure that they contained sufficient information regarding users’ perceptions, concerns, or actions regarding the 3CN. We excluded comments that contained insufficient information (e.g., a comment on [55] that stated “*same issue*”). We stopped searching and collecting comments when data saturation was reached (§3.2). Users whose comments we included in the study are referred to as OC-users (online comment users).

Demographic Categories		# of Participants
Gender	Male	11
	Female	11
Age	19–29	6
	30–39	7
	40–49	4
	50–59	3
	60 or above	2
Educational level	High school	2
	Bachelor	9
	Community college	2
	Master	6
	Post-graduate	1
	University below bachelor	1
Occupations	Apprenticeship	1
	Student	2
	Retired	2
	Software developer	2
	Accountant	1
	An intervention worker	1
	Occupation therapist	1
	Theater technician	1
	Product developer	1
	Sport official	1
	Stay-at-home mom	1
	Business intelligence manager	1
	Dermatologist	1
	Business owner	1
	Landscaper	1
	Farmer	1
	Project manager	1
	IT specialist	1
	Unemployed	1
Salesperson	1	

Table 1: Summary of participants’ demographics

3.1.2 Interviews

After gaining a sense of users’ issues with the 3CN, we conducted semi-structured interviews with users who had received such a notification. Participants were recruited using Facebook advertisements. They were asked to fill out an eligibility survey.² To be eligible to participate in the study,

²See the screening survey at https://github.com/AUXResearcher/SOUPS102/blob/main/Screening_Survey.pdf.

they had to have received a 3CN within the two weeks before filling out the survey. This study was approved by UBC’s research ethics board. Note that we did not recruit our interview participants from among OC-users.

The interviews served as a complementary approach to better explore users’ reasons for their concerns, challenges, and actions (if any) regarding the notification. During each interview, we asked open-ended questions to facilitate in-depth discussion with the participants [27, 82]. We focused on exploring participants’ reasoning about their concerns, challenges, and actions (if there were any) regarding the 3CN. For instance, during the interviews, we were able to explore participants’ reasoning for not acting on the notification. Specifically, some OC-users did not change passwords for accounts they perceived as unimportant. We discovered through interviews that participants viewed accounts that do not have personal or financial information as unimportant (§4.6).

Our interviews focused on four topics.³ First, we gained a basic understanding of how users interact with Chrome to manage their credentials. We asked such questions as, “For what kinds of accounts do you save your credentials using Chrome and why?” and “For which accounts do you reuse your passwords and why?” Second, we explored participants’ experiences of receiving the 3CN by asking such questions as, “What is your impression of the 3CN?” Next, we explored users’ understanding of 3CN, their concerns about it (if there were any), and their actions afterwards. We asked such questions as, “How do you think Chrome finds out about your breached credentials?” To better explore users’ reasoning behind their concerns and actions, we asked follow-up questions. For instance, when a participant chose to change passwords for only some accounts, we explored their reasons behind such an action. Finally, to further explore users’ unmet needs, we asked participants whether there was anything they would want to know regarding 3CN.

3.2 Data Analysis

We qualitatively analyzed users’ comments. Similar to many prior studies (e.g., [19, 38]), we qualitatively analyzed the comments using thematic analysis. Thematic analysis is a widely used form of analysis within qualitative research that allows patterns (i.e., themes) within the data to be identified [8, 109]. Specifically, we copied each relevant comment into a spreadsheet with the username of the person who posted the comment (referred to as “OC-user”), the time the comment was posted, the content of the comment, and other information we found relevant to the study (e.g., the screenshot of the pop-up warning the user shared). We then analyzed the comments by generating codes mapped to relevant and important pieces of information in the comments. This allowed us to develop a codebook. Once all the comments were coded, we sorted

³See the interview guide at https://github.com/AUXResearcher/SOUPS102/blob/main/Interview_Guide.pdf.

and grouped similar codes into themes. Then we reviewed and revised the themes to ensure that each one was accurately represented in the data. At this stage, we merged or broke down themes as necessary [109].

We also conducted a thematic analysis of the interview data. We started interview coding with the codebook developed from analyzing the comments. Following the same steps, we identified new codes and newly emerged themes. The combination of online comments and interviews allowed us to capture a more extensive picture of users' challenges and concerns, as well as their reasons behind them.

3.3 Limitations

Our study has several limitations. First, while we are confident that we reached data saturation during our analysis, we reviewed comments from a limited number of sources. There is also a chance that people used different usernames and went on different sites asking for help about the same issues. We might have missed web pages that were not returned by the search engine because of our choice of search keywords.

Second, because of the nature of interviews, our data are self-reported, which is always subjective [80] and may introduce selective memory bias [92]. Further, due to the nature of qualitative research, our study and our data are not amenable to generalizable quantification, such as the extent of the concerns in the target population. Our results point only to the existence of the identified concerns.

Last, with the end-goal of informing the future design of 3CN to help users better protect their online accounts, we focused on exploring the interview participants' considerations of the notification, instead of participants' individual differences (e.g., cultural background, educational background, or previous experience with data breaches). Future studies could be conducted to investigate whether and how people's individual differences correlate with their perceptions of the 3CN.

4 Results

4.1 Data Description

We collected 539 online comments from 81 sources. Each comment was posted using different usernames. Sources included 48 Google Chrome Help Center pages, 5 IT support sites, 3 Q & A websites, 4 news websites, and 20 Reddit posts. The earliest comment was posted on December 17, 2019, and the last on July 8, 2021. The longest comment contained 524 words, while there were 5 words in the shortest. We stopped the analysis when we reached thematic saturation after 493 comments [49, 57]. We coded 46 more comments to make sure no new codes were identified. Overall, we generated 139 codes and organized them into 10 themes.⁴ As we focus on

⁴See the list of all identified themes at <https://github.com/AUXResearcher/SOUPS102/blob/main/Themes.pdf>.

reporting the challenges users experienced and the concerns they had regarding the 3CN in this manuscript, we excluded the findings that were less relevant (e.g., users' strategies of creating credentials). We describe our reported five themes in Appendix B.

We recruited a diverse set of 22 interview participants from North America. The sample varied in age, occupation, and education level. Interview participants (referred to as "participants") were 20 to 74 years old (mean 40 and median 37), 11 of them identified as female (see the summary of participants' demographic information in Table 1). Interviews were conducted between August 2021 and January 2022. The interviews lasted an average of 26 minutes. Each participant was compensated with CAD 15. Data saturation was reached after 19 participants. We continued interviewing three participants and obtained no new codes [39]. We assigned 178 new codes in addition to those from the analysis of online comments and generated 11 new themes. In this manuscript, we reported 3 of the 11 new themes and related codes that are related to users' challenges with 3CN (see reported themes and codes in Appendix B). During the interview, some participants needed to review the UI to answer our questions. Upon their request, we showed them screenshots of the 3CN by the lead researcher sharing her screen.

In the rest of this section, we report the challenges and concerns identified regarding 3CN. We found that users' issues with 3CN are mainly associated with five major core aspects of the notification: the authenticity of the notification, data breach incidents, Google's knowledge of users' compromised credentials, multiple accounts being associated with one notification, and actions recommended by the notification. In the following, we explained how users' detailed challenges and concerns are associated with the identified aspects of the 3CN (see Table 2). The mapping between our findings and the identified themes is presented in Figure A.1.

4.2 Authenticity of the Notification

Believing the notification was a mistake. OC-users believed the notification was shown to them even though there were no security vulnerabilities. They therefore questioned the authenticity of the notification. To illustrate, OC-user128 commented: "*It [i]s wrong! ... I only get this on a website that only asks me for characters never the full password and chrome can [no]t store it.*" OC-user341 reported the same issue: "*I'm getting this from one[-]time password entries. ... [I] think you guys need to reconsider the implementation.*"

Misunderstanding that the cause of the notification was nothing related to compromised credentials. Some OC-users and participants believed the problems with their credentials were not about the credentials being compromised. Instead, they believe that the notification alerts them about having weak passwords in general. For instance, OC-user337 commented: "*The problem with this popup is weak passwords.*"

Core Aspects of 3CN	Users' Perceived Challenges and Concerns
Authenticity of the notification	Believing the notification was a mistake
	Misunderstanding that the cause of the notification was nothing related to compromised credentials
Data breach incidents	Lack of information about the "data breach incidents"
	False assumption that the breach occurred on the website on which the notification appeared
	Misunderstanding about Google being breached
	Security concerns about Google
Google's knowledge of users' compromised credentials	Lack of explanation of how Chrome finds users' compromised credentials
	False assumption that Chrome learns about users' plaintext credentials
	Misunderstanding about Google checking users' non-saved credentials
	Privacy concerns about Google's management of users' data
	Concerns about losing control over own data
Multiple accounts being associated with one 3CN	Lack of an explanation of why more than one account was found insecure with one 3CN
	Notification appears on many websites
Actions recommended by 3CN	Lack of information about the severity of the risks
	Lack of justification of the recommended action
	Lack of motivation to take the recommended action
	Challenges in managing new passwords
	Lack of instructions for discontinued accounts

Table 2: The core aspects of 3CN and the detailed challenges OC-users and participants experienced and their concerns about each aspect. Contents in the gray background indicate the identified concerns, and contents in the blank background indicate the discovered challenges.

The end. It has nothing to do with breaches." Another example is OC-user69, who stated: "If I had to guess [the reason for me getting the notification], Google is probably just pointing out that your password is too simple, and trying to light a fire under your ass to try to get you to change it." There are OC-users who believed the reason for them getting the notification was that the website where the notification appears had security problems: "[The issue] is [the] website not having their SSL certificates or the site itself has been detected for malware and phishing"[OC-user155].

4.3 Data Breach Incidents

Lack of information about the "data breach incidents."

OC users and participants were unable to find information about the data breach in which their credentials were leaked. The information was perceived as important for users to verify the incident's authenticity, understand the incident, and act on it. OC-users and participants wanted information such as when the breach occurred, where it happened, who was responsible for the incident, and what measures were taken by the responsible party as a response to the incident. To illustrate, OC-user88 stated: "I find it very frustrating that no additional info[rmation] is provided in regard to the data breach. I [would] like to know more about the breach, and how my info was compromised and what logic was used to determine [that] I need to update passwords. This feels a bit non-transparent on google's part." Another example is P4, who also wanted to know more about the data breach regarding where it happened: "I would like to know more if the data breach happened on any of the trusted websites. Because they are always the targets. Then, I will definitely change my password."

False assumption that the breach occurred on the website on which the notification appeared. Because the source of the breach was perceived as unclear, OC-users and par-

ticipants started making assumptions that the website where they received the notification was breached. Although it was possible that the website issuing the notice was also breached, this was not always the case. Assuming the source of the data breach was the website was a misinterpretation. For instance, P6 stated: "I assumed it is because that company's information [was] breached, like there was a data breach and maybe they were held at ransom for people's personal information and included their passwords." P16, who also had such a misinterpretation, wanted an explanation from the company who owns the website: "I want to know what the company did about [the breach incident]. When did they find out [that] they had a data breach? Why is Google telling me and why did not the company tell me [about the incident]?"

This misinterpretation led participants to trust the website less and/or stop visiting the company's website. When explaining her perception of the website after getting the notification, P16 stated: "I guess I trust them a little less. It makes me a little more careful about the data I put into different websites. Sometimes, I stopped going to the website altogether. Sometimes I unsubscribed from the newsletter."

As a response, OC-users tried to contact the website to verify the source of the breach. For instance, OC-user123 described her actions: "I contacted the websites that Google Chrome indicated had my passwords breached. They replied that my passwords and accounts had NOT been breached and warned me against this "third party" that was sending me misinformation perhaps to scam me."

On the other hand, the organization's IT support technicians reported clients had asked about the notification they received on the website. They believed that misleading information in 3CN had caused unfounded concerns among their clients and harmed their business. To illustrate, OC-user138 commented: "I have clients who are now deeply concerned about their security and they now somewhat distrust our work when they

see a message [about] ‘a data breach on-site’.”

Misunderstandings about Google being breached. By interpreting the notification, some OC-users misunderstood that Google was breached. For instance, some OC-user235 stated: “[G]oogle [has been] breach[ed] or it is an affiliate of theirs. ... Google has cookies everywhere for tracking and advertising purchases. [I]t is no wonder there are so many breaches when these companies require and share so much of our information while charging us to use many of their services.” This kind of misunderstanding has caused OC-users to stop using some features of Chrome: “If chrome is going to tell me every few weeks [that I] need to change passwords then [I] will turn off the save passwords and just type them in myself from now on. ... [C]hrome is said to be so safe and now [I] see all my password saved on chrome have been compromised!!”[OC-user144].

Security Concerns about Google. Because OC-users misunderstood that Google was breached, they expressed security concerns about Google. For instance, OC-user324 commented: “This is a very convoluted feature. Makes me think Chrome has bad security and gets hacked regularly. ... [It] seems like a reoccurring problem, and changing the password will do nothing.” Another OC-user blamed Google for not keeping users’ passwords safe and complained: “How could Google keep saying it is safe to store passwords in chrome while they just had a data breach? how could they have a data breach of our data and not even spend the effort to publi[sh] it and explain who, when, where and WHY and what are the strict mitigation actions they put in place?????”[OC-user47]

As a response to the perception that Google was breached, OC-users decided to stop saving credentials on Chrome or avoid using Chrome. To illustrate, OC-user531 remarked: “Is Google Chrome security THAT frigging weak?? I no longer want to save passwords to my Chrome account.” OC-user269, on the other hand, decided to change to another browser because “[on the other browser] I do not need to worry about ‘security breaches.’”

4.4 Google’s Knowledge of Users’ Compromised Credentials

Lack of explanation of how Chrome finds users’ compromised credentials. Users wanted more clarification about how Google knows their credentials were leaked. For instance, OC-user108 asked: “Does this mean that [G]oogle are sending my username/password (even hashed) to a third site without notification?” This perceived non-transparency reduced OC-users’ trust in Google: “Google is simple fear mongering, probably just to convert more users to Chrome. If [G]oogle truly cared or thought they were being helpful, they wouldn’t go through great lengths to hide the details of their operation” [OC-user41]. Participants believed that more knowledge of how Chrome learns about users’ compromised credentials could help them build trust in Google and moti-

vate them to take the proposed measures: “[The information] will increase my knowledge. And if I know [Google] is taking good care of our data, maybe in the future, I would be more comfortable sharing information with them” [P10].

False assumption that Chrome learns about users’ plaintext credentials. Poorly informed users formed a hypothesis that Chrome checks users’ plaintext credentials to facilitate the 3CN. For instance, OC-user427 stated: “Is Google decrypting [users’ credentials] to compare [them with] known list of compromised credentials? ... not certain I feel safe knowing that [G]oogle has a plain text version of my password to process even if it is for my better.”

Misunderstandings about Google checking users’ non-saved credentials. Some OC-users and participants believed Google checked their credentials even if they were not saved in Google accounts. For instance, OC-user521 stated: “If [G]oogle can find your password online; it means it is reading and processing your password before encrypting and storing. I think it is a terrible idea to save passwords on [G]oogle.”

Participants’ past experiences with similar security incidents on Google played a role in this misunderstanding. During the interviews, we carefully explored how users developed such misconceptions. Previous work suggested that past experiences with similar incidents may reduce users’ perception of the threat [95]. We, however, found that their past experiences contributed to participants’ misunderstanding of 3CN. For instance, P10 explained that she had received a “suspicious sign-in prevented” email from Google. Through the email, she learned someone was trying to log in to her account from an unauthorized device. Based on this previous incident, she concluded that: “Google keeps tracking of everything you are doing on your laptop or on your mobile. So, I think nothing is hidden from Google.”

Privacy concerns about Google’s management of users’ data. Believing Google tracks users’ non-saved credentials, OC-users and participants raised corresponding privacy concerns. To illustrate, OC-user244 stated: “Why is google tracking what I type for login credentials that I have not saved to Google? ... Getting the message about a breach might seem helpful, but considering how the warning came and what Google has to be doing to issue the warning, it is just really creepy.” Further, some participants wanted Google to be more transparent about how users’ data was treated, such as “who has access to [users’ data] and how easily accessible is it for someone else?”[P13] and “if users’ data are encrypted or if [users’ data are] in the cloud or on a server”[P16].

Participants adopted acceptance as the strategy to mitigate this privacy concern. To illustrate, P11 explained that taking a trade-off was the reason for not acting to stop Google from checking all his credentials: “If something is being offered for free as a service, then you are the product.”

Further, several OC-users believed that Google facilitated scams by sharing users’ data with other parties. To illustrate, OC-user486 commented: “But, isn’t it kind of fishy that

Google would know that my old useless account was compromised in a data breach, but yet, no way to know which it is. In other words, Google yet again supports malicious scams through their services and records data of the [s]ite, email, and passwords you are creating in real-time.” ... [G]oogle can now create a database of all email/pass[word] combos and the sites they are used on, for their users, to then “release” through planned data breaches to victimize more people.”

Concerns about losing control over own data. Several OC-users disliked Google checking their credentials without asking for consent first. For example, OC-user112 remarked: “Why on Earth does [Google] feel it [i]s appropriate to be doing this password/username background comparison without asking for explicit consent?” In addition, OC-user453 felt this default feature took away users’ ability to control their data: “Almost all other security features are toggleable. It i[s] not an unreasonable request for this feature to be optional.” OC-user496 believed users should be given more control over their own data: “My issue is that the user should have the ability to control Google’s desire to enhance the user’s security!!!”

4.5 Multiple Accounts Being Associated with One 3CN

Lack of an explanation of why more than one account was found insecure with one notification. After receiving one notification, users were surprised to see that there were many accounts shown to be insecure. When the user’s single username-password pair was leaked, all accounts that share the same credentials became insecure. Chrome’s browser-based credential check service examines all the accounts users saved in their Google accounts. Since people often reuse their passwords [23, 104], when users receive a notification of a compromised username-password pair, they most likely will find a long list of accounts with the same breached credentials. But there is no explanation about the link between the identified accounts, so users tend to be confused and panicked when learning that many of their accounts were listed as insecure. As a result, some OC-users questioned the authenticity of the breach and resisted changing the passwords: “1 day they are all fine and the next day 99 passwords are compromised. I still would like to know how. Because this is a lot of work to change all these passwords. ... No way someone hacked me on 90 sites”[OC-user379].

Notifications appear on many websites. Another challenge is that OC-users reported the 3CN pop-up on many websites. The notification appears when users sign up/log in to an account with the breached credentials [43]. Suppose users reuse their breached credentials across accounts; whenever they try to sign in to the accounts, they receive a notification. However, without such knowledge, OC-users were confused with many notifications showing up on many websites: “It pops up for EVERY webpage. I do [no]t want to live in password paranoia forever”[OC-user46].

4.6 Actions Recommended by 3CN

Lack of information about the severity of the risks. 3CN was perceived as not communicating the severity of the risks to users. Such information was perceived as a contributor for users to take mitigation strategies. Specifically, OC users and participants wanted to know if it would be a significant risk if they decided not to change the breached passwords: “I mean, how risky is it if I do not change my password?”[P13] In addition, P18 wanted to know if there could be other security problems by not changing the compromised password: “Is there a way to put some malware [in my device]? Will it be possible [that not changing the password] could compromise even the other sites?” Further, the risk level was perceived as helpful for users to decide if it is worth making a lot of effort to change the passwords: “Google is telling me [that] I have compromised passwords. How serious is this? ... I also really do not want to have to change my passwords if I do not need to. Because I have more than a hundred spread across many forums and sites”[OC-user520].

Lack of justification of recommended actions. Participants wanted more clarification about why changing the password is the best practice and what risks would be avoided by doing this. Such information could influence their risk management behaviors. For instance, P6 stated: “I would like to know if the best you can do is to just change [the password]. Or is it you just do the best you can and then, fingers crossed, hope for the best situation? ... I think it would be helpful to know what does [changing the password] actually mean for users.” Further, participants asked whether and how changing the password could mitigate the existing damage (i.e., breached credentials). P22 asked: “If there has already been a data breach, what is the point of changing the passwords? I would like to know if [the breached credentials] are completely out of your control at this point or [if] changing the password can help with that.” P22 was also unclear about why changing the password was suggested and nothing else: “... but they only tell you to change the password. That got me thinking maybe my username is Ok. But if not, why do not they ask us to change [the username] too?”

Lack of motivation to take the recommended action. OC-users and participants argued that the notification alerted users about something (i.e., account hijacking) that may not happen. Therefore, they tend to delay or not take action until harm has occurred [125]: “I read the message more and realized it was not saying my account had been compromised. It was just a warning, like there is a risk [that my account being compromised] may happen. So, I did not change my password”[P7]. Several OC-users shared the same opinion: “Randomly trying those compromised credentials in an account is like a 1 in a million shot, more actually, 1 in a billion probably”[OC-user39].

Further, even if an adversary found the accounts with compromised credentials, the damage is perceived to be limited

because users have additional authentication methods set up. To illustrate, OC-user41 explained: “[A]ny respectable website worth accessing (like a bank’s website) is going to employ [usually multiple] additional traditional authentication methods - be it pin numbers, one-time passwords, 2-factor authentication, image recognition, geolocation, device recognition, etc. You can not simply bypass these and gain access with a simple username and password.”

Unimportant accounts were not worth the effort. Some OC-users and participants suggested that they change the passwords for “important accounts.” Such accounts contained their personal information (e.g., pictures, medical record, social security number, and taxpayer ID number) or financial information (e.g., “PayPal account” [P14], “HSBC account” [P17], and “eBay account” [P18]). To illustrate, P11 explained his process of changing the passwords: “I just went through the list [of insecure accounts] to see where could I have my credit card [information] saved. So, if it is like Home Depot, I probably bought something from [it]. It probably has my credit card. ... But if it is like a news site. I would just leave it there.” OC-user180 also decided not to change the passwords for accounts they did not consider important: “They are not [the] websites I care about people getting my info. What are they going to do? Go on Carvana and buy a car for me?”

They further justified their action by indicating that their passwords for the important accounts were different from those for non-important accounts (e.g., “Fandom account” [P3]). Therefore, even if the unimportant accounts were breached, it would not harm them. However, research shows that 33% of the time, it was possible to use a common password list and the user’s password created in a “lower level” account to successfully guess their “higher-level” account passwords [52]. Therefore, if the passwords of non-important accounts are public, there is a risk that users’ important accounts could be hijacked.

Challenges in managing new passwords. Participants struggled with creating new passwords. Through the interview, we found that participants were uncertain about whether the new passwords were “good enough” to resist being breached again. None of the participants recalled receiving any suggestions on Chrome in creating new passwords [44]. Similar to previous findings [50], our participants used the same strategies to create new passwords, such as making a slight change to their current password (e.g., adding “!” at the end of their current password). For instance, P18 explained his strategy of creating new passwords as using “Same configurations. Not exactly the same. I just add different stuff. ... I am not sure if they are more secure. I hope so. ... I would like some kind of indicators saying that they are strong enough, like [the password] will not be breached again.” However, participants’ new passwords are most likely vulnerable to credential tweaking attacks, where the attacker tries different variations of the leaked password [23, 115].

Lack of instructions for discontinued accounts. OC-

users and participants wanted instructions about what to do when the accounts were not in use or when they no longer had access. For instance, P13 had some old accounts that she no longer used. She did not know the appropriate step regarding the breached credentials of such accounts: “A lot of these [accounts] are like 10 years ago, I do not actually use them anymore. I do not think I have access to them anymore. Now, you are saying [the passwords] need to be changed. ... I am not sure what to do. What if I just delete the accounts? Will that get me in trouble?”

5 Discussion

5.1 Novelty of Our Findings

We have contributed to the body knowledge in four ways.

First, to the best of our knowledge, ours is the first study to investigate users’ perceptions of browser-based compromised credential notifications. Specifically, compared to a notification of breached credentials for a certain account (e.g., a Facebook account [95]), we captured the unique challenges users experienced in managing multiple accounts through a browser-based password manager. For instance, we discovered that OC-users and participants found it confusing that they received a 3CN on many websites, and that one 3CN might indicate that many accounts were in danger (§4.5).

Second, we contributed new findings on users’ challenges in comprehending data breaches. Prior work regarding data breaches has focused on exploring people’s familiarity with the data breaches [4, 110], their perception of the risks caused by the data breaches [65], and their behaviors after the data breaches [65, 125]. Our work highlighted the perceived critical information that contributed to users’ comprehension of the data breach. We also discovered that the missing critical information played a part in users’ misinterpretation of the source of the breach. Furthermore, users’ misunderstanding of the data breach may result in them having unjustifiable concerns (§4.3). We therefore offer design recommendations aimed at improving the 3CN design to help users gain an accurate understanding of it (Recommendation 2 in §5.3).

Third, we not only corroborated previous findings indicating that few users act on the security warnings [40, 99], but also investigated their reasons for failing to take action and the challenges they experienced when they did act on a notification (§4.6). We provide suggestions for how notification instructions can be improved in several ways (§5.4).

Finally, we discovered the privacy and security concerns that OC-users and participants had regarding the notification (§4.3 and §4.4). Because of these concerns, they failed to mitigate the risk effectively. At the same time, the concerns resulted in some negative perceptions of Google. Recommendation 4 in §5.5 aims to address these concerns.

5.2 Layers of Information

Critical information about the credentials leaks was perceived as missing from the notification. Prior research on security warnings has offered many insights into the need to comprehensibly communicate various risks [10, 28], such as the consequences of not complying with a suggested action [10]. However, we discovered that the 3CN failed to communicate certain types of critical information to its users (§4.3 to §4.6). The missing information led OC-users and participants to be confused and make additional efforts to verify the authenticity of the risk and the need to take action to mitigate it.

Missing information is not easily accessible. For instance, an explanation of how Google learns about breaches in users' credentials is available [45], but this information is not linked to the process that users go through when responding to 3CN. In other words, users must search for such details proactively and may not find what they need.

Recommendation 1: Provide important information in a layered form. Prior work has suggested that the message in a security warning should specify the underlying risk clearly [10] but provide only the essential information to avoid overwhelming users [28, 53]. However, previous work in other fields (e.g., group decision making [123]) has also shown that having more information improves people's decision making. *Therefore, there is a trade-off between the amount of information that should be included to enable users to understand the notification and the perceived effort required to read and process it.* As the information identified as missing was perceived to be essential, we suggest that a notification should include all such missing details listed in Table 2.

A layered approach has been proposed and evaluated as a way to present information about privacy and security to users, such as a privacy notification for IoT devices [21, 30]. The results of previous studies suggest that a layered approach allows users to obtain prompt, detailed, and accurate information about the privacy protection of an IoT device [21].

A layered approach can potentially provide the following benefits: First, it would enable the 3CN to convey a large amount of relevant information to its users without overwhelming them. The initial layer of the notification contains the most essential information [93]. Subsequent layers would each provide additional important information (such as the information we identified as missing in Table 2). The design for each layer would observe the well-known principles of risk communication [14, 36], such as using as few technical words as possible [53]. The pathway from one layer to the next should be made clear and straightforward [35]. Second, with all the relevant information linked directly through the layered approach, users could find the answers to all their questions without seeking help elsewhere. Another potential advantage of the layered approach is that it can benefit different types of users (e.g., the tech-savvy and the novice). Each user could decide how much information they want when

learning about the notification.

However, the huge amount of information [53] may overwhelm (novice) users [28] and possibly push them away from responding to the notifications. Therefore, the usability and users' perceptions of such a layered approach will require further evaluation.

5.3 Correct and Adequate Understanding

We identified several challenges that users face when understanding 3CN. Knowledge enables both recognition and interpretation to occur [97]. Without knowledge, understanding is impossible [76, 79]. Therefore, we include our findings of the knowledge gaps in discussing users' perceptions of 3CN.

An example is the comprehension of the "data breach." Here, three types of challenges emerged: lack of information about the "data breach," false assumption of "data breach" due to being poorly informed, and having misunderstandings regarding the "data breach" (§4.3). Different approaches may be needed to solve each type of challenge. For the first type, more information can be provided to users to help them develop a better understanding of the notification (see our Recommendation 1 in §5.2).

The second challenge is that users' lack of knowledge results in misinterpretations. In other words, users were unclear about certain aspects of the 3CN. They started forming the wrong assumptions. Providing more information to users can potentially clear up some of these misinterpretations (Recommendation 1 in §5.2). However, when users have already formed their own hypotheses, a deeper explanation may be needed to correct a misinterpretation.

The third challenge is users developed misunderstandings of certain aspects of 3CN by interpreting the information they received (e.g., Google is breached §4.3). Getting additional information about the notification may not be enough to correct these users' misunderstandings. Once established, mental models (i.e., users' understanding of how something works) can be surprisingly hard to change, even when they are aware of contradictory evidence [113]. Instead of providing more information, explaining certain aspects of the notification may be necessary to dispel such misunderstandings.

Recommendation 2: Consider explaining certain aspects of the notification to dispel the misconceptions. Prior studies suggest that users may improve their understanding if a system makes its *reasoning transparent*, such as its purpose of accessing a particular type of users' information [67, 72]. Therefore, we suggest correcting users' misunderstandings by providing detailed explanations. For example, instead of saying that Google does not access users' plaintext passwords, 3CN can focus on clarifying how Google learns that users' credentials are leaked without accessing their passwords. This explanation should be direct and easy to understand without too many technical terms and jargon [84, 91]. Assessing the effectiveness of such an approach requires future evaluation.

5.4 Action Recommendations

Instructions that merely suggest changing passwords were not perceived as helpful. As explained in Section 4.6, OC-users and participants experienced many challenges regarding taking the recommended action. These challenges resulted in some OC-users and participants being unsure about whether to take action, and if so, what that action should be.

Recommendation 3: Provide more details in the instructions. To better help users mitigate the risk of 3CN, we suggest that more explanations should be provided in the instructions to justify the necessity of changing the passwords. For example, we recommend explaining why it is necessary to change the breached password, but not the username, what risks can or cannot be mitigated by this action, and what risks the user may face if they do not change their passwords.

Additionally, more instructions could be provided on how to create new passwords. The focus can be on why a slight modification of an old password might not be effective in mitigating the data breach risks [23, 115]. Also, users can be assisted in understanding the quality of their new passwords (e.g., through a password strength meter [25, 98]). Other instructions [33] for creating unique passwords, such as not reusing passwords across accounts [33], could also be helpful for users. More research is needed to evaluate whether more detailed explanations in the instructions are more beneficial in persuading users to act effectively and protect their accounts.

Due to the similarities in the design of instructions provided by other browser-based PMs (e.g., Firefox Password Manager) and standalone PMs (e.g., LastPass and 1Password) and the design of 3CN, we believe we believe Recommendation 3 can also bring insights into these PMs' future designs. To illustrate, both Firefox and 1Password ask their users to change their passwords without providing more details [17, 107], such as the severity of the risks of not changing the passwords. There is a chance that their users find this instruction unhelpful as well. We suggest that these PMs also consider providing more information in the instructions to help their users better manage their credentials.

5.5 More Control and Data Transparency

Some users' security and privacy concerns were specifically related to Google. They criticized the company for having too much control over users' data, not being transparent about managing their data, and facilitating scams (§4.3 and §4.4). These concerns resulted in some of the OC-users refusing to use Chrome password manager or abandoning Chrome entirely. These concerns may be addressed by clarifying how Google detects breached credentials (see our Recommendation 2 in §5.3). In addition, providing more transparency about how users' data is protected might also help mitigate concerns and build trust in the company [95, 123].

Recommendation 4: Replace the one-or-nothing model

by giving users more control over their data. Another step further would be to give users the ability to select and deselect accounts they want to receive notifications about breaches. Providing greater control to users might help address users' concerns and build their trust in the company [96, 112]. For instance, provided they are clear about the possible risks of certain behaviors (e.g., changing passwords for certain accounts, not changing passwords at all, or slightly changing passwords) (see Recommendation 3 in §5.4), users could be given a choice as to whether or not they wanted to be notified about breached credentials or not. Currently, users can either get notifications of all accounts with breached credentials or not get any notifications (by turning off the feature). This approach clearly does not work for all users. Our proposed approach could potentially motivate users to manage their credentials without being bombarded with notifications. However, the effectiveness of the proposed approach would need to be evaluated in future studies.

Similarly, we found that other PMs (e.g., Firefox Password Manager, LastPass, and 1Password) also check all users' saved credentials to alert them of compromised ones. Due to this similar all-or-nothing design, we suggest that these PMs also consider providing more control to users over deciding which accounts will receive a notification.

We want to clarify that we reviewed only the UI of other PMs and identified several aspects of the design similar to 3CN. As our users experienced challenges regarding these aspects, we believe our Recommendations 3 and 4 to improve the design of these aspects can also benefit other PMs. However, to what extent our recommendations will benefit the design of other PMs requires further research.

6 Conclusion

We report the challenges users experience and their concerns about the Chrome compromised credentials notification. Our findings suggest that developers consider improving the design of various aspects of the notification to support users in better protecting their online accounts.

Acknowledgments

This research has been supported by a gift from Scotiabank to UBC. We would like to thank members of the Laboratory for Education and Research in Secure Systems Engineering (LERSSE) who provided their feedback on the reported research. Our anonymous reviewers and shepherd provided important feedback and suggestions to improve the paper. Stylistic and copy editing by Eva van Emden helped to improve readability of this paper.

References

- [1] Netmarketshare: Market share statistics for internet technologies. <https://netmarketshare.com/browser-market-share.aspx>. Accessed: 2022-01-18.
- [2] What do you think of google chrome now warning you if your web passwords have been stolen? <https://www.quora.com/What-do-you-think-of-Google-Chrome-now-warning-you-if-your-web-passwords-have-been-stolen>, 2021. Accessed: 2022-05-24.
- [3] Compromised passwords warning - what does this mean? https://www.reddit.com/r/chrome/comments/i7kcb5/compromised_passwords_warning_what_does_this_mean/, 2022. Accessed: 2022-01-31.
- [4] Lillian Ablon, Paul Heaton, Diana Catherine Lavery, and Sasha Romanosky. *Consumer attitudes toward data breach notifications and loss of personal information*. Rand Corporation, 2016.
- [5] Mustafa Emre Acer, Emily Stark, Adrienne Porter Felt, Sascha Fahl, Radhika Bhargava, Bhanu Dev, Matt Braithwaite, Ryan Slevi, and Parisa Tabriz. Where the wild warnings are: Root causes of chrome https certificate errors. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1407–1420, 2017.
- [6] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A Large-Scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 257–272, Washington, D.C., August 2013. USENIX Association.
- [7] Hazim Almuhammedi, Adrienne Porter Felt, Robert W Reeder, and Sunny Consolvo. Your reputation precedes you: History, reputation, and the chrome malware warning. In *10th Symposium On Usable Privacy and Security (SOUPS) 2014*, pages 113–128, 2014.
- [8] Elaine Barnett-Page and James Thomas. Methods for the synthesis of qualitative research: a critical review. *BMC medical research methodology*, 9(1):1–11, 2009.
- [9] Brian Barr. Everyone loves credentials: Highlights from the verizon 2021 data breach investigations report. <https://spycloud.com/highlights-from-the-verizon-2021-data-breach-investigations-report/>. Accessed: 2021-05-13.
- [10] Lujó Bauer, Cristian Bravo-Lillo, Lorrie Cranor, and Elli Fragkaki. Warning design guidelines. *CMU-CyLab-13*, 2, 2013.
- [11] Sruti Bhagavatula, Lujó Bauer, and Apu Kapadia. (how) do people change their passwords after a breach? *arXiv preprint arXiv:2010.09853*, 2020.
- [12] Sruti Bhagavatula, Lujó Bauer, and Apu Kapadia. What breach? measuring online awareness of security incidents by studying real-world browsing behavior. In *European Symposium on Usable Security 2021*, pages 180–199, 2021.
- [13] Ann Bostrom, Cynthia J Atman, Baruch Fischhoff, and M Granger Morgan. Evaluating risk communications: completing and correcting mental models of hazardous processes, part ii. *Risk analysis*, 14(5):789–798, 1994.
- [14] Cristian Bravo-Lillo. *Improving computer security dialogs: an exploration of attention and habituation*. PhD thesis, Carnegie Mellon University, 2014.
- [15] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. Your attention please: Designing security-decision uis to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, pages 1–12, 2013.
- [16] José Carlos Brustoloni and Ricardo Villamarín-Salomón. Improving security decisions with polymorphic and audited dialogs. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 76–85, 2007.
- [17] Emily Chioconi. Received a data breach notification in Ipassword? take these 5 steps. <https://blog.1password.com/what-to-do-when-you-get-a-data-breach-notification/>. Accessed: 2022-05-27.
- [18] Catalin Cimpanu. 2021 databreach investigation report. <https://www.verizon.com/business/resources/reports/dbir/>, 2021. Accessed: 2022-01-18.
- [19] Carl J Clare. *Understanding the factors that influence the effectiveness of online customer reviews: a thematic analysis of receiver perspectives*. PhD thesis, Manchester Metropolitan University, 2012.
- [20] Stephanie Condon. Okta offers free multi-factor authentication with new product, one app. <https://www.zdnet.com/article/okta-offers-free-multi-factor-authentication-with-new-product-one-app/>. Accessed: 2018-05-23.
- [21] Lorrie Faith Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 10:273, 2012.

- [22] Ciaran Daly. Google Chrome warning as millions of users told to change their passwords, November, 02, 2021. <https://www.dailystar.co.uk/tech/google-chrome-hacker-warning-millions-25355589>.
- [23] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The tangled web of password reuse. In *NDSS*, volume 14, pages 23–26, 2014.
- [24] Sanchari Das, Jacob Abbott, Shakthidhar Gopavaram, Jim Blythe, and L Jean Camp. User-centered risk communication for safer browsing. In *International Conference on Financial Cryptography and Data Security*, pages 18–35. Springer, 2020.
- [25] Xavier de Carné de Carnavalet and Mohammad Mannan. From very weak to very strong: Analyzing password-strength meters. In *Network and Distributed System Security Symposium (NDSS 2014)*. Internet Society, 2014.
- [26] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*, pages 79–90, 2006.
- [27] Alison Doyle. What Is a Semi-Structured Interview?, June 27, 2020. <https://www.thebalancecareers.com/what-is-a-semi-structured-interview-2061632>.
- [28] Serge Egelman. *Trust me: Design patterns for constructing trustworthy trust indicators*. Carnegie Mellon University, 2009.
- [29] Serge Egelman and Stuart Schechter. The importance of being earnest [in security warnings]. In *International Conference on Financial Cryptography and Data Security*, pages 52–59. Springer, 2013.
- [30] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring how privacy and security factor into iot device purchase behavior. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.
- [31] Adrienne Porter Felt, Alex Ainslie, Robert W Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettis, Helen Harris, and Jeff Grimes. Improving ssl warnings: Comprehension and adherence. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2893–2902, 2015.
- [32] MM Fischhoff, G Baruch, A Bostrom, L Lave, and CJ Atman. Communicating risk to the public. *Environmental Science Technology*, 26(11), 1992.
- [33] Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. Pushing on string: The ‘don’t care’ region of password strength. *Communications of the ACM*, 59(11):66–74, 2016.
- [34] Tomas Flotyn. Chrome now warns you if your password has been stolen. <https://www.forbes.com/sites/thomasbrewster/2019/12/10/google-chrome-will-now-warn-you-if-your-web-passwords-have-been-stolen>, 2019. Accessed: 2022-01-26.
- [35] Center for Long-Term Cybersecurity. Designing risk communications: A roadmap for digital platforms. <https://cltc.berkeley.edu/2020/12/15/designing-risk-communications-a-roadmap-for-digital-platforms/>, 2020. Accessed: 2022-01-18.
- [36] Steven M Furnell, Adila Jusoh, and Dimitris Katsabas. The challenges of understanding and using security: A survey of end-users. *Computers & Security*, 25(1):27–35, 2006.
- [37] Shirley Gaw and Edward W Felten. Password management strategies for online accounts. In *Proceedings of the second symposium on Usable privacy and security*, pages 44–55, 2006.
- [38] Emma L Giles, Matthew Holmes, Elaine McColl, Falko F Sniehotta, and Jean M Adams. Acceptability of financial incentives for breastfeeding: thematic analysis of readers’ comments to uk online news reports. *BMC pregnancy and childbirth*, 15(1):1–15, 2015.
- [39] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364, 1968.
- [40] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. "what was that site doing with my facebook password?" designing password-reuse notifications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1549–1566, 2018.
- [41] Google. Google Help Center. <https://support.google.com/chrome/?#topic=9796470>.
- [42] Google. Manage warnings about unsafe sites. <https://support.google.com/chrome/answer/99020?hl=en&co=GENIE.Platform%3DAndroid>. Accessed: 2022-01-26.
- [43] Google. Password manager. <https://passwords.google.com/>. Accessed: 2022-01-22.

- [44] Google. Tired of memorizing p4ssw0rd\$? the new chrome has your back. <https://www.blog.google/products/chrome/chrome-password-manager/>, 2018. Accessed: 2022-02-10.
- [45] Google. Better password protections in chrome - how it works. <https://security.googleblog.com/2019/12/better-password-protections-in-chrome.html>, 2019. Accessed: 2022-02-08.
- [46] Google. Protect your accounts from data breaches with password checkup. <https://security.googleblog.com/2019/02/protect-your-accounts-from-data.html>, 2019. Accessed: 2019-02-06.
- [47] Google. How chrome protects your passwords. <https://support.google.com/chrome/answer/10311524#zippy=%2Chow-password-protection-works>, 2022. Accessed: 2022-01-20.
- [48] Joshua Gray, Virginia NL Franqueira, and Yijun Yu. Forensically-sound analysis of security risks of using local password managers. In *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, pages 114–121. IEEE, 2016.
- [49] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006.
- [50] Hana Habib, Pardis Emami Naeni, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. User behaviors and attitudes under password expiration policies. In *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*, pages 13–30, 2018.
- [51] Rybe Hager. Google is making it easier to check if your passwords have been compromised in a data breach. <https://www.theverge.com/2019/10/2/20892854/google-password-checkup-hack-detection-now-available>, 2019. Accessed: 2022-01-26.
- [52] SM Taiabul Haque, Matthew Wright, and Shannon Scielzo. A study of user password strategy for multiple accounts. In *Proceedings of the third ACM conference on Data and application security and privacy*, pages 173–176, 2013.
- [53] Marian Harbach, Sascha Fahl, Polina Yakovleva, and Matthew Smith. Sorry, i don’t get it: An analysis of warning message texts. In *International Conference on Financial Cryptography and Data Security*, pages 94–111. Springer, 2013.
- [54] Google Account Help. Getting a compromised password warning, but no passwords are showing as compromised. huh? <https://support.google.com/accounts/thread/89361664/getting-a-compromised-password-warning-but-no-passwords-are-showing-as-compromised-huh?hl=en>, 2022. Accessed: 2022-01-20.
- [55] Google Chrome Help. Compromised password warning. <https://support.google.com/chrome/thread/73069988/compromised-password-warning?hl=en>, 2020. Accessed: 2022-01-18.
- [56] Maria Henriquez. Apple’s new requirement puts additional focus on consumer and data privacy.
- [57] Nicolas Huaman, Sabrina Amft, Marten Oltrogge, Yasemin Acar, and Sascha Fahl. They would do better if they worked together: The case of interaction problems between password managers and websites. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1367–1381. IEEE, 2021.
- [58] Jun Ho Huh, Hyoungshick Kim, Swathi SVP Rayala, Rakesh B Bobba, and Konstantin Beznosov. I’m too busy to reset my linkedin password: On the effectiveness of password reset emails. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 387–391, 2017.
- [59] Troy Hunt. Have i been pwned? <https://haveibeenpwned.com/Passwords/>. Accessed: 2022-01-19.
- [60] Retail Insider. How much are online reviews actually worth? <https://retail-insider.com/retail-insider/2020/04/how-much-are-online-reviews-actually-worth/>, 2020. Accessed: 2022-01-26.
- [61] Insurance Information Institute. Facts + statistics: Identity theft and cybercrime. <https://www.iii.org/fact-statistic/facts-statistics-identity-theft-and-cybercrime>, 2022. Accessed: 2022-01-24.
- [62] Iulia Ion, Rob Reeder, and Sunny Consolvo. {“... No} one can hack my {Mind”}: Comparing expert and {Non-Expert} security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, 2015.
- [63] Jeffrey L Jenkins, Bonnie Brinton Anderson, Anthony Vance, C Brock Kirwan, and David Eargle. More harm than good? how messages that interrupt can make us vulnerable. *Information Systems Research*, 27(4):880–896, 2016.

- [64] Ambarish Karole, Nitesh Saxena, and Nicolas Christin. A comparative usability evaluation of traditional password managers. In *International Conference on Information Security and Cryptology*, pages 233–251. Springer, 2010.
- [65] Sowmya Karunakaran, Kurt Thomas, Elie Bursztein, and Oxana Comanescu. Data breaches: user comprehension, expectations, and concerns with handling exposed data. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS) 2018*, pages 217–234, 2018.
- [66] Soyun Kim and Michael S Wogalter. Habituation, dishabituation, and recovery effects in visual warnings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 1612–1616. SAGE Publications Sage CA: Los Angeles, CA, 2009.
- [67] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–10, 2012.
- [68] LastPass. Lastpass for chrome.
- [69] Rachel Leist. How to Do Keyword Research for SEO: A Beginner’s Guide, January 7, 2020. <https://blog.hubspot.com/marketing/how-to-do-keyword-research-ht>.
- [70] Xiaozhou Li, Zheyang Zhang, and Kostas Stefanidis. Mobile app evolution analysis based on user reviews. In *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pages 773–786. IOS Press, 2018.
- [71] Stacy Liberatore. More than 500,000 zoom user credentials have been stolen and sold on the dark web for less than a penny each. <https://www.dailymail.co.uk/sciencetech/article-8218723/More-500-000-Zoom-user-credentials-sold-dark-web-PENNY-each.html>, 2020. Accessed: 2020-04-14.
- [72] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: understanding users’ mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 501–510, 2012.
- [73] Debin Liu, Farzaneh Asgharpour, and L Jean Camp. Risk communication in security using mental models. *Usable Security*, 7:1–12, 2008.
- [74] Gang Liu, Shaoqing Fei, Zichun Yan, Chia-Huei Wu, and Sang-Bing Tsai. An empirical study on response to online customer reviews and e-commerce sales: from the mobile information system perspective. *Mobile Information Systems*, 2020, 2020.
- [75] Zhiwei Liu and Sangwon Park. What makes a useful online review? implication for travel product websites. *Tourism management*, 47:140–151, 2015.
- [76] Ephrat Livni. It’s better to understand something than to know it. <https://qz.com/1123896/its-better-to-understand-something-than-to-know-it/>, 2017. Accessed: 2022-01-18.
- [77] Ragnar E Löfstedt. Risk communication and management in the 21st century. *Available at SSRN 545724*, 2004.
- [78] Sanam Ghorbani Lyastani, Michael Schilling, Sascha Fahl, Michael Backes, and Sven Bugiel. Better managed than memorized? studying the impact of managers on password strength and reuse. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 203–220, 2018.
- [79] Emelda M. Difference between knowing and understanding. <http://www.differencebetween.net/language/difference-between-knowing-and-understanding/>, 2018. Accessed: 2022-01-18.
- [80] Fiona MacKellar. Subjectivity in Qualitative Research. <https://www.sfu.ca/educ867/hm/subjectivity.htm>.
- [81] Behnood Momenzadeh, Shakthidhar Gopavaram, Sanchari Das, and L Jean Camp. Bayesian evaluation of privacy-preserving risk communication for user android app preferences. *Information & Computer Security*, 2021.
- [82] Gemma Morgan. Semi-structured, narrative, and in-depth interviewing, focus groups, action research, participant observation, 2016. <https://www.healthknowledge.org.uk/public-health-textbook/research-methods/ld-qualitative-methods/section2-theoretical-methodological-issues-research>.
- [83] Shivali Best & Daniel Morrow. Android users can check to see if their password has been hacked by scammers. <https://www.dailyrecord.co.uk/news/science-technology/android-users-can-check-see-22551580>, 2020. Accessed: 2022-01-20.
- [84] Emily Newman. Avoiding use of jargon with customers. <https://corp.yonyx.com/customer-service/avoiding-use-of-jargon-with-customers/>.

- [85] Gilbert Notoatmodjo and Clark Thomborson. Passwords and perceptions. In *Proceedings of the Seventh Australasian Conference on Information Security-Volume 98*, pages 71–78. Citeseer, 2009.
- [86] Joy Okumoko. "chrome password breach warning: How to check and fix asap. <https://www.maketecheasier.com/fix-chrome-password-breach-warning/>, 2021. Accessed: 2022-01-20.
- [87] Joy Okumoko. Chrome Password Breach Warning: How to Check and Fix ASAP, August 24, 2021. <https://www.maketecheasier.com/fix-chrome-password-breach-warning/>.
- [88] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let's go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 295–310, 2017.
- [89] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why people (don't) use password managers effectively. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 319–338, 2019.
- [90] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.
- [91] Anastasia Philopoulos. How cutting out jargon can help you achieve clear communication. <https://www.shopify.ca/partners/blog/108716102-how-cutting-out-jargon-can-help-you-achieve-clear-communication>.
- [92] James H Price and Judy Murnan. Research limitations and the necessity of reporting them. *American Journal of Health Education*, 35(2):66, 2004.
- [93] TechRadar Pro. The case for a privacy nutrition label. <https://www.techradar.com/news/the-case-for-a-privacy-nutrition-label>, 2020. Accessed: 2022-01-18.
- [94] Jennifer Pullman, Kurt Thomas, and Elie Bursztein. Protect your accounts from data breaches with password checkup, 2019.
- [95] Elissa M Redmiles. "should i worry?" a cross-cultural examination of account security incident response. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 920–934. IEEE, 2019.
- [96] Harvard Business Review. With big data comes big responsibility, November, 2014. <https://hbr.org/2014/11/with-big-data-comes-big-responsibility>.
- [97] Irvin Rock. Perception and knowledge. *Acta Psychologica*, 59(1):3–22, 1985.
- [98] Tobias Seitz and Heinrich Hussmann. Pasdjo: quantifying password strength perceptions with an online game. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 117–125, 2017.
- [99] David Sharek, Cameron Swofford, and Michael Wogalter. Failure to recognize fake internet popup warning messages. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 557–560. SAGE Publications Sage CA: Los Angeles, CA, 2008.
- [100] Richard Shay, Iulia Ion, Robert W Reeder, and Sunny Consolvo. "my religious aunt asked why i was trying to sell her viagra" experiences with account hijacking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2657–2666, 2014.
- [101] Jeff Shiner. Finding compromised passwords with 1password. <https://blog.1password.com/finding-pwned-passwords-with-1password/>, 2022. Accessed: 2022-01-18.
- [102] James Simmons, Oumar Diallo, Sean Oesch, and Scott Ruoti. Systematization of password manager use cases and design paradigms. In *Annual Computer Security Applications Conference*, pages 528–540, 2021.
- [103] Radames Cruz Moreno Sreekanth Kannepalli, Kim Laine. Password monitor: Safeguarding passwords in microsoft edge. <https://www.microsoft.com/en-us/research/blog/password-monitor-safeguarding-passwords-in-microsoft-edge/>, 2022. Accessed: 2022-01-18.
- [104] Elizabeth Stobert and Robert Biddle. A password manager that doesn't remember passwords. In *Proceedings of the 2014 New Security Paradigms Workshop*, pages 39–52, 2014.
- [105] Elizabeth Stobert and Robert Biddle. Expert password management. In *International Conference on Passwords*, pages 3–20. Springer, 2015.
- [106] Joshua Sunshine, Serge Egelman, Hazim Almuhammedi, Neha Atri, and Lorrie Faith Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX security symposium*, pages 399–416. Montreal, Canada, 2009.

- [107] Mozilla Support. Firefox password manager - alerts for breached websites. <https://support.mozilla.org/en-US/kb/firefox-password-manager-alerts-breached-websites>. Accessed: 2022-01-21.
- [108] TeamPassword. What happened with the zoom credentials hack? <https://www.teampassword.com/blog/what-happened-with-the-zoom-credentials-hack>. Accessed: 2021-08-10.
- [109] James Thomas and Angela Harden. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC medical research methodology*, 8(1):1–10, 2008.
- [110] Kurt Thomas, Frank Li, Ali Zand, Jacob Barrett, Juri Ranieri, Luca Invernizzi, Yarik Markov, Oxana Comanescu, Vijay Eranti, Angelika Moscicki, et al. Data breaches, phishing, or malware? understanding the risks of stolen credentials. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 1421–1434, 2017.
- [111] Kurt Thomas, Jennifer Pullman, Kevin Yeo, Ananth Raghunathan, Patrick Gage Kelley, Luca Invernizzi, Borbala Benko, Tadek Pietraszek, Sarvar Patel, Dan Boneh, et al. Protecting accounts from credential stuffing with password breach alerting. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1556–1571, 2019.
- [112] Allison Schoop Timothy Morey, Theodore “Theo” Forbath. Customer data: Designing for transparency and trust, May, 2015. <https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust>.
- [113] Joe Tullio, Anind K Dey, Jason Chalecki, and James Fogarty. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 31–40, 2007.
- [114] Christopher Vendome, Diana Solano, Santiago Liñán, and Mario Linares-Vásquez. Can everyone use my app? an empirical study on accessibility in android apps. In *2019 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 41–52. IEEE, 2019.
- [115] Ding Wang, Zijian Zhang, Ping Wang, Jeff Yan, and Xinyi Huang. Targeted online password guessing: An underestimated threat. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1242–1254, 2016.
- [116] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. Understanding password choices: How frequently entered passwords are re-used across websites. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 175–188, Denver, CO, June 2016. USENIX Association.
- [117] Brian Wentz, Dung Pham, Erin Feaser, Dylan Smith, James Smith, and Allison Wilson. Documenting the accessibility of 100 us bank and finance websites. *Universal Access in the Information Society*, 18(4):871–880, 2019.
- [118] Wikipedia. Credential stuffing. https://en.wikipedia.org/wiki/Credential_stuffing. Accessed: 2022-01-22.
- [119] Wikipedia. Keychain (software). [https://en.wikipedia.org/wiki/Keychain_\(software\)](https://en.wikipedia.org/wiki/Keychain_(software)).
- [120] Graham Wilson, Harry Maxwell, and Mike Just. Everything’s cool: Extending security warnings with thermal feedback. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 2232–2239, 2017.
- [121] Michael S Wogalter. Purposes and scope of warnings. *Handbook of warnings*, 864, 2006.
- [122] Weining Yang, Aiping Xiong, Jing Chen, Robert W Proctor, and Ninghui Li. Use of phishing training to improve security warning compliance: evidence from a field experiment. In *Proceedings of the hot topics in science of security: symposium and bootcamp*, pages 52–61, 2017.
- [123] Paul Zarnoth and Janet A Snizek. The social influence of confidence in group decision making. *Journal of Experimental Social Psychology*, 33(4):345–366, 1997.
- [124] Zhijie Zhao, Jiaying Wang, Huadong Sun, Yang Liu, Zhipeng Fan, and Fuhua Xuan. What factors influence online product sales? online reviews, review system curation, online promotional marketing and seller guarantees analysis. *IEEE Access*, 8:3920–3931, 2019.
- [125] Yixin Zou, Abraham H Mhaidli, Austin McCall, and Florian Schaub. “i’ve got nothing to lose”: Consumers’ risk perceptions and protective actions after the equifax data breach. In *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*, pages 197–216, 2018.

Appendices

A How Are the Findings Associated with the Reported Themes?

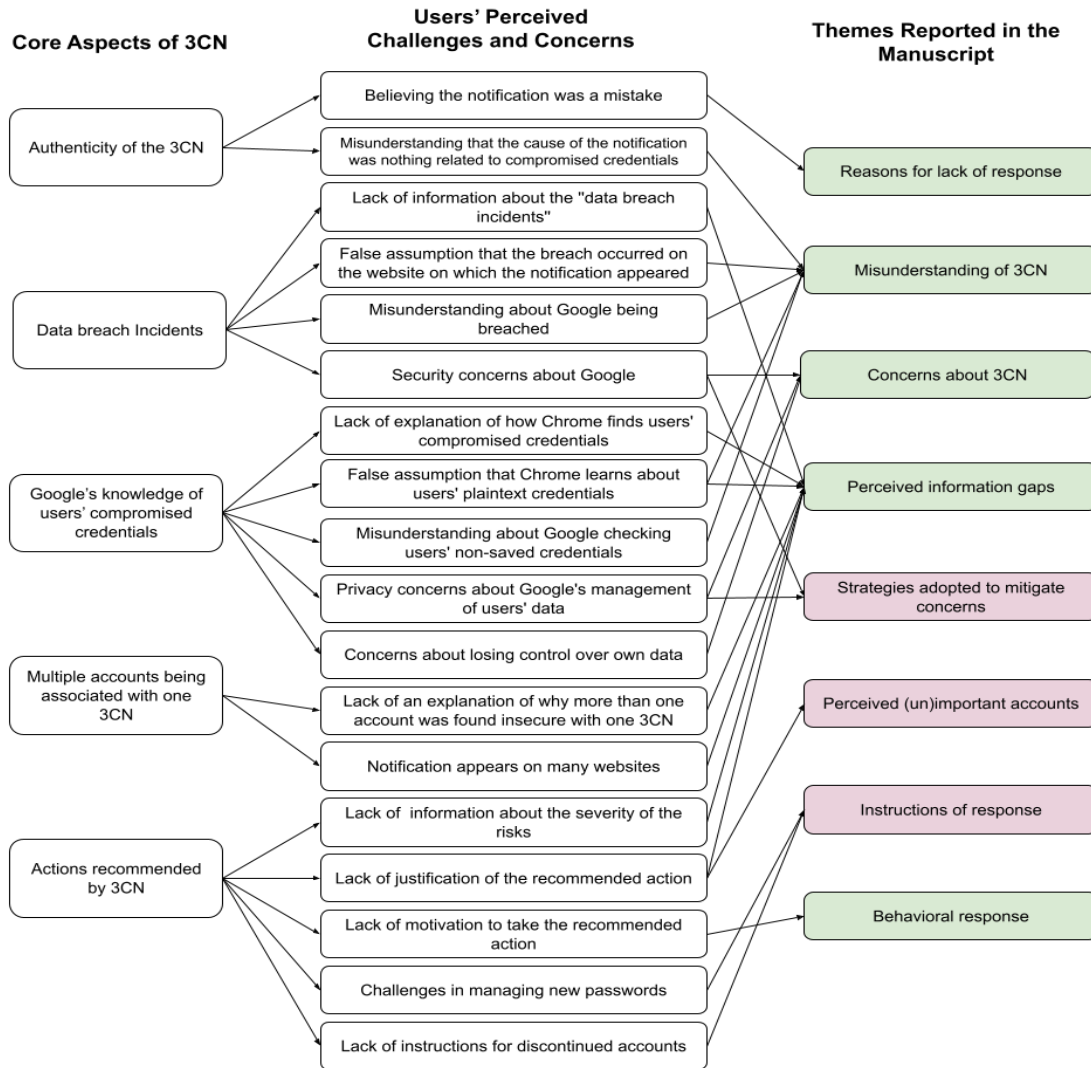


Figure A.1: A mapping between the core aspects of the app, the challenges and concerns OC-users and participants expressed, and the themes reported in this manuscript. The contents with a green background represent the themes that were identified by analyzing online comments, while the contents with a pink background indicate the themes were new themes identified through interviews.

B Themes and Related Codes Reported in the Manuscript

Themes	Codes	Number of online comments (N)	Number of participants (n)
Information Gaps	Risks of not changing passwords	4	2
	Severity of the risks	3	4
	Other security problems with the 3CN	3	1
	Whether changing the password is the best option?	2	3
	What other methods can take?	4	4
	When did the breach happen?	5	10
	Where did the breach happen?	25	8
	Who is responsible for the breach?	13	2
	Why did the company not notify users?	3	2
	What had been done as a response to the breach?	3	1
	Why there is no relevant news about the breach?	3	2
	Why do users receive so many 3CNs?	19	0
	Why does one 3CN represent so many insecure accounts?	14	0
	Why do users receive 3CN even after changing the passwords?	38	3
	How does Chrome know the about credentials being compromised?	30	15
	What information does Chrome check (credential or password)?	9	3
	What breached credentials does Chrome compare users' credentials with	8	6
	Does Chrome know users' plaintext passwords?	4	3
	Does Chrome check users' non-saved credentials?	7	5
	Why is changing the password suggested?	0	3
How effective is changing the passwords to mitigate risks?	0	3	
Why is changing usernames not suggested?	0	2	
Can changing passwords mitigate existing damage?	0	2	
Misunderstanding of the 3CN	The problem behind the 3CN is weak passwords	7	3
	Google's strategy to get people update passwords	6	2
	The website where the 3CN appears has security problems	9	2
	Google has been breached	16	0
	Chrome checks users' plaintext passwords	3	3
	Chrome checks non-saved credentials	3	3
Behavioral response	Click 3CN to learn more about it	6	4
	Disable the feature	4	2
	Lack of action	10	6
	Check other online sources to verify the data breach	8	3
	Email IT professional to learn more about the 3CN	5	2
	Ask friends/family about 3CN	4	4
	Search information about 3CN online	12	6
	Change all compromised passwords	3	2
	Change passwords for important accounts	8	6
	Delete stored credentials	13	3
	Ask help from Google live chat	1	0
	Change browser	7	0
	Contact the company where the 3CN appeared	6	0
	Run virus scan	2	0
	Intend to sue the company for not protecting data	1	0
	Changed some passwords then gave up due to too much effort	3	2
	Stop saving passwords on browser	3	4
	Decided to use other password managers	4	0
	Used Chrome suggested password as new passwords	2	1
	Stop visiting the websites where the 3CN appears	2	3
Examined each account and decide whether to change the passwords	0	4	
Close the notification	0	5	
Reasons for lack of response	3CN looks suspicious/not legitimate	20	2
	The message on 3CN is unclear/confusing	31	2
	Belief that no breach occurred	8	0
	Accounts are not important	8	6
	Perceived low chance of the account being taken	3	3
	Perceived low risk even if the account is hijacked	2	2
	Too much effort to change passwords for unimportant accounts	10	8
	Notification keeps appearing even after changing the passwords	38	3
	Unclear about how to deal with discontinued accounts	3	3
	3CN is alerting about something that has not happened	8	2
	3CN is exaggerating the risk	2	3
	Setting up additional protection methods	8	5
	Believing one should have the right to use any passwords they like	2	0
	Believing the passwords are complex enough	0	3
	Do not remember having such a compromised account	0	1
	The damage is already done	0	2
	Being too lazy to take action	0	2
Concerns	Google is breached and fails to protect users' data	23	0
	Google checks users' data without asking for permission first	8	2
	Google shares users' data with other parties	11	0
	Losing control over own data	12	0
Ways to steal users' new passwords	6	1	
Expected instructions of response	How to avoid being breached in the future	0	3
	How to create new passwords?	0	6
	Whether newly created passwords are secure enough	0	4
	Whether it is OK to use the same username	0	2
	How to deal with accounts that are no longer in use	0	2
	Whether certain accounts are riskier than others	0	1
Whether more methods are needed to increase account security level	0	3	
Perceived (un)important accounts	Accounts associated with financial information	0	9
	Accounts associated with personal information	0	9
Strategies to mitigate concerns	Accept the privacy-utility trade-off	0	3

Table B.1: Reported Themes and Codes. Themes and codes in pink are identified through interviews. We use “N” to indicate the number of online comments for each code and “n” to indicate the number of participants.

Exploring User-Suitable Metaphors for Differentially Private Data Analyses

Farzaneh Karegar
Karlstad University

Ala Sarah Alaqra
Karlstad University

Simone Fischer-Hübner
Karlstad University,
Chalmers University of Technology

Abstract

Despite recent enhancements in the deployment of differential privacy (DP), little has been done to address the human aspects of DP-enabled systems. Comprehending the complex concept of DP and the privacy protection it provides could be challenging for lay users who should make informed decisions when sharing their data. Using metaphors could be suitable to convey key protection functionalities of DP to them. Based on a three-phase framework, we extracted and generated metaphors for differentially private data analysis models (local and central). We analytically evaluated the metaphors based on experts' feedback and then empirically evaluated them in online interviews with 30 participants. Our results showed that the metaphorical explanations can successfully convey that perturbation protects privacy and that there is a privacy-accuracy trade-off. Nonetheless, conveying information at a high level leads to incorrect expectations that negatively affect users' understanding and limits the ability to apply the concept to different contexts. In this paper, we presented the plausible suitability of metaphors and discussed the challenges of using them to facilitate informed decisions on sharing data with DP-enabled systems.

1 Introduction

Differential privacy (DP) is a mathematically rigorous definition of privacy initially formalized in 2006 by Cynthia Dwork [20] for the calculation of statistics on a dataset. DP places a formal bound on the leakage of information from these statistics about individual data points within dataset.

Informally, for each person who submits their data to a differentially private data analysis, DP assures that the output of such analysis will be approximately the same, regardless of the contribution of their data to the data sample under analysis. Differentially private mechanisms perturb data in a controlled manner. This allows quantifying privacy through a privacy loss parameter ϵ , thereby fulfilling the assurance. Although leading to more privacy, lower privacy loss parameter values negatively affect the accuracy of the results. Consequently, there is a trade-off between privacy and accuracy in differentially private data analyses.

Within the past few years, several large companies, including Apple [42], Google [24], Microsoft [19], Uber [28] and LinkedIn [30], integrated differentially private mechanisms into their systems. The United States Census Bureau also adopted DP to prevent information disclosure in the summary statistics it released for the 2020 Decennial Census [5]. Further, different variants and extensions of DP have been proposed for other types of data analysis scenarios, such as local DP or DP for federated learning. Variants have local or central security models, and the choice of model has a considerable impact on the types of adversarial behaviour the system can tolerate.

Given the growing deployment of differentially private mechanisms in different variants and contexts, there is a need to address the human aspects of DP-enabled systems. In this work, we focus on conveying differentially private data analyses to data subjects who would share their data with systems deploying DP. The data subjects are mainly lay users without any expertise or knowledge about privacy or DP. However, they need to make informed privacy decisions about sharing their data when confronted with DP-enabled systems and services. Usable transparency of the functionality of the underlying differentially private mechanisms could help data subjects form correct mental models of how their data is protected, thus facilitating their decisions. Researchers have shown that how DP is described in practice is insufficient to help users make informed decisions [17]. Therefore, we need to explore how and to what extent the differentially private mechanisms

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

should be explained to users. Further, the issues surrounding their understanding and mental models of differentially private data analyses, their perceptions of the privacy provided, and their trust in such systems should be addressed. Therefore, in this work our main objective is to contribute to the body of knowledge on usable DP and to investigate how to effectively explain the underlying differentially private data analyses to lay users to facilitate their decisions.

In explaining a system to users, design models can employ familiar metaphors [14]. When the aim is to explain or represent a complex, abstract or novel domain (i.e. target domain), it is often helpful to use metaphors or analogies and make a vivid comparison to a familiar and often simpler domain (i.e. source domain) that people already understand. Some researchers argue that while in metaphor-based descriptions the mapping of ideas involves transformation, with analogies a direct transferal is made between existing knowledge and a novel domain [9]. However, in this paper, we do not make a distinction between these two concepts. Using an appropriate metaphor, people's understanding can be enhanced by encouraging them to use their pre-existing knowledge of the source domain to structure their thinking about the target of explanation [13]. Consequently, we assume that metaphors can be used to convey the concept of DP and its privacy functionality, as comprehending the complex concept of DP and the privacy protection it provides could be challenging for lay users. Nonetheless, suitable metaphors for complex concepts need to be generated with care and evaluated for their effectiveness in each context.

Previously, Demjaha et al. [18] benefited from the framework proposed by Alty et al. [9] to generate and evaluate explanatory metaphors for E2E encryption. We employed and adapted this framework to generate and evaluate metaphors for DP in the context of different data analyses. Our focus is on pictorial metaphors elaborated with short, simple text. Our approach consists of three phases: 1) metaphor generation, 2) metaphor analytical evaluations based on expert analysis and 3) metaphor empirical evaluations involving lay users. The first two phases resulted in adapted metaphors, their analytical evaluations and a functionality list that can be used to analytically evaluate the suitability of metaphors to convey the privacy functionality of differentially private data analysis to users. The details of the first two phases have been published in [29]. This paper focuses on the third phase but briefly describes the other two phases for clarity. In the third phase, we empirically evaluated the metaphors from phase 2 and addressed the following research questions.

RQ1. What information about the underlying differentially private systems is required by users to decide about using such systems (i.e. sharing their data)?

RQ2. What are users' perceptions about the data privacy provided by the proposed metaphors of DP?

RQ3. To what extent are our proposed metaphors suitable for conveying the concept of DP to lay users in the context of

different differentially private data analyses?

To address our questions, we conducted 30 online interviews. We defined three differentially private data analysis scenarios in the context of eHealth for local DP, typical central DP, and central DP for federated learning. Each interviewee was exposed to one of the scenarios and the related adapted metaphor(s) from phase 2. Interviewees responded to questions about their opinions and understanding of the metaphors.

We extended the previous findings on how DP should be explained to data subjects to facilitate their data sharing decisions. Our empirical evaluations provide information on the extent of the suitability of our proposed metaphors to explain DP to lay users and confirm the (plausible) suitability of metaphors while revealing specific challenges that must be addressed.

2 Background

Definition of DP. As defined by Dwork et al. [21], a randomized mechanism A is ϵ -differentially private, where $0 \leq \epsilon$, iff for any two data sets D and D' that differ in at most one record, and any set R of possible outputs of A , we have $Pr[A(D) \in R] \leq e^\epsilon * Pr[A(D') \in R]$. The definition prevents an attacker who knows all but one record in a database from inferring the last one after viewing the output. Simply put, DP mechanisms guarantee the stability of the output of a function based on changes that may happen in the input. Such a guarantee can facilitate releasing statistics on a database while preserving individuals' privacy in the database.

Different models. Differentially private mechanisms can be implemented as local or central (aggregate-level) models. In central models, a trusted data analyst (data curator/aggregator) gathers data from individual users and processes the data in a way that satisfies DP before publishing the aggregate statistics, similar to the original definition of DP [21]. In local models, users do not need to trust the entity responsible for analysis because their data get perturbed before being shared. The information disclosure risks differ substantially between these two models. However, in communicating with users, industry and media outlet DP descriptions do not clearly distinguish between central and local models, as reported by Cumming et al. [17]. Therefore, to address the limitation of existing descriptions, we defined three scenarios of differentially private data analyses in the context of eHealth.

Data analysis scenarios. The first scenario (SC1) is related to the local model of DP (Figure 3a in Appendix A) in which user data gets perturbed before being shared with the health company, which might not be trusted. For central models, in one of the scenarios (SC2) we have one data aggregator, a health company that conducts differentially private data analysis on actual information it collects from users and combines (Figure 3b in Appendix A). The other aggregate-level scenario (SC3) is related to differentially private federated learn-

ing where we have several data aggregators (different health companies) that collaboratively make an improved machine-learning model. They train a model collaboratively with the help of an Internet-based analyser (IBA) while preserving the privacy of their users (Figure 3c in Appendix A).

3 Related work: Usable differential privacy

Although technical literature on differentially private mechanisms and how to enhance them abound (e.g. [22,31,33]), just a small body of work focuses on the ethical, legal [15,16,36], and Human-Computer Interaction (HCI) implications of DP [12,17,47]. Among the considerable body of work on privacy communications (e.g. [6,34,39,41,48]), only a limited amount of research has focused on the communication of DP with different types of users. For instance, DPComp [26], PSI [25], Overlook [43], DPP [27], and ViP [35] provide interfaces for interacting with DP. However, the target groups of such tools are, for example, data curators/data analysts who may decide about the privacy budget based on their privacy and utility requirements. To the best of our knowledge, only three works have focused on explaining DP to end users (i.e. data subjects), which is more closely related to what we aimed for in this paper.

Bullek et al. [12] used a virtual spinner to describe the randomized response technique (RRT), a specific local DP technique, to users. They investigated whether users trust the RRT mechanism and if they adjust their privacy decisions when they see more details of the privacy promises implied by the RRT. Bullek et al. [12] reported that users vastly preferred the most anonymous spinner, although some participants preferred the most truthful spinner because they thought it minimized the ethical consequences of lying. We also use the spinner metaphor to describe DP in a local model. However, our spinner metaphor conveys the privacy-accuracy trade-off. Consequently, our results regarding users' preferences for which spinner to use differ from what Bullek et al. reported. In addition, our study reveals the shortcoming of the spinner metaphor for lay users and how it can be improved.

Xiong et al. [47] analysed the effects of using different short textual descriptions to inform users that their information is protected with DP on their willingness to share different types of information (sensitive and nonsensitive). They slightly modified and adapted descriptions from the companies/organizations that deployed and communicated DP to the public. Their results show that although users struggled to understand the DP descriptions, the descriptions explaining implications, that is, what happens if the aggregator's database is compromised, could facilitate people's data-sharing decisions and their comprehension of the local and central models.

Cummings et al. [17], in a series of online surveys, exposed their people to short verbal DP descriptions derived from publicly available descriptions of DP and investigated respondents' privacy expectations of DP-enabled systems and their

willingness to share data in such systems and showed that common privacy concerns can be addressed by DP. However, how DP is described in the real world haphazardly raises privacy expectations that may mislead users about the systems' privacy features. Results of studies in [17,47] show the need for better DP descriptions for users.

To the best of our knowledge, no attempts have yet been made to generate, test and compare metaphors for conveying the underlying differentially private data analysis (both local and central model) to lay users.

4 Method

Figure 1 shows an overview of our approach which is based on a framework proposed by Alty et al. [9]. The framework provides suitable tools and techniques for metaphor design for interactive systems. Demjaha et al. [18] previously adapted the framework and analytically and empirically evaluated the efficacy of their explanation metaphors for E2E encryption. Due to contextual differences, to reach our objective, we applied the adapted and extended version of the framework. Particularly, two rounds of analytical evaluations are included and the steps related to the integration of metaphors into the user interfaces of real systems are excluded. More details on phases 1 and 2 and the design of interviews as part of phase 3 are provided in Section 4.1 and Section 4.2, respectively.

4.1 Phase 1 and Phase 2

Phase 1: metaphor generation. We used both the *extension* and the *design metaphor* techniques proposed by Alty et al. [9] to generate metaphors in our work.

To begin with, we reviewed literature and media outlets to see how others conveyed the concept of DP to users using metaphors or analogies. Our literature review uncovered that, for the first time, Warner [45] proposed randomization of responses by a spinner to improve the reliability of them to sensitive questions. The spinner metaphor was later used by Bullek et al. [12] to convey DP to lay users. The spinner has also been used in media outlets to convey how DP works [2]. We searched the Web for *differential privacy* alone and in combination with the keywords users, people, definition and introduction. We examined each of the first five pages of the results to find explanations (in any format, including videos) describing the concept at a high level. Our search on media outlets showed that DP is explained to people using an example of tossing a coin for changing user responses [1], noisy sound waves of radio channels [4] and a noisy portrait [3]. Investigating how companies described DP to their users did not result in any other metaphors we could use in our study.

In phase 3, we monitored and analysed users' language when they talked about their perception, and opinions of DP and the metaphors to which they were exposed to see whether further metaphors could be derived.

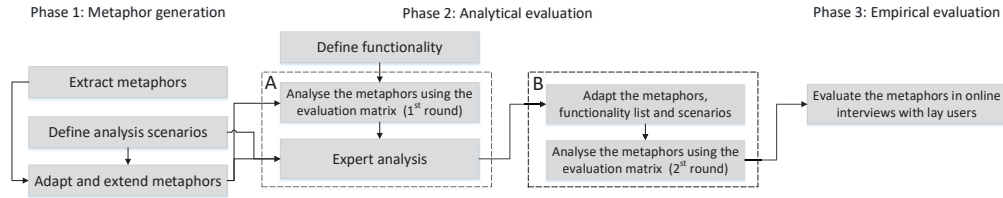


Figure 1: The steps followed to address the research questions.

Phase 2: analytical evaluation. Metaphor-system pairing is the step in Alty et al.’s framework [9] to analytically evaluate metaphors for which system functionality should be defined and then compared to features implied or conveyed by a metaphor. Therefore, we defined general privacy features of differentially private analyses and analysed metaphor-system pairings based on a metaphor evaluation matrix adapted from [18]. The matrix helps categorize the comparison of DP features with the features a metaphor (M) supports into desirable (DP+M+), undesirable (DP+M-), and very undesirable (i.e. conceptual baggage – DP-M+) groups of features. The template of the metaphor evaluation matrix we used is provided in [29].

Eight privacy experts knowledgeable about DP from academia and industry reviewed our materials in step A of phase 2 (see Figure 1), including the description of scenarios, the original functionality list, the resulting metaphors in phase 1 and our first round of analytical evaluation. The purpose of the expert review was to improve the validity and to check the authenticity of our materials. We reached the experts through personal contacts and ongoing collaborations within joint projects. Based on the reviews, we adapted the metaphors, functionality list and scenarios and re-analysed the metaphors (step B of phase 2). Section 5 presents our functionality list and briefly describes the resulting metaphors from phase 2, depicted in Figure 2, which we tested in our interviews.

4.2 Phase 3: Interviews

To evaluate the metaphors in Figure 2 and address our RQs, we conducted online interviews (via Zoom) with lay users. The interviews differed based on the data analysis scenarios (SC1 to SC3 described in Section 2) and the related metaphor(s) to which the interviewee was exposed. The interviews had three stages: 1) a prelude session, 2) a main session and 3) an epilogue session. In the prelude session, after a brief introduction to the study the interviewees were asked for their consent and were provided with a link to answer optional demographic questions (age group, gender, educational background). The main session had two parts: 1) scenario introduction and gauging expectations and opinions (before exposure to metaphors) and 2) metaphor introduction and gauging perceptions and opinions.

The first part of the main session started with an introduc-

tion of a persona followed by the data analysis scenario. A persona was used to avoid the disclosure of personal information. The interviewees were exposed to the related pictorial presentation of a data analysis scenario (Appendix A). Simultaneously, the interviewer read the scenario description. The interviewees were informed about the general privacy problem in the scenario and the existence of DP to mitigate the problem. However, the information on how DP would work was not revealed yet. The description of SC1 read to participants is reported as an example in Appendix B. After the scenario introduction, participants played the role of the persona and were asked to make a decision on sharing their data based on the scenario. They then answered the interview questions and provided input on reasons for their decisions, requirements for further information on DP, perceptions of the benefits and risks if they agreed to share data, expectations of privacy protection in the scenario and factors that would help them their trust in DP to protect their privacy.

In the second part, each interviewee was exposed to the related pictorial metaphor(s) for the scenario (Figure 2). At the same time, the interviewer read the simplified description of DP as the accompanying information defining the metaphor. The exact accompanying information for each metaphor is provided in Appendix C. Afterwards, participants were requested to review the decision they previously made regarding sharing their data and to elaborate on their decisions to see how the DP description could have affected their decisions. They then provided their opinions on the understandability of the metaphor and how it could be improved. Questions about users’ perceptions of distortion/perturbation and privacy provided by DP were asked. This was followed by questions to check whether the metaphor could convey the features in the functionality lists, including the privacy-accuracy trade-off and users’ perceptions and preferences. Participants were prompted to elaborate on their opinions about the remaining privacy risks, whether they would trust DP to protect their data, to describe DP in their own words, and to suggest alternative descriptions of DP. The main session ended here for the SC2 and SC3 interviews. However, for SC1, half of the participants were first exposed to the spinner metaphor and then to the noisy picture metaphor, while the other half were exposed in the opposite order. After being exposed to the second metaphor, participants answered questions about their perceptions of the second metaphor. They also answered

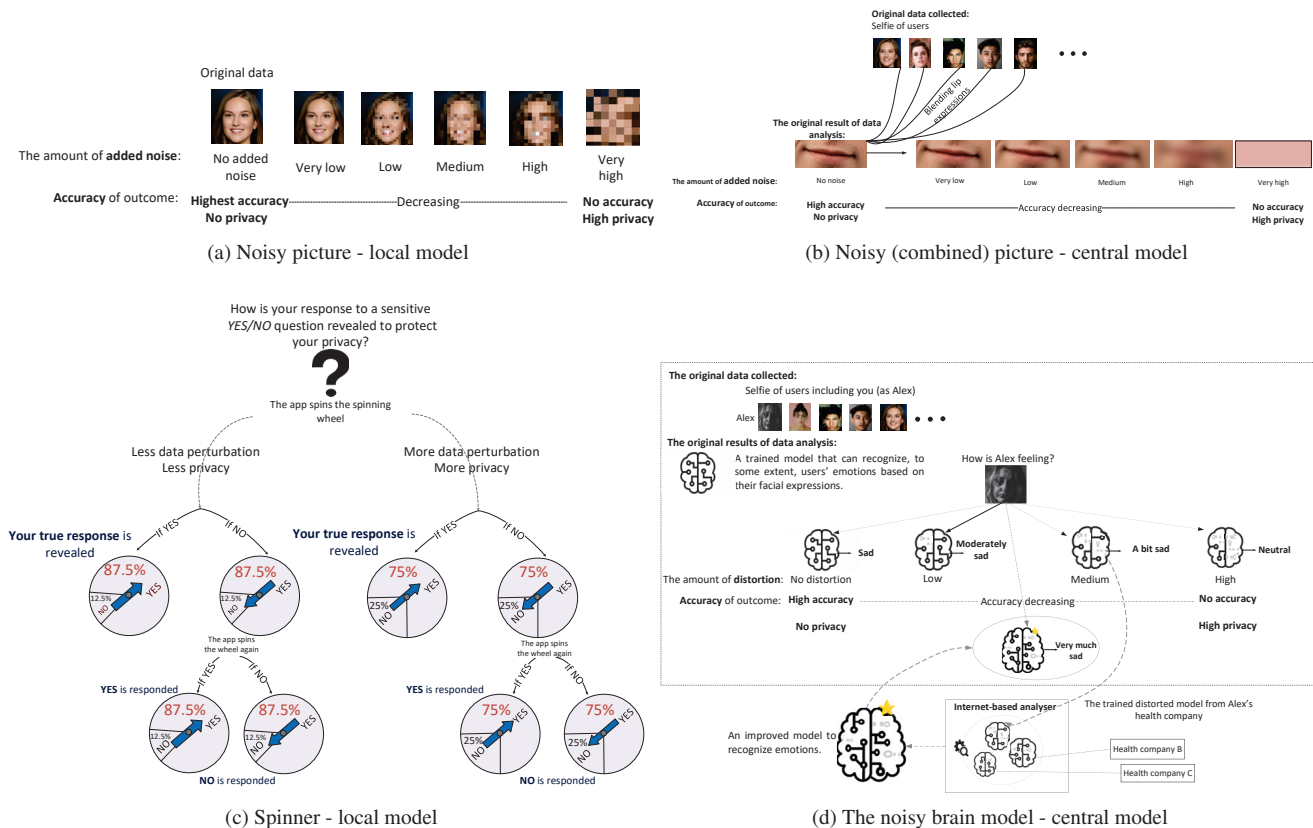


Figure 2: The metaphors to which participants were exposed in our interviews.

questions about which of the metaphors was easier to understand, conveyed privacy-accuracy trade-off in a better way, associated perturbation/distortion with privacy protection in a better way and which they preferred to be exposed to for DP description.

The interview was concluded in the epilogue session in which we asked participants to reflect on any issues we did not discuss in the interview. All interviewees who completed the interview were compensated with 20 GBP. The interview guide is provided in Appendix D.

Sampling and recruitment. We recruited 30 interviewees (10 for each scenario) from the Prolific platform. We used Prolific prescreening filters to recruit people whose current countries of residency were EU countries, EEA countries, the UK and Switzerland due to the scope of our funders. To sample lay participants without knowledge of DP, we excluded those with an educational background related to engineering, computing (IT) and computer science. Furthermore, at the beginning of the interview, we asked a few questions to gauge participants' initial familiarity with privacy-enhancing tools (PETs) and DP. We also conducted three pilot interviews (one for each scenario), which did not result in any major changes. Our participants' demographics are reported in Section 6.

Data analysis. We analysed our empirical data using thematic

analysis [11]. Our data included approximately 36.6 hours of audio recordings from the interviews (SC1=13.6 h, SC2=13 h, SC3=13 h). A research assistant manually transcribed the audio recordings. All authors read and familiarized themselves with the content of the transcripts. Using NVIVO (software for qualitative data analysis), one author analysed the transcripts; this resulted in 1257 excerpts for SC1, 1255 for SC2 and 1142 for SC3. The other authors then reviewed the excerpts. All authors met afterwards for a workshop to discuss the codes and the code book and to agree on terminologies and resolve conflicts (disagreement percentage was 2%). We then finalized the code books for each scenario that were used as a basis for the following rounds of the analysis. Because we went through several iterations to discuss the codes and resolve conflicts, our disagreement percentage was calculated based on the final round of codes. Next, each author independently led a preliminary categorization (thematic analysis) of one scenario, and then reviewed the other two scenarios. All authors met afterwards in a second workshop to discuss the main themes of all scenarios. Following agreement about the main themes, the second round of categorization of codes into the agreed-upon themes and further sub-themes was conducted. A third workshop was conducted to finalize the analysis and results.

Ethical considerations. The study was approved by the ethical advisor at Karlstad University. Interview data, including Zoom session recordings, were collected on the legal basis of informed consent given by the participants, and all data were processed in compliance with the General Data Protection Regulation (GDPR). Participants were instructed to use a non-identifying pseudonym as their Zoom name and to turn off their cameras during the recording to prevent the leakage of any identifying information. During the interviews, we introduced the scenarios in terms of a fictional user (persona) called Alex and asked the interviewees to answer questions from the perspective of Alex or in general and thus NOT to reveal any sensitive personal data, such as data related to their personal health conditions.

5 Functionality list and adapted metaphors

The following is the adapted functionality list after receiving feedback from experts:

- (F1)** A differentially private analysis, that is, a mechanism, bounds and quantifies the probability of additional privacy risk any individual would face because of their participation in a data analysis.
- (F2)** The privacy of a differentially private analysis is controlled by tuning a privacy loss parameter.
- (F3)** The smaller the value of the privacy loss parameter, the better the privacy guarantee for an individual.
- (F4)** The smaller the value of the privacy loss parameter, the less accurate the results of data analysis.
- (F5)** A differentially private analysis randomly perturbs data on an aggregate level (i.e. the results of the analysis) or individual level (i.e. the input data), depending on the context.
- (F6)** The amount of perturbation is controlled by the underlying differentially private analysis.
- (F7)** A differentially private analysis is resistant to privacy attacks based on auxiliary information, i.e. any past, present and future information an attacker may have.
- (F8)** A differentially private analysis does not promise unconditional freedom from privacy risks.

F1 can be interpreted in different ways. For the central model, it should convey that the results of a differentially private data analysis do not significantly depend on any particular individual's data. F1 can also be rephrased in terms of plausible deniability for a particular data record in the local model and participation in data analysis in the central model. Although a metaphor may not directly convey F1, it may imply one of its interpretations.

Considering our target group, we did not focus on the privacy loss parameter but on the role of perturbation in providing privacy and the effects of perturbation on the accuracy of the results. Therefore, if a metaphor conveys that more perturbation leads to better privacy but less accuracy we assume it covers F3 and F4. Further, we avoided including additional details to keep the features simple.

Table 1: Features of functionality list covered or not (Y/N) by each metaphor.

Metaphor/feature	F1	F3	F4	F5	F6	F7	F8	Context
Spinner	Y	Y	Y	Y	Y	Y	Y	Scenario 1
Noisy single picture	N	Y	Y	Y	Y	N	Y	Scenario 1
Noisy picture - combined	Y	Y	Y	Y	Y	Y	Y	Scenario 2
Distorted brain model	Y	Y	Y	Y	Y	Y	Y	Scenario 3

The development stages of our metaphors are defined in [29] in detail. In sum, in phase 1 we adapted and extended our initial metaphors (described in Section 4.1) because they were not necessarily suitable for all scenarios and all models of DP. In phase 2, based on the results of the expert analysis and our analytical evaluation, we excluded the metaphor of noisy sound waves of a radio channel due to features categorized as conceptual baggage and adapted our preliminary spinner metaphor to better communicate F3, F4 and F6.

The metaphors shown in Figure 2 were all defined for an eHealth application where users' stress levels are determined by analysing their face or lip expressions in pictures (selfies) contributed by the users. For SC1, we chose the metaphor of a *noisy picture* showing different levels of added noise with different degrees of pixelation (Figure 2a) and an adapted spinner metaphor showing two spinners with different biased outcomes mediating different levels of perturbation (Figure 2c). For SC2, a noisy combined picture metaphor was used to convey that noise is added to the aggregated data (i.e. the combination of pictures with lips is pixelated, Figure 2b) and not directly to individual records (users' selfies). For SC3, we used a *distorted brain*, for which some of the neural connections are greyed out, as a metaphor of a differentially private trained model (Figure 2d).

Table 1 shows whether each of our adapted metaphors conveys or implies the features in the functionality list, although it is subject to the validation of users. Features F3 to F6 and F8 are conveyed by all four metaphors. Until completely distorted, we can still have a useful analysis that may carry a risk of revealing information about individuals. F1 is implied by the spinner metaphor. However, the noisy picture metaphor for the local model (Figure 2a) does not cover F1 and F7. The noisy combined picture metaphor may convey F1 and F7, depending on the combination of all pictures picked for that metaphor. In addition, users' understanding of, for example, how much the aggregate-level picture might be revealing and if and how the added noise can circumvent privacy leakage from a combined picture may play a significant role. The distorted brain metaphor (Figure 2d) is quite abstract, and whether it conveys F1 and F7 is greatly dependent on what users know or understand from the concept of a model.

6 Interview results: Phase 3

Among the 30 participants (P1–P30), 13 identified themselves as female (SC3=2, SC2=7, SC1=4), 16 as male (SC3=8,

SC2=3, SC1=5), and one did not answer demographic questions. Our interviewees were relatively young; 18 were aged 18–25, 8 were aged 26–35 and 3 were aged 36–45. They had diverse academic backgrounds, including medicine, chemistry, psychology, cooking, international business and architectural design. Most were pursuing higher education studies. However, four of the interviewees indicated they were high school graduates or students. While the participants were not generally knowledgeable about PETs, some were aware of technologies or tools that help protect privacy. Encrypted messaging in specific communication apps, cookie consent forms, basic pseudonymization with reference to what Prolific does to hide users' real identity (e.g. using codes instead of emails/full names) and virtual private networks (VPNs) were mentioned in all three groups. Further, the participants were not knowledgeable about DP and had not heard about it before, meaning they were non-experts in privacy.

In total, our analysis resulted in 12 main themes. The main themes are indicated by a (T) and a unique number (X). Sub-themes follow the format TX.X. When reporting a scenario-specific sub-theme, the scenario number follows the theme number in this format: (TX.X-SCX). If a sub-theme was common between all scenarios we omitted the scenario number. An overview of all themes is provided in Appendix E. We skip the theme number in the number format of a sub-theme when we report a sub-theme in this section for readability purposes. For example, instead of reporting (T1.1) we simply report (1). For SC1, the order of being exposed to the two metaphors (spinner/noisy picture first) had no significant difference in the results. The first four themes (T1–T4) are *pre-explanation* themes and the rest (T5–T12) are *post-explanation* themes. The explanation refers to the introduction of the DP metaphors relating to each scenario (see Section 4.2).

Information needed for trust and data sharing: Themes T1–T4 and T6–T8 address RQ1, as they shed light on the information affecting users' trust in and decisions to share their data with a DP-enabled system. The results show that the mere presence of a privacy technique is seemingly enough to persuade users to share their data. However, lack of transparency about DP leads to varied expectations and interpretations of who gets access to actual (raw) data, different assumptions (correct/incorrect) about DP and negative impacts on willingness to share data with and trust in a DP-enabled system. Most participants required usable transparency of DP, for example, to know how DP works, protects, and uses personal data and to know about the risks of identification.

T1: Factors affecting sharing of data. In all three scenarios, participants mentioned positive (1) and negative (2) factors affecting their decision to share their data with DP-enabled systems. Positive factors are the existence of a protection technique (1.1), transparency of DP (1.2), providing reassurance regarding data safety and reliability (1.3), the specific type of data and data processing purposes (1.4), good reputation/location of the company (1.5-SC2,3), the existence of DP

as a trust factor (1.6-SC2,3), contribution to the improvement of the health app (1.7-SC1,2) and being anonymous (1.8-SC1). The claimed existence of a privacy technique was important and enough for several participants to decide to share their data. In SC1, where the company does not have to be trusted, anonymity was mentioned more often than in SC2 and SC3, where the reputation of the company mattered for trust. P36 mentioned the following reason for deciding to share data: *"Because the site has a good reputation so I- I think my data is safe"*. Participants had concerns about different kinds of privacy risks that negatively affected their sharing decisions, including the involvement of third parties and data/purpose misuse risks (2.1), identification risks (2.2-SC1,2) and data leakage/security risks (2.3-SC1,2). In addition, incorrect assumptions about DP (2.4-SC1,2), such as being reversible, negatively affected the decision to share. Before being exposed to how DP works, participants had the opportunity to make assumptions about its functionality (see also T3). Although the existence of a protection technique motivates people to share their data, the lack of transparency regarding DP (2.5-SC1,2) negatively affects their decisions to share data. Other hindering factors included not trusting the company (2.6-SC1), not trusting DP to protect privacy (2.7-SC1,2) and a general lack of trust (2.8-SC3) due to the belief of the persistent possibility of data leakage.

T2: Expressed needs for more privacy information. Across all three scenarios, most participants expressed a need for more information related to privacy protection (1) and more specifically related to DP (2) that should be provided in an understandable way (usable transparency) (3). P3 indicated that concrete examples should be given to illustrate the protection and risks of using DP: *"I might want to know what exactly they would protect, like what goes under the protection model and what doesn't [...] the data that they do protect is sleep cycles, but they don't protect the um... information about maybe the steps I'm taking"*. The main needs were for information about the provided privacy functionality (1.1), further specific privacy protection information (1.2), data storage information (1.3), whether sensitive data is processed and with whom it is shared (1.4), information about anonymity/re-identifiability when sharing data (1.5) and information about protection against breaches and risks (1.6). Furthermore, the need for more DP-related information (2) was also expressed, including information on how DP works (2.1), how DP uses and protects data (2.2), the accuracy of personal data that the company receives (2.3-SC1) and information on how trustworthy DP is (2.4-SC3).

T3: Expectation of claimed protection (data access). Our results show that the mere claim that DP protects data without further information on how it works can lead to different assumptions about DP (1) and its privacy features. It can also lead to varied expectations and interpretations regarding access to actual (raw) data (2) by different entities involved in data analysis. Such assumptions and expectations may prevent

users from sharing their data if they incorrectly assume that a specific entity (e.g. the health company in SC1) gets access to their data as they disclose them. DP has been associated with anonymization/pseudonymization (1.1) or with encryption (1.2-SC1,2). Several participants still perceive the risks of identification or data leakage/security risks (1.3-SC1,2) even with DP in place, and/or (incorrectly) think that DP is reversible (1.4-SC2) or assume that analysing data requires access to actual raw data (1.5-SC2,3) or simply associate DP with lower accuracy of data (1.6-SC1).

Assumptions about DP (1) played a significant role in participants' perceptions and expectations of the claimed protection and access to data by different entities. In all scenarios, participants who associated DP with pseudonymization expected that the raw data would be shared with different entities, depending on the context. For example, P2 stated that when sharing with the health company: *"I just assumed that some more personal information would be anonymous, and the rest would be like the raw data"*. In SC2 and SC3, participants assumed that medical researchers and the IBA needed access to raw data to analyse data, which is a false assumption. Likewise, in SC1, doubts about where the protection technique comes into force and the fact that the app is provided by and belongs to the company contributed to users' confusion and wrong incorrect assumption about access to raw data.

T4: Expressed trust factors of DP protecting data. In all scenarios, transparency of DP (1), transparency of data processing types and purposes (2) and good reputation of the company and its history of securing data (3) appeared as factors affecting users' trust in DP to protect their privacy. Trust factors also include being legally (GDPR) compliant (4-SC1,2), having unlinkability features (5-SC1,2), the existence of different privacy assurances and guarantees (6-SC2,3), trusting the company (7-SC1,3), having accountability measures in place (8-SC2) and being a standardized technique (9-SC3). Interestingly, although for the local model (SC1) the health company does not have to be trusted, the trustworthiness of the company appeared as a trust factor. P9 elaborates: *"if I see that the company itself has been trustworthy for several years and has not had major controversy with previous products"*. In addition, incorrect assumptions about DP impacted user trust (10-SC1,2). This included the assumption of DP being reversible, which negatively impacted trust, and associating it with encryption, which positively impacted trust.

T6: Varied impact of DP descriptions on decisions to share. The exposure to metaphoric descriptions of how DP works had a varied impact on the participants' willingness to share their data. The metaphoric DP descriptions either supported/increased the willingness to share (1) or decreased the willingness to share (2). Some participants indicated that privacy concerns are not critical for the decision to share. For example, P7 stated: *"considering I agreed earlier on my data to be shared, I don't think that would be that much of a problem but this would be at the back of my mind"*. There

were four participants in SC1, six in SC2, and five in SC3 who decided to share their data and persisted in sharing after exposure to the related metaphor. A few participants decided to share, contrary to their previous decisions, or became more inclined to do so (three in SC1, one in SC2 and two in SC3). Trust in having privacy protection and safety due to DP (1.1), the existence of distortion for privacy protection (1.2), transparency of DP (1.3-SC1,3), trusting the company receiving the data (1.4-SC2), the type of data requested (not perceived as sensitive) (1.5-SC1) and perceived common good benefits of sharing (1.6-SC2) were the factors that supported/increased willingness to share data. Interestingly, misconceptions about DP can also have a positive impact on data sharing (1.7-SC2). The perception of aggregation being secure enough for privacy protection increased the willingness to share in SC2. For example, P11 stated that: *"the first image with no noise is a mixture of the selfies [...] there is some sort of privacy cause it's not my actual picture."* Trading accuracy for privacy (2.1) and the type of data requested (2.2) were the factors in all scenarios that negatively impacted the willingness to share. Participants were mostly not happy to share the type of data they considered very personal. Many voiced the need for more information (2.3-SC1,2) or concerns about the risk of identification (2.4-SC1,2), which were other factors that decreased their willingness to share. Further, misconceptions about DP once again negatively impacted users' perception of its privacy protection (2.5-SC1). For example, after being exposed to the metaphor, P7 stated: *"I cannot guarantee about the privacy which I'm letting it go [...] I mean if I had some noise it's already blurred, but there are many ways which we can, you know, remove the noises."*

T7: Perceptions of information provided/missing. Most participants (eight in SC1, eight in SC2 and five in SC3) perceived the metaphors as easy to understand. In SC3, participants expressed confusion about the model and distortion. They desired more information about what distortion is, how it happens and what its role is in privacy protection and more concrete examples. In all scenarios, most participants expressed interest in receiving more detailed information on how DP works but in simple and clear language. For all three metaphors, people thought there was a lack of information on how distorted/perturbed data can be useful for the analysis and wanted to know if they would have control over the level of distortion. For the noisy picture metaphor, they specifically wanted to know if the process was reversible and thought the levels of accuracy/privacy shown needed elaboration. Participants also suggested some improvements for the spinner. Some indicated that the "YES/NO" on the first spinner was confusing and suggested replacing it. P4 stated: *"YES/NO you're not sure what they're talking about...that can maybe be mistaken as yes or no question"*. In SC1, most participants believed the noisy picture was easier to understand compared to the spinner metaphor; it was appreciated because of its brevity, clarity, simplicity and graphical visualization.

T8: Expressed trust factors (post-explanation). Most participants stated they would generally trust DP to protect their privacy. Transparency of DP (1), type of data/purposes of processing (2), accuracy (accurate results) (3) and understanding of protection provided by DP (4) were the common trust factors in all scenarios. Having control of the distortion level (5-SC1,2), a balanced trade-off (6-SC2) and aggregated data (7-SC3) were also factors indicated to enhance users' trust. Misconceptions about DP were reported to negatively impact users' trust (8-SC2). Many shared the misconception of DP being reversible, which led to distrusting DP. P16 stated: "I don't trust this because it's very easy to reverse it...it can be made by humans so we can reverse it" and P6 stated: "pixels themselves are related to the maths and how the math ... aids the encryption and I'd be worried if it's done by maths can the process be reversed".

Perceptions of privacy features of DP and the extent of the suitability of metaphors: Themes T5 and T9–T12 relate to RQ2 and RQ3, as they specifically reveal users' perceptions about the claimed data protection of DP and their understanding of its different privacy features implied or conveyed by our metaphors. In sum, participants correctly perceived that perturbation leads to privacy protection. They also understood, to varying degrees in all scenarios, that perturbation protects against identifiability and provides plausible deniability. However, in all scenarios most of the participants understood the trade-off between accuracy and privacy protection. An analysis of users' perceptions of privacy features of DP revealed several misconceptions, including reversibility of the process (e.g. data distortion) and the perception of DP as encryption. People also had varied perceptions about protection against adversaries with auxiliary information, preferences for the level of distortion and acceptance of and perceptions about remaining risks across all scenarios.

Table 2, which is an updated version of Table 1 based on the themes relating to RQ3, summarizes the extent of the suitability of our metaphors. Y in the table implies that the feature was understood by the majority of participants (80% or more), while N means that the feature was not understood by most of them (20% or less). P shows diversity in understanding, that is, an indication that the feature was perceived by some of the participants. P* means that although the auxiliary information was perceived to be of no help for re-identification by some participants, the reasons behind it were related to the misconception that aggregation would sufficiently protect their privacy.

Table 2: Features of functionality list understood (or not) by data subjects: Yes (Y), No (N), Partially (P)

Metaphor/feature	F1	F3	F4	F5	F6	F7	F8	Context
Spinner	Y	Y	Y	Y	Y	Y	P	Scenario 1
Noisy single picture	P	Y	Y	Y	Y	N	P	Scenario 1
Noisy picture - combined	P	Y	Y	P	Y	P*	P	Scenario 2
Distorted brain model	P	Y	Y	P	Y	P*	P	Scenario 3

T5: Perceptions of claimed protection of DP. Analysing

users' perceptions of claimed privacy protection that they assumed was conveyed by the metaphors revealed their misconceptions of DP (1) and their perception of claimed protection by distortion (2). The only common misconception among all scenarios was the perception of DP (noise addition/perturbation) being reversible (1.1). However, in SC1 the reversibility of DP was triggered by the noisy picture metaphor and not by the spinner metaphor. Other common misconceptions, at least between two of the scenarios, include the perception of DP enabling selective disclosure (1.2-SC1,2), the perception of perturbation on individual data records instead of on the aggregate level in SC2 or on the model in SC3 (1.3- SC2,3). Further, there was the perception that aggregation provides enough privacy (1.4-SC2,3). For example, P15 stated: "I believe that the picture is safe enough because it is a combination and it's not linked to any specific person". Some participants believed that distortion would selectively add noise to parts of data or exclude sensitive parts of data and share the rest; for example, P14 stated: "But since they can't hide everything using this system some of my other data probably, which are not this important, can be probably leaked". In SC1, based on the noisy picture metaphor, the description was taken literally (1.5-SC1) and led to the perception of distortion as pixelation of data, or as P9 expressed it: "I think they also try to either blur or in this case the classic mosaic censorship". Further, the pixelated picture metaphor led to the perception of DP as encryption (1.6-SC1). In SC2, it was assumed that how DP works was a secret, which led to the misconception that knowledge of DP by someone could reveal information about individuals (1.7-SC2) if that person accessed differentially private results of analysis. For example, P12 stated: "Because they know the algorithms and the mathematical equation that are needed to get this level of distortion. They could reverse it".

Almost all participants in SC1 and half of the participants in SC2 and SC3 perceived that perturbation protects privacy (2.1). However, in SC1 participants' opinions varied regarding the metaphor that better conveyed that distortion protects privacy. While almost half of the participants believed that the noisy picture better showed the amount of distortion and how it protected privacy, two believed that the spinner metaphor better communicated the unidentifiability feature.

Further, distortion was believed to protect against identifiability or to provide plausible deniability (2.2) to a varying degree in all scenarios. While in SC1 the majority understood it well, in SC2 and SC3 few perceived it correctly. However, using the example of having a unique feature in a population resulted in helping participants (almost all in SC2) to perceive the need for distortion even when aggregation is in place and that it can protect against identifiability, even with unique features (see also 7). The metaphor in SC3 led to confusion about distortion and privacy protection (2.3-SC3). The brain icon often contributed to participants' confusion and was partly misinterpreted and taken literally as images

of users' brains. People had different perceptions of what a model was and what it meant to distort a model. In SC1, a comparison of opinions on two metaphors revealed there were different perceptions on the level of privacy protection based on the metaphors (2.4-SC1). The spinner was perceived to provide better privacy protection. This was among the reasons why almost half of the participants expressed a preference to be exposed to a system illustrated by the spinner metaphor than to one illustrated by the noisy picture metaphor. Interestingly, the results in SC1 revealed that the perception of distortion (gained from the metaphor) is not easily transferable/applicable to other contexts (2.5-SC1). Although it generally made sense to the participants that distortion could protect privacy, it was hard to understand what distortion was and how it would affect data and its accuracy if we had data types other than pictures or YES/NO questions.

Perceptions about the claimed protection after exposure to the metaphors showed varied perceptions about data access by different actors (3) across all scenarios. Understanding of what the company could access contributed to people's correct perception about data access by different actors in SC1. However, in SC2 and SC3 the misconception of how DP works and confusion about the concept of a model and its distortion resulted in only about half of the participants having a correct perception about data access by different actors.

People also had various perceptions about protection against adversaries with auxiliary information (4) across all scenarios. In SC1, based on the noisy picture metaphor, the auxiliary information was perceived to be helpful for the identification (4.1-SC1). However, based on the spinner metaphor, the auxiliary information was perceived to be of no help for re-identification (4.2) of users, given users could lie in the answers perturbed by the spinner. Almost all participants (9/10) in SC1 believed that no one could distinguish actual and random answers from each other. In SC2 and SC3, auxiliary information was mostly perceived to be of no help for re-identification. However, the reason for this perception was the misconception that aggregation would sufficiently protect their privacy and no one with or without extra information about users could identify them.

T9: Perceptions of the accuracy-privacy trade-off. There were various perceptions about the accuracy-privacy trade-off of DP (1) among participants in all three scenarios. Most participants in all scenarios understood the trade-off. In SC1, everyone understood the trade-off for the noisy picture metaphor, and the majority stated that the trade-off is better conveyed by the noisy picture than by the spinner; that is, it shows a clearer progression of noise and its effects on accuracy. However, problems in understanding different terminologies or trade-off elements (2) were reported, which contributed to the misunderstanding of the trade-offs. There were different perceived consequences of trade-offs (3) among the participants in all three scenarios. Several consequences of a

lack of accuracy regarding the expense of privacy protection were perceived, including misguided or inaccurate information (3.1-SC1, SC2, SC3), service dissatisfaction (3.2- SC1,3), unreliable recommendations (3.3- SC1,3), application uselessness (3.4- SC1,3) and trust concerns (3.5-SC3). In addition, it was noted for SC1 that the context matters when it comes to trade-offs. For example, P4 stated: *".. because this is health issue so it's not always good to share the wrong information"*. Furthermore, in SC3 it was noted that distortion in long term could lead to false results and would provide no benefits.

T10: Preferences about distortion levels. The general preferences about distortion levels varied across the scenarios. In SC1, participants' preferences regarding the noisy picture varied from no noise to high noise. However, there was a consensus in SC1 regarding the spinner picture; all of the participants preferred the spinner with less probability of revealing true responses. In SC2, four participants indicated that a balance between privacy and utility is important. For example, regarding distortion preferences P12 stated: *"in the medium distortion I think there is the perfect balance"*. Likewise, in SC3 five participants indicated a preference for a medium level of distortion to balance privacy against utility. In addition, it was indicated in SC1 that the level of perturbation/distortion depends on context (i.e. health) and the amount of data to be shared.

T11: Varied acceptance/perceptions of remaining risks. There were five, six and five responses in SC1, SC2 and SC3, respectively concerning the remaining risks the participants perceived (1). Many indicated their perceptions of risks were part of their general perception of privacy risks online, such as through hacker attacks or through the possible misuse of the vast amount of personal data collected about people. For example, P9 stated: *"Every single minute of our life.. we are being tracked be it by the Internet browsing history or Google Maps... It's a privacy concern but nothing new"*. However, when it came to accepting the remaining risks, only one in SC1 refused to accept the remaining risks. There were five responses from SC1 and seven responses from each of SC2 and SC3 that indicated the participants would accept the remaining risks (2). There were three, one and three participants, in SC1, SC2 and SC3, respectively, who indicated that they either have no concerns about or no knowledge of any remaining risks (3) and that they trust the mechanism to protect their data. However, most participants across all scenarios indicated that information about the remaining risks is needed for decision making (4).

T12: Users' input/suggestions for DP alternatives. Distortion was described in different ways (1). In all scenarios, several participants described distortion as the change of original data to protect privacy (1.1). Distortion was also described as something that masks/hides data (1.2 - SC1,2) or filters/removes data (1.3-SC3). Nonetheless, how people described the privacy features of DP varied in different scenarios (2). In SC1, all those who were exposed to the spinner

metaphor first and asked to describe DP in their own words highlighted the plausible deniability without referring to the privacy-accuracy trade-off (2.1-SC1,2,3). However, most of those who were exposed to the noisy picture metaphor first referred to the trade-off and the effects of distortion on the accuracy of data (2.2-SC1,2). In SC2, half of the participants referred to the trade-off (2.2-SC1,2). The rest just highlighted the privacy protection features of distortion. They confirmed the importance of including the trade-off feature in their description only after being prompted by the moderator. In SC3, most participants highlighted the protection/security that DP provides and did not mention the trade-off (2.1-SC1,2,3). Participants were confused about the meaning of distortion in the context. For example, P29 said: “*They won’t send our face but what they are sending?*”. When asked to describe DP, four participants still referred to distortion on individual selfies than distortion on an aggregate level.

The participants’ alternatives to the metaphor to describe DP (3) include DP as pseudonymization (3.1) in SC1–SC3, DP as a generalization (e.g. using ranges instead of single data points) (3.2- SC1,2), and DP as encryption or a technique that mixes data in SC1 and SC2 (3.3-SC1,2). Further, participants in SC1 and SC3 suggested text-based metaphors/examples (3.4-SC1,3) to describe distortion and the trade-off between privacy and accuracy. For example, P30 stated: “*That some of the words are... made completely meaningless*”. Analysing users’ descriptions and suggested alternatives did not result in any suitable new metaphors for DP.

7 Discussion

Metaphors can influence how people think about a wide range of issues (e.g. [40]), concepts and experiences [23, 37]. However, metaphorical descriptions may come with specific problems. For instance, metaphorical mappings are partial. They highlight some features of a target domain and de-emphasise others [44] or imply features that do not exist [9]. Cognitive, affective and social-pragmatic factors also moderate the power of metaphors [44]. Our interview results showed the plausible suitability of our metaphors, each to a varying degree (see Table 2), to convey the privacy features of DP to lay users. However, at the same time, our study reveals and confirms several challenges that require further attention if we intend to use metaphors:

Privacy-accuracy trade-off in focus. Because the feature of accuracy loss is prominently demonstrated by the metaphors, some participants defined DP as accuracy loss and/or emphasised the accuracy loss characteristic more than the privacy protection features of DP. This also contributed to participants’ accuracy loss-related concerns regarding DP and was a factor for not trusting DP.

Earlier work on *differential identifiability* [10, 32] suggests that information on identification risk reduction is of more relevance for policy makers than information on how to ap-

proach the trade-off; therefore, it should be in focus when explaining DP in an understandable way. Our interview results confirmed that identification risks are of special interest and are a general concern even for lay users. Therefore, we recommend future research on the effects of DP explanations (in metaphoric or other forms) that emphasise the reduction of identification risks when explaining DP to different groups of users. This is in line with the recommendation of Wu et al. [46] based on a related study on mental models of encryption. They suggest improved risk communication focusing on the *why* in terms of benefits for the user rather than on *how* the technology works. Our metaphors mainly convey how the technology works by showing privacy protection through the addition of noise. In addition to communicating the benefits of reduced identification risks (and thus emphasising the “why”), users should be guided regarding adequate identification risks per context and the implications (similar to what is also suggested by [35]).

Conveying the feature of plausible deniability. In contrast to the privacy-accuracy trade-off, other features of DP, such as plausible deniability (F1), were not as clearly conveyed to the participants, with an exception of the spinner metaphor. Plausible deniability can be perceived as a benefit by users for accepting differentially private data analyses. Therefore, it should preferably be communicated to users in accordance with Wu et al.’s recommendation of focusing on the benefits for users [46]. However, an illustrative example for the noisy picture metaphors (pixelating pictures) provided in a follow-up question (to Q24) during the interviews helped several participants understand that even people with uniquely identifiable features should not stick out in differentially private data releases. The noisy picture metaphors could be improved by directly integrating the following illustrative example as a metaphor extension for SC2: “*One of the pictures shows a person with a unique characteristic (e.g. a spot on the lips), which is still visible in the combined picture, while not visible any longer in the perturbed combined picture*”. This extension helps clarify plausible deniability and also shows that aggregation alone is not sufficiently protective. Therefore, our suggested improvement can also address another common misconception and incorrect threat model concerning statistical inference attacks that several participants had for SC2. Previously, Wu et al. [46] recommended explaining the strength of a technology in terms of the capabilities of likely attackers; our proposed improvement follows this recommendation.

Misconceptions based on digital world analogies. Misconceptions about DP are likely triggered by participants’ knowledge of security technologies that they are familiar with and that they relate to DP by assuming that DP would have the same features. For instance, the noisy picture metaphor based on pixelation could be related to encryption and could lead to the assumption that DP is reversible, a misconception that largely appeared for the noisy picture and brain metaphors in

all scenarios but was not observed for the spinner metaphor. Similarly, two of the participants (out of four) who were familiar with VPNs and their feature of hiding IP addresses perceived DP as selectively hiding data (“black out”, “filter out” data), and one participant who heard of firewalls understood DP as a means of access control. Similar issues with digital world analogies that are made and that may impact the users’ mental models of new privacy technologies they are unfamiliar with were observed earlier [7, 8]. Hence, besides considering real-world analogies, DP metaphors should address the challenge of catering for digital world analogies that users may make.

Usable transparency: challenges and possible solutions.

Transparency about various aspects related to DP was named a trust factor by participants, who demanded information in a clear and easy-to-digest form. However, not all the aspects of interest, including all essential privacy features of DP, can be conveyed well by a single metaphor. In addition, interest in transparency of DP may vary significantly among different people. While about half of our participants stated their interest in how DP works, others were only interested in its privacy functionality, remaining risks and consequences. Further, our results showed that individuals had varied and not always correct perceptions of the different privacy features of DP (e.g. F5, F7, F8). In addition, our findings confirmed the problem that individuals lack clear and correct mental models of threats, which was also highlighted by [38] in a study on metaphors for E2E encryption.

Therefore, we suggest complementing metaphor illustrations with additional information when suitable. The additional information should highlight important aspects not sufficiently conveyed by the metaphor and should allow users to easily access additional information of their choice (e.g. by using multi-layered policy statements with links to sub-pages with various information and varying details on DP). In conformance with the recommendation by [38], future work should focus on finding information and complementing metaphoric illustrations that can change mental models and correct persistent misconceptions that individuals commonly have.

Metaphorical explanations: a quandary. Finally, our study also demonstrated and confirmed that metaphoric explanations inherently suffer from several shortcomings that we need to consider and counteract when we use metaphors to explain privacy technologies to users. Complementing metaphors with suitable additional information, as suggested above, can be one way to counteract these shortcomings.

Problems to abstract: Users might either take metaphors literally or have problems applying the explained features to another context. For instance, our study revealed that the noisy picture metaphor for distortion was generally understood for pictures as data types. However, when asked to apply the concept of distortion to numbers, several participants literally applied it by hiding/blurring numbers.

Different perceptions of the level of privacy protection across metaphors: Two metaphors for the same concept may result in different perceptions of the level of privacy protection. Half of our participants in SC1 preferred to be exposed to the spinner metaphor because they assumed it provided a better privacy-accuracy trade-off, although almost all believed that this trade-off was easier to understand with the noisy picture. The diverse levels of abstractions of the underlying system as a result of using different metaphors impose the risk of different (inaccurate) perceptions of privacy protection.

Conceptual baggage: Our interviews confirmed that metaphors may convey negative or positive features that the system does not have. Such features, if positive, may create an incorrect sense of privacy protection or, if negative, may affect trust and data sharing decisions. For example, our interview results revealed that a noisy picture metaphor conveyed that people with auxiliary information could identify users. Our results likewise revealed that adding noise to pictures could have resulted in the perception that this process was reversible. This conceptual baggage of the noisy picture metaphor negatively impacted our participants’ trust and data sharing.

Limitations. We conducted the interviews online with participants’ cameras turned off to preserve their privacy, which could have hindered our observations of their attentiveness. However, all participants appeared to be very engaged and attentive in the interviews. Further, our sample mainly consisted of young educated participants, which could have contributed to their understanding of the privacy technique described. However, it could have also negatively impacted their understanding of metaphors, that is, misconceptions due to associating DP with other familiar techniques.

8 Conclusion

This article presents our investigation of the suitability of metaphors to explain differentially private data analyses to lay users to facilitate their informed decisions. We highlight that there is a high interest in usable transparency of DP and privacy protection in general, with different preferences for various aspects (privacy functionality and/or structural information on how DP works) and levels of detail. Our results showed the plausible suitability of the metaphors presented to explain some privacy features of DP to users. We also discuss the misconceptions that result from the metaphors and the challenges of using them. While some of the issues can be addressed by improving the metaphors, others are rooted in the inherent limitations of metaphors. Further research is needed to address these challenges and investigate the type of information that should be provided to lay users to complement metaphoric illustrations to explain the functionalities of DP and correct common misconceptions.

Acknowledgments

This work was funded by the H2020 Framework of the European Commission under Grant Agreement No. 786767 (PA-PAYA project) and by the Swedish Knowledge Foundation (TRUEdig project). The work was also partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

- [1] Differential privacy - simply explained. <https://www.youtube.com/watch?v=gI0wk1CX1sQ>. Accessed: 2022-02-16.
- [2] Differential privacy, an easy case. <https://accuracyandprivacy.substack.com/>. Accessed: 2022-02-16.
- [3] Explaining differential privacy in 3 levels of difficulty. <https://aircloak.com/explaining-differential-privacy/>. Accessed: 2022-02-16.
- [4] What is differential privacy? <https://www.youtube.com/watch?v=-JRURYTfBXQ>. Accessed: 2022-02-16.
- [5] John M Abowd, Gary L Benedetto, Simson L Garfinkel, Scot A Dahl, Aref N Dajani, Matthew Graham, Michael B Hawes, Vishesh Karwa, Daniel Kifer, Hang Kim, et al. The modernization of statistical disclosure limitation at the US Census Bureau. <https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf>.
- [6] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Ninth Symposium on Usable Privacy and Security (SOUPS)*, pages 1–11, 2013.
- [7] Ala Sarah Alaqra, Simone Fischer-Hübner, and Erik Framner. Enhancing privacy controls for patients via a selective authentic electronic health record exchange service: Qualitative study of perspectives by medical professionals and patients. *Journal of medical Internet research*, 20(12):e10954, 2018.
- [8] Ala Sarah Alaqra, Bridget Kane, and Simone Fischer-Hübner. Machine learning-based analysis of encrypted medical data in the cloud: Qualitative study of expert stakeholders' perspectives. *JMIR human factors*, 8(3):e21810, 2021.
- [9] James L. Alty, Roger P. Knott, Ben Anderson, and Michael Smyth. A framework for engineering metaphor at the user interface. *Interacting with computers*, 13(2):301–322, 2000.
- [10] Daniel Bernau, Günther Eibl, Philip W Grassal, Hannah Keller, and Florian Kerschbaum. Quantifying identifiability to choose and audit ϵ in differentially private deep learning. *arXiv preprint arXiv:2103.02913*, 2021.
- [11] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- [12] Brooke Bullek, Stephanie Garboski, Darakhshan J. Mir, and Evan M. Peck. Towards understanding differential privacy: When do people trust randomized response technique? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 3833–3837. ACM, 2017.
- [13] Herbert H Clark. *Using Language*. Cambridge University Press, 1996.
- [14] Louise Clark and M. Angela Sasse. Conceptual design reconsidered: The case of the Internet session directory tool. In Harold Thimbleby, Brid O'Conaill, and Peter J. Thomas, editors, *People and Computers XII*, pages 67–84. Springer, 1997.
- [15] Aloni Cohen and Kobbi Nissim. Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2020.
- [16] Rachel Cummings and Deven Desai. The role of differential privacy in GDPR compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.
- [17] Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. "I need a better description": An investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3037–3052, 2021.
- [18] Albese Demjaha, Jonathan M Spring, Ingolf Becker, Simon Parkin, and M Angela Sasse. Metaphors considered harmful? an exploratory study of the effectiveness of functional metaphors for end-to-end encryption. In *Proc. USEC*, volume 2018, 2018.
- [19] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.
- [20] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12. Springer, 2006.

- [21] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- [22] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.
- [23] Kristen C Elmore and Myra Luna-Lucero. Light bulbs or seeds? how metaphors for ideas influence judgments about genius. *Social Psychological and Personality Science*, 8(2):200–208, 2017.
- [24] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.
- [25] Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. PSI (Ψ): A private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016.
- [26] Michael Hay, Ashwin Machanavajjhala, Gerome Miklau, Yan Chen, Dan Zhang, and George Bissias. Exploring privacy-accuracy tradeoffs using DPComp. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2101–2104, 2016.
- [27] Mark F St John, Grit Denker, Peeter Laud, Karsten Martiny, Alisa Pankova, and Dusko Pavlovic. Decision support for sharing data using differential privacy. In *2021 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 26–35. IEEE, 2021.
- [28] Noah Johnson, Joseph P Near, Joseph M Hellerstein, and Dawn Song. Chorus: Differential privacy via query rewriting. *arXiv preprint arXiv:1809.07750*, 2018.
- [29] Farzaneh Karegar and Simone Fischer-Hübner. Vision: A noisy picture or a picker wheel to spin? exploring suitable metaphors for differentially private data analyses. In *European Symposium on Usable Security 2021, EuroUSEC '21*, page 29–35, New York, NY, USA, 2021. ACM.
- [30] Krishnaram Kenthapadi and Thanh TL Tran. Pripearl: A framework for privacy-preserving analytics and reporting at LinkedIn. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2183–2191, 2018.
- [31] Jong Wook Kim, Kennedy Edemacu, Jong Seon Kim, Yon Dohn Chung, and Beakcheol Jang. A survey of differential privacy-based techniques and their applicability to location-based services. *Computers & Security*, 111:102464, 2021.
- [32] Jaewoo Lee and Chris Clifton. Differential identifiability. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12*, page 1041–1049, New York, NY, USA, 2012. ACM.
- [33] Meng Li, Liehuang Zhu, Zijian Zhang, and Rixin Xu. Achieving differential privacy of trajectory data publishing in participatory sensing. *Information Sciences*, 400:1–13, 2017.
- [34] Vivian Genaro Motti and Kelly Caine. Towards a visual vocabulary for privacy concepts. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 1078–1082. SAGE Publications Sage CA: Los Angeles, CA, 2016.
- [35] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. Visualizing privacy-utility trade-offs in differentially private data releases. *arXiv preprint arXiv:2201.05964*, 2022.
- [36] Daniel L Oberski and Frauke Kreuter. Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review: HDSR*, 2(1):1–21, 2020.
- [37] Shani Robins and Richard E Mayer. The metaphor framing effect: Metaphorical reasoning about text-based dilemmas. *Discourse Processes*, 30(1):57–86, 2000.
- [38] Leonie Schaewitz, David Lakotta, M Angela Sasse, and Nikol Rummel. Peeking into the black box: Towards understanding user understanding of E2EE. In *European Symposium on Usable Security 2021*, pages 129–140, 2021.
- [39] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Eleventh Symposium on Usable Privacy and Security (SOUPS)*, pages 1–17, 2015.
- [40] Aaron M Scherer, Laura D Scherer, and Angela Fagerlin. Getting ahead of illness: Using metaphors to influence medical decision making. *Medical Decision Making*, 35(1):37–45, 2015.
- [41] Awanthika Senarath, Nalin AG Arachchilage, and Jill Slay. Designing privacy for you: A practical approach for user-centric privacy. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 739–752. Springer, 2017.
- [42] ADP Team et al. Learning with privacy at scale. *Apple Mach. Learn. J.*, 1(9), 2017.

- [43] Pratiksha Thaker, Mihai Budiu, Parikshit Gopalan, Udi Wieder, and Matei Zaharia. Overlook: Differentially private exploratory visualization for big data. *arXiv preprint arXiv:2006.12018*, 2020.
- [44] Paul H Thibodeau, Teenie Matlock, and Stephen J Flusberg. The role of metaphor in communication and thought. *Language and Linguistics Compass*, 13(5):e12327, 2019.
- [45] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.
- [46] Justin Wu and Daniel Zappala. When is a tree really a truck? exploring mental models of encryption. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 395–409, 2018.
- [47] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. Towards effective differential privacy communication for users’ data sharing decision and comprehension. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 392–410. IEEE, 2020.
- [48] Yixin Zou, Shawn Danino, Kaiwen Sun, and Florian Schaub. You ‘might’ be affected: An empirical analysis of readability and usability issues in data breach notifications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

A Data analysis scenarios - Figures

Figure 3 shows the three data analysis scenarios in our study.

B Example of scenario description

Scenario 1. “The app notifies its users, including Alex, that it is possible to receive supportive recommendations to help them cope with stressful conditions if they want and agree.

To do so, the health company needs to: a) receive stress-related information from different users. Stress-related information may include, for example, users’ responses to daily questions about their moods or users’ selfie pictures on different occasions when they feel stressed or not. b) Combine and analyse the information it collects from different users to gain insights into stressful conditions and provide remedies and assistance to cope with stress.

In this scenario, Alex trusts the wearable device and her phone but not the health company. Therefore, the information the health company receives from users through the app can negatively affect Alex’s (and other users’) privacy. Thus there

is a privacy problem. The health company may learn about Alex’s stress problems and stressful situations.

To protect users’ privacy and mitigate the privacy problem, the app applies a privacy mechanism on Alex’s (or any other user’s) input data before the personal data leaves Alex’s (or any other user’s) device. This satisfies so-called differential privacy, a formal notion of privacy that provides provable privacy assurances. To a certain extent, this differentially private mechanism prevents leakage of Alex’s (or any other user’s) actual stress records.”

C Metaphor descriptions

The metaphors were accompanied by descriptions that the moderator read to the interviewees.

Description of metaphor in Figure 2c: “Now imagine, in the scenario described, that the health app requests its users including you (as Alex) to answer some sensitive YES/NO questions about their stress conditions. The health company then receives the responses and can use the responses to analyse, for example, the proportion of users who say YES or NO to each question. The health app will protect users’ privacy by using a differentially private mechanism, as depicted in this figure.

Your health app on your phone uses a spinner wheel to perturb (change) your responses to the questions with a controlled and known probability based on the underlying mechanism before sharing them with your health company. The app spins the wheel. If it lands on YES, your true response will be revealed. If it lands on NO, it will spin the wheel again. If it lands on YES the second time, it will reveal YES and if it lands on NO, it will reveal NO regardless of your true responses. The purpose of perturbing your responses is to assure your privacy. The mechanism guarantees that what the health company can infer about your true responses is limited and negligible. You can deny, to a certain extent defined by the mechanism, if a given YES or NO response is your true response.

This figure is not a precise representation of the underlying mathematical mechanism that perturbs users’ data; it is just a simplified example of what perturbation means. Note that the outcome of the spinner, whether it lands on YES or NO, remains hidden from the health company. Although the health app deliberately perturbs its users’ responses, the health company can still benefit from the collected responses to infer the proportion of users who said YES or NO to each question.”

Description of metaphor in Figure 2a: “Now imagine, in the scenario described that the health app requests its users including you (as Alex) to share their selfies with the health company. The health company can then use the selfies and analyse, for example, the most common facial expressions of users when they are stressed. The health app will protect users’ privacy by using a differentially private mechanism, as depicted in this figure. First, you share your selfie with the

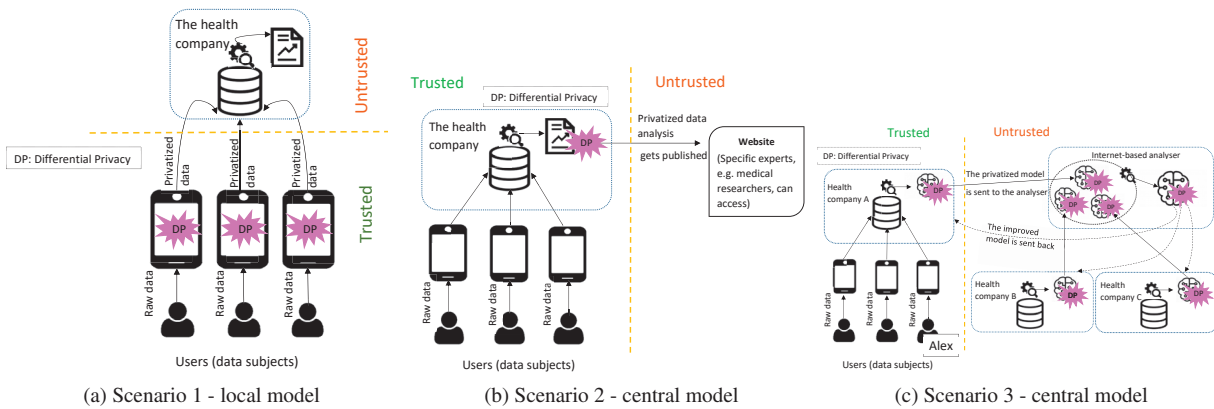


Figure 3: Data analysis scenarios.

app. Then your health app perturbs (carefully distorts) the details of your picture by adding a specific amount of noise to it (e.g. a medium amount of noise) based on the underlying mechanism before sharing it with your health company. The purpose of distorting your selfie is to assure your privacy. The mechanism guarantees that what the health company can infer about your true selfie is limited. You can deny, to a certain extent defined by the mechanism, if a shared selfie is your true selfie.

This figure is not a precise representation of the underlying mathematical mechanism that perturbs (distorts) users' data; it is just a simplified example of what perturbation means.

Although the health app deliberately adds noise to its users' selfie pictures, the health company can still benefit from the collected selfies to infer the most common facial expressions of users when they are stressed."

Description of metaphor in Figure 2b: "Now imagine your health company wants to analyse what the common lip expression is when people do (not) feel stressed and it requests users, including you as Alex, to share their selfies so it can combine and analyse them. The health company will protect its users' privacy by using a differentially private mechanism to analyse their data, as depicted in this figure.

Before revealing the common lip expression, the health company perturbs (carefully distorts) the details of the common lip expression based on the underlying mechanism by adding a specific amount of noise to it (e.g. a medium amount of noise). The purpose of distorting the common lip expression is to assure users' privacy by limiting the effects of each individual's selfie on the analysis results, i.e. the common lip expression. Therefore, the mechanism guarantees that the likelihood of privacy harm users may face by being identified as a result of sharing their selfies and having their selfies analysed together with those of other users is limited and insignificant.

This figure is not a precise representation of the underlying mathematical mechanism that perturbs the results of data analysis; it is just a simplified example of what perturbation means.

Note that although the health company deliberately perturbs the results of data analysis, in this case the common lip expression derived, the distorted results of the analysis can still be useful for the receivers, for example, the health researchers."

Description of metaphor in Figure 2d: "Now imagine your health company wants to create a model that can recognize a user's emotion from his/her facial expression. Again, note that you can think of a model as an artificial brain that learns from its inputs. In other words, a model can be trained based on the characteristics of its inputs to do a special thing. The health company requests its users, including Alex, to share their selfies and then uses the selfies to train a model so the model can recognize emotions based on facial expressions. For example, the model can predict if a user is very happy, sad, confused, stressed, furious, etc. In this figure, you see a trained model based on users' selfies. Now if the trained model receives a user's selfie as its input, it can predict the user's emotion.

As mentioned previously, the health company protects the users' privacy by using a differentially private mechanism to train the model, as depicted in this figure. Before sharing its locally trained model with the analyser, the health company perturbs (carefully distorts) the trained model based on the underlying mechanism. This means that the health company distorts the information the model has learned from selfies randomly but in a controlled way, for example, using the medium level of distortion. The purpose of distorting the trained model is to assure users' privacy by limiting the effects of each individual's selfie on what the model has learned from the selfies. Therefore, the mechanism guarantees that the likelihood of privacy harm users may face by being identified as a result of uploading their selfies and having their selfies analysed with other selfies to train the model is limited and insignificant.

This figure is not a precise representation of the underlying mechanism that distorts a trained model; it is only a simple example of what distortion means.

Although each health company deliberately distorts its

trained model, the final model is better than each of the locally trained models at recognizing the emotions. The final model made by the analyser is also a distorted model that protects users' privacy."

D Interview guide

This is our interview guide for the first scenario (depicted in Figure 3a). As mentioned in Section 4.2, half of the interviewees assigned to this scenario were first exposed to the spinner metaphor (Figure 2c) and then the metaphor of the noisy picture (Figure 2a). The other half were exposed to the same metaphors but in the reverse order. The interview questions for other scenarios were adjusted to fit the context.

Welcome and introduction. Participants are welcomed and instructed about the setup of Zoom and turning off their video. An introduction to the study, the goal and the different parts of the interview are provided. The consent form is given to the interviewees, and once they agree the session starts and the recording commences.

Scenario introduction and expectations discussion. The moderator first describes the persona and then describes one of the scenarios to the interviewee by showing the related figure (see Figure 3) and reading the related description provided in C. The following questions are asked.

Q1. Have you heard about any privacy protection techniques (techniques to guarantee users' privacy and to improve it)? Have you ever heard about differential privacy?

Q2. In what context did you hear about it?

Q3. Do you know what differential privacy is? Can you explain it in your own words?

Participants are then told to pretend they are Alex, who is using a wearable device, and have received the notification in the scenario while answering the following questions.

Q4. Would you agree to share your data to be analysed in the way described? What factors did play a role in the decision for Alex?

Q5. How did the differential privacy mechanism play a role in your decision? Would it matter if another mechanism were used to protect your privacy instead of differential privacy?

Q6. What should have been different so you would agree?

Q7. What do you want to know about the mechanism applied (the differentially private mechanism) to protect your privacy? What information would you like to be added to the scenario?

Q8. What would be the benefits for you if you agreed? What would be the risks for you?

Q9. In this scenario, from whom do you expect your actual stress-related data to be hidden? (follow-up: could your health app see your actual stress-related data? What about your health company?)

Q10. In this scenario, it was mentioned that your privacy is protected against potential privacy risks using a specific

mechanism. What factors do play a role for you to trust this mechanism to protect your data?

Metaphor introduction and perceptions gauging. The moderator shows a specific metaphor depending on the data analysis scenario for the interview and reads the description of the metaphor to the interviewee. The descriptions of the metaphors are provided in Appendix C. Participants are told to consider the description of differential privacy and the scenario when answering the following questions.

Q12. Would you change the decision you made on behalf of Alex in the previous step after receiving more information about differential privacy? Why?

Q13. In general, do you think that receiving information about the underlying privacy techniques a system uses would help you decide to use a system? How (in what way) could it be helpful?

Q14. Is the description of differential privacy understandable and easy to grasp for you? What is not clearly described or is missed in the description? How can the description be improved?

Q15. Is there any information that is surprising to you—you did not expect? Please elaborate.

Q16. Would you like to know more about the technical and mathematical details of the underlying differentially private mechanism? Why?

Q17. The mechanism perturbs (changes) your data in a controlled way. Can you explain in your own words what it means to perturb your responses? How does data perturbation protect your privacy?

Q18. How would your privacy be better protected—using a spinner with a bigger area for YES or a smaller area for YES? What happens if the area for YES is zero and is 100% for NO?

Q19. Which of the spinners do you prefer to be used to perturb your data? Why?

Q20. Can you explain whether there is a trade-off between the accuracy of the data analysis results and the privacy of your data?

Q21. How would you as Alex be affected if the data analysis results are not accurate? Would you rely on the recommendations the app gives you to cope with stress? Why?

Q22. Imagine that as Alex you agreed to share your stress-related information in the scenario. Which of the following entities would be able to see your (Alex's) actual stress-related data? (Why do you think so?): a. Hackers who access the database of the health company. b. People who know how the differentially private mechanisms work if they access the perturbed data.

Q23. Would the health company be able to prove that a YES answer is your true answer? Would a close friend (if she/he gets access to your perturbed answers) be able to prove that a YES answer is your true answer?

Q24. Imagine you agreed and that the health company analysed the proportion of users who said YES or NO to

each question, based on the perturbed responses it received. If you did not agree to share your responses, how would it affect the proportions of YES/NO responses to a question that is calculated by the health company based on the perturbed responses it receives? (follow-up: Do you think the proportion of users who said YES/NO to each question greatly depend on your decision to share your responses? What if the responses were not perturbed?) (Follow-up (SC2): Imagine you as Alex have a feature that no other user has. For example, you have a dark spot on your lip. Therefore, the common lip expression derived will include a dark spot as well. How would it change if your selfie was not included? Now imagine that we distort the common lip expression so that the dark spot is not shown. How would the distorted lip expression change if you did not agree to share your selfie?)

Q25. How would you describe the likelihood of remaining privacy risks? Would you accept the remaining risks? Would more information about the remaining risks be of your use in deciding to share your data or not?

Q26. Now that you know more about differential privacy, would you trust this method in general to protect your privacy? Why? (If the answer is NO:) What are your concerns in this regard?

Q27. How would you describe differential privacy to someone who does not know about it? Can you think of any alternative description/example of perturbation (data changes)

other than the one we used to describe the concept?

Second metaphor introduction and discussion. The moderator reads the second metaphor and then asks:

Q28. Can you describe how your privacy (as Alex) will be protected by perturbing (distorting) your data? How would your privacy be better protected, by adding more noise or by reducing the noise?

Q29. How can data distortion affect the accuracy of data the health company receives?

Q30. Which of these two description better conveys the trade-off between accuracy and privacy? Why?

Q31. Which of these two descriptions is easier for you to understand? Why?

Q32. Which one do you prefer to be exposed to when you want to decide to use a differentially private system? Why?

Q33. Which of these two descriptions better helps you to relate data perturbation (changing your data or distorting your data) to privacy protection? (follow-up: What are the shortcomings of this description? How it can be improved?)

Feedback and thanks. The moderator asks the participants if they have any comments and/or questions. Then the moderator thanks the participants for their participation.

E Themes



Figure 4: Themes with the sub-themes (note that the themes without sub-themes are not listed.)

An Empirical Study of a Decentralized Identity Wallet: Usability, Security, and Perspectives on User Control

Maina Korir
University of Bedfordshire*
maina.korir@beds.ac.uk

Simon Parkin
TU Delft
s.e.parkin@tudelft.nl

Paul Dunphy
OneSpan
paul.dunphy@onespan.com

Abstract

User-centric digital identity initiatives are emerging with a mission to shift control over online identity disclosures to the individual. However, there is little representation of prospective users in discussions of the merits of empowering users with new data management responsibilities and the acceptability of new technologies. We conducted a user study comprising a contextual inquiry and semi-structured interviews using a prototype *decentralized identity* wallet app with 30 online participants. Our usability analysis uncovered misunderstandings about decentralized identifiers (DIDs) and pain points relating to using QR codes and following the signposting of cross-device user journeys. In addition, the technology did not readily resolve questions about whether the user, identity provider, or relying party was in control of data at crucial moments. We also learned that users' judgments of data minimization encompass a broader scope of issues than simply the technical provision of the identity wallet. Our results contribute to understanding future user-centric identity technologies from the view of privacy and user acceptance.

1 Introduction

Identity fraud impacts around 10 million Americans per year [70] and costs the global economy \$5 trillion per year [57]. In addition, over 90% of American consumers believe they have lost control over how their personal information is collected and used [60]. At the same time, a groundswell of new digital

infrastructures [51, 75] and political initiatives are creating a renewed vigor to explore new and better ways to transact online using our identity. A common goal is to leverage user-centric identity technologies while improving access to vital services, including those provided by governments, healthcare providers, travel hubs, and financial institutions. The European Union pursues a mission to create a European Digital Identity [29]. National governments are drafting identity governance frameworks, e.g. United Kingdom [69], Canada [21], and the United States [17]. Large companies also modify their product offerings [40, 52] to accommodate privacy-friendly decentralized identity (also referred to as self-sovereign identity) [59].

One technology that is fast emerging as a cornerstone of most, if not all, future proposals for user-centric digital identity schemes is the *identity wallet: a tool that enables end-users to prove aspects of their identity online in a secure and privacy-respectful manner*. An identity wallet enables users to have meaningful control over the transfer and disclosure of verified personal information when identity is federated between online services. One core function of the technology is to collate cryptographic attestations of personal attributes (e.g. age, name) or entitlements (e.g. right to work) from an identity provider in a form verifiable by a second online service. German law already permits identity wallets to be legally used within anti-money laundering regulations to access financial services [22] and the federal government has already deployed its own identity wallet [6]. Private companies have also created identity wallet technologies for deployment in their products [30, 45, 46].

While the velocity of design and rollout of identity wallets is increasing, we lack knowledge about the characteristics of a successful user-centric identity wallet. We see three reasons we must further investigate these new technologies. Firstly, an identity wallet is a complex technology that integrates multiple processes that pertain to security and privacy; secondly, there is an untested assumption that the perception of enhanced control over the disclosure of personal data will drive user acceptance. Finally, while identity wallets are still in a

*This work was led by the first author during an internship at OneSpan

formative stage, there are few reported trials or experiments focused on the user experience.

In this paper, we report the results of a user study of an identity wallet prototype designed using tools for decentralized identity [59] – the most privacy-respecting vision for online user-centric identity. We conducted a user study comprised of a *contextual inquiry* and *semi-structured interview* with 30 participants recruited from the United Kingdom and the United States. Our findings cut across the domains of usability, security, and privacy. For example, while the most ambitious vision of decentralized identity requires user autonomy for identity data and credential storage, there was a dominant expectation that (at least) one trusted party provides account recovery if wallet data were lost or corrupted. The root of this finding is that participants reported not being fearful of losing an identity wallet. Also, the accuracy of user judgments about the oversight held by external parties was mixed, which is a concern if the assumed benefit of identity wallets to users is an acute understanding of data control. More generally, we learned that today, participants are dependent on paper-based methods to identify themselves to identity-critical services. However, we also found that while poor experiences onboarding with paper documents are common, participants increasingly have experiences with improved technology for document scanning, data parsing, biometric checks, etc. Therefore there appears to be an arms race between new user-centric identity methods that preserve privacy and more efficient ways to capture and parse privacy-invasive data for identity purposes. The contribution of this paper is as follows:

- We present insights into the user perceptions and acceptance of the key components of a decentralized identity wallet: decentralized identifiers (DIDs), verifiable credentials (VCs), and identity proofs. More specifically, we shed light on the perceptions of *user control* delivered by identity wallets in the context of decentralized identity, in that technically constructed privacy benefits might constitute small drivers for uptake.
- We propose a method to capture users' mental models of security and privacy in the context of identity wallets that is also applicable to other user-centric technologies. Our lightweight mental model scale prompted participants to express their intuition and understanding of the technology and geography of data. The technique informs approaches for determining how well users' understanding of user-centric services impacts acceptance, an important issue with, e.g. FIDO2 authentication [43].
- We detail usability measures and user journey challenges inherent to decentralized identity wallets. We also draw parallels to similar issues inherent to other user-centric technologies, such as FIDO [16] and FIDO2 [43], where learnability is a particular challenge for end-users. Finally, we propose improvements for identity wallet technology.

The remainder of the paper is organized as follows: Section 2 details related work, and Section 3 introduces key concepts of identity wallets and the design of our prototype. Our user study design is detailed in Section 4, followed by presentation of results (Section 5), limitations (Section 6) and discussion (Section 7). Our concluding remarks are in Section 8.

2 Related Work

2.1 Federated Identity Management

Federated Identity Management (FIM) is the technology and process to transfer trustworthy attributes from one security domain to another. FIM techniques and technologies are standard in orchestrated and closed deployments (e.g. a workplace), but it is a more significant challenge to achieve FIM on the open Internet where, back in 2001, users had an average of 16 accounts to manage [63]. In FIM deployments, the number of parties involved in an identity transaction increases from two parties to three, and we get what is known as the *trust triangle* [59] that has three roles: issuer, verifier, and holder (or, identity provider, relying party, and user).

The seven laws of identity [14] are heuristics that support the evaluation of identity schemes and are particularly relevant to FIM. Microsoft designed CardSpace around 2006 to instantiate those seven laws and create a universal identity layer for the Internet. Indeed, one claimed design priority of the Microsoft Infocard system was the user experience [15].

Landau and Moore [42] propose that FIM is a technology of great promise whose wider adoption has so far been disappointing, and also describe some of the economic *tussles* that can make or break FIM in a specific application. They propose that so far, identity providers and service providers have tussled about who controls user data rather than the provision of benefits to users. Gov.Verify is one British government system that federates citizen identity across government services. Gov.Verify is beset by privacy concerns [8] along with citizen concerns about interacting with the government via private companies [12].

2.2 Web Single-Sign On

Single-sign on is one critical application of FIM, and this exists on the open web, most commonly in the form of the standards-based *OpenID Connect* (OIDC), or OAuth 2.0 [1]. Google provides an OIDC compatible sign-on, but *Facebook Connect* provides a proprietary sign-on technology that leverages OAuth 2.0. While technically different, Facebook's single sign-on mechanism is conceptually similar. One significant difference is that OIDC offers a taxonomy of the attributes and data formats that an application can provide and consume, whereas OAuth 2.0 does not [18]. Facebook proposed in 2010 that there were more than 250 million users

of Facebook Connect [71], and research has expressed concern that users were not making informed consent for sharing attributes with online services [27].

A study of web single sign-on relying parties suggested Facebook followed by Google were dominant identity providers, and that 75% of relying parties request more than authentication state from identity providers [18]. One reason for a relying party to prefer one identity provider relates to the *attributes* that an identity provider can provide to a relying party. These attributes could be trustworthy to different degrees. For example, a first name and surname may not be reliable from Google. However, Facebook performs some basic validation of names, which might make Facebook more desirable if an application requires the user's "real" name [31].

2.3 User Centricity and Decentralized Identity

User centricity is a crucial framework in Federated Identity Management (FIM) because it forces reflection on *how* to implement FIM to respect the privacy of the end-user. There are three dimensions to user-centricity: user control, architecture, and usability [7]. For example, technologies such as those compliant with FIDO standards [35] are user-centric.

Decentralized identity – also known as *self-sovereign identity* (SSI) [59] – is borne out of dissatisfaction with the privacy properties and power dynamics inherent to some user-centric identity technologies. Decentralized identity manifests as principles to reinforce the goal that the user is central to the administration of their identity [3]. Furthermore, several specific elements are commonly associated with decentralized identity: (i) an eco-system of multiple identity providers, (ii) a decentralized *trust registry* [56] - a root of trust that contains tamper-resistant shared records and has no single point of failure, and (iii) an identity wallet for end-users that stores personal information and provides cryptographic techniques for privacy-friendly information disclosure. For the latter, Decentralized Identifiers (DIDs) [65] are user-generated identifiers that decouple identifiers from identity providers and are verifiable through public-key cryptography. A *verifiable credential* (VC) [72] can digitally represent attributes found in physical identity documents such as name or date of birth and new things that have no physical equivalent, such as ownership of a bank account. In addition, VCs contain digital signatures, which makes their authorship verifiable and their contents tamper-resistant. Zero knowledge proofs [34] are also considered to be relevant techniques. Candidate schemes that embody these techniques have already been proposed [48] and the user experience of constituent technologies will be critical for future uptake [25, 26].

3 Identity Wallets

The design of user-centric identity infrastructures requires the existence of a means to provide the user with control of

personal data and disclosures. In most cases, this necessitates the existence of a conceptual or visual control panel where the user can inspect the status of their entitlements and data and provide consent to, and initiate, information disclosure. There are numerous examples of approaches to design this control panel. This could, for instance, be a simple user interface displayed by a website requesting consent to disclose information to another party. For example, in the Microsoft InfoCard project, the *identity selector* [15] was built into the Windows operating system and provided a point of control where the user can select which cards (credentials) to disclose.

In the context of decentralized identity [59], this control panel takes the form of an identity wallet which stands to inherit additional complexity than seen in previous user-centric systems such as InfoCard for multiple reasons. For example, the integrity and authenticity of identity information depends upon public-key cryptography secured primarily by the wallet, the design of an identity wallet is geared to portable devices which might be lost, the user journeys cut across multiple devices and workflows are asynchronous, and the wallet must also interact with a decentralized trust registry. Furthermore, the wallet software may not be controllable by an identity provider or a relying party.

We wanted to understand the dominant design approaches inherent to decentralized identity wallets. Therefore we firstly gathered publicly available decentralized identity wallets that we could find, namely: uPort [46], Connect.Me [30], Lissi [45], ShoCard [58], and SelfKey [64]. None of these apps appeared authoritative. Therefore, we abstracted the user journeys to create an identity wallet app template (which can be seen in Figure 1). We learned that there are three key journeys envisioned in identity wallet apps:

1. **Connect Identity Wallet** - The wallet scans a QR code created by the online service and looks up the public key of the online service from a decentralized registry using its W3C Decentralized Identifier (DID) [65]. The wallet generates a new DID and shares this with the online service, and negotiates a shared key with the online service using its public key. This process results in a secure connection between the identity wallet and the online service.
2. **Obtain Credential** - The wallet requests a W3C Verifiable Credential [72] from the identity provider for attributes that were verified apriori. The identity provider sends a *credential request* to the identity wallet. The user must read and accept this credential request, and then the credential – digitally signed by the identity provider – is sent to the wallet for secure storage.
3. **Enrol Using Proof** - The end-user navigates to a new online service and selects to enrol using an identity proof. The online service sends a *proof request* JSON structure to the identity wallet that lists the attributes that the user

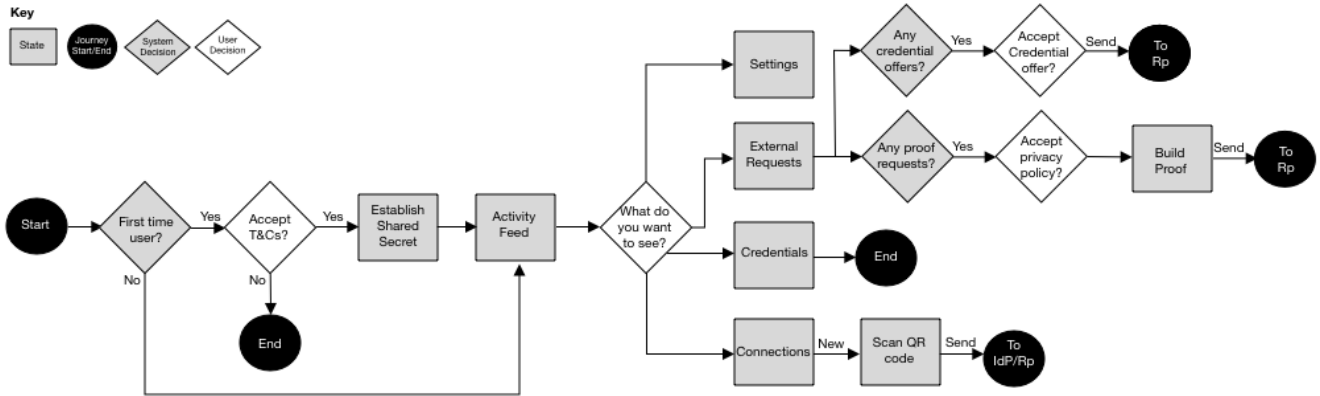


Figure 1: In order to understand the characteristics of existing decentralized identity wallets, we evaluated the user journeys of several publicly available wallets. We found that there were three journeys that apps had in common: (i) Connect Identity Wallet, (ii) Obtain Credential, and (iii) Enrol Using Proof. We also found that user journeys within the identity wallet are brief and usually involve a task switch to interact with the system of an identity provider (IdP) or relying party (Rp) using a different app or device.

must evidence before enrolment. The user then responds by matching a credential with each attribute and sending the proof, along with cryptographic proof that the wallet owns the credential(s).

We also learned that each user journey in the app is brief and requires the switch to another app or device to interact with the online system of the identity provider or relying party.

3.1 Open Challenges

An identity wallet combines processes that are individually challenging according to user-centered security and privacy research, such as understanding privacy policies, obtaining informed user consent, personal information storage, and cryptographic key management. Prior work has also highlighted the specific challenges facing user uptake of identity management technologies such as an unclear user proposition [20], lack of perceived urgency to adopt identity management technology [68], and a focus from the technology designers on owning data of the user [42]. However, it remains unclear if identity wallets will solve, or suffer from, the same issues.

Moving data-sharing processes online generally brings challenges, for instance, in how users can be supported not to over-disclose personal information when interacting with services [41]. Identity wallets act then as a consolidated tool to manage how data is shared with requesters, removing web interfaces as a potential source of confusion or friction. Identity wallets are, in essence, an attempt to provide a user-centric solution for individuals’ data-sharing practices. Encompassed in this challenge is how to encourage adoption of complex yet well-intentioned technologies while providing the necessary assurances, as with encrypted communications (e.g. [32, 67]).

Secondly, the design promise of an identity wallet is that it should deliver enhanced end-user control [7] over the storage and disclosure of identity attributes when compared to incumbent, paper-based methods. However, it is unclear whether the dominant technical framing of user control will constitute a driver of uptake for end-users. Finally, while the need for identity wallets is widely assumed, their current state as a concept means we cannot yet enumerate the challenges they will present to the security and privacy of end-users. Therefore there is a pressing need to research these challenges before large-scale deployments occur. For example, one aspiration is that 80% of Europeans will be using identity wallets by 2030 [29].

4 User Study

We conducted a user study to explore our overarching research question: *What are the user-centered privacy and security challenges facing decentralized identity wallets?* We scoped our interest in this broad research question through three sub-questions: i) Which problems do users have today to prove identity online? ii) How are the privacy properties of the technology valued by end-users? iii) What are the usability properties of identity wallets?

4.1 Methodology

To explore our research questions, we performed a *contextual inquiry* [61] which is a well-established method in human-computer interaction for uncovering requirements and problems relating to a context of use. Our contextual inquiry was composed of three tasks for participants to complete,

where data included insights from "thinking aloud" and a *semi-structured interview*. The most challenging aspect of addressing the research questions was to gather experiences of a technology that is nascent and where end-to-end implementations are not openly available. Therefore, we needed to develop our own prototype that was broadly representative of identity wallets that can be found today to simulate the experiences that users might have in practice, and draw conclusions for that entire class of technology. Future-oriented prototypes can be of great value in usable security and privacy research as they can facilitate *problem-scoping* and *problem-solving* [49].

Conducting a contextual inquiry (including think-aloud techniques) to gauge potential user acceptance of future-facing prototypes is a well-established practice in user-centered security and privacy research. For example, Lyastani et al. [37] at IEEE S&P 2020 instrumented a dummy online service with FIDO2 authentication libraries in order to collect usability insights on FIDO2 passwordless authentication, which was not widely deployed at that point in time. Sun et al. [68] at SOUPS 2011 instrumented a prototype to simulate a browser-enabled version of OpenID, though behind the scenes their prototype contained a man-in-the-middle proxy to relay login details (since websites were not compatible with the technology). Brostoff et al. [12] explored the use of federated government ID to access healthcare information using low fidelity prototypes before such a service had seen deployment.

In order to explore our interest in perceptions of privacy features, particularly user control, we created a concise *mental model* scale designed to try to hone in on security and privacy perceptions of a specific identity wallet component and also to reveal participants' intuition about their control over personal data. This mental model scale sought to check the *functional* mental model participants had of some properties of an identity wallet and how they relate to tasks; this is as opposed to a *structural* mental model of the underlying details of how the system works [23]. Such approaches are helpful to probe users' understanding of the properties of, for example, end-to-end encryption (E2EE) [19]. Similar approaches to relate beliefs to the functional workings of security-related technologies can probe, for example, beliefs about the safety of online browsing and the use of dedicated security software [74]. Here we probe a mix of functional and sentiment-driven perceptions.

Finally, we chose to situate the context of the study in the banking and financial services sector since the industry faces multiple problems in verifying customer identity in the face of anti-money laundering regulation [36]. In addition, novel proposals aim to provide a digital identity infrastructure specifically for banking [28, 76]. Furthermore, banking is a use case of importance to members of the public that we hoped would provoke curiosity and insight.

4.2 Prototype

Our prototype worked end-to-end and had three components: the identity wallet mobile app, the backend distributed ledger network, and the end-user facing websites to depict the identity provider (IdP) and the relying party (Rp). Screenshots of the Android prototype are in Figure 2. We created the identity wallet app for the Google Android platform and used the Hyperledger Indy SDK (herein Indy SDK) [39]. The Indy SDK provides functionality for several fundamental components of a decentralized identity wallet: W3C Decentralized Identifiers (DIDs), W3C Verifiable Credentials, and data retrieval from a Hyperledger Indy ledger.

We created websites that resembled fictional banks: Alpaca Bank (IdP) and Bank of Carpathia (Rp). These sites had plausible domain names for financial institutions and LetsEncrypt [44] TLS certificates. We hosted the service providers on Amazon Web Services, and both entities could read and write to the Indy ledger. A snapshot of the user interface of these services is in Figure 2.

We also created a five-node Hyperledger Indy network [39]. Indy records references to verifiable credentials and provides functions for revocation. The IdP can write to the Indy ledger to add a credential reference to a cryptographic accumulator, and the Rp can read the same cryptographic accumulator to verify the validity of credentials bundled in a proof.

4.3 Method

Before the study, we sent the participant a URL to the study information sheet and captured their consent to participate using an eSignature tool. Then, after agreeing on a convenient time for the study, we asked the participant to install the identity wallet app on their mobile device via a private URL in the Google Android Play Store.

We conducted the study as follows: The participant connected to the video call; the experimenter then checked that the participant joined the meeting on both a laptop and a mobile device. The experimenter firstly gave a brief verbal description of the study and allowed the participant to ask any questions. Next, the experimenter led the participant through a semi-structured interview that generally covered their online identity experiences. We then asked the participant to screen share on their mobile device while connected to Zoom, and to carry out individual steps for the following tasks while *thinking aloud* [61]: making a connection, obtaining a credential from the IdP, and building a proof. The experimenter noted the critical incidents encountered using Nielsen's usability incident taxonomy [54] as the participant thought aloud. The experimenter then led the participant through a second part of the semi-structured interview, using a mental model scale to identify their perceptions of privacy and the identity wallet concept, and the System Usability Scale [11] to assess the subjective usability of the identity wallet app.

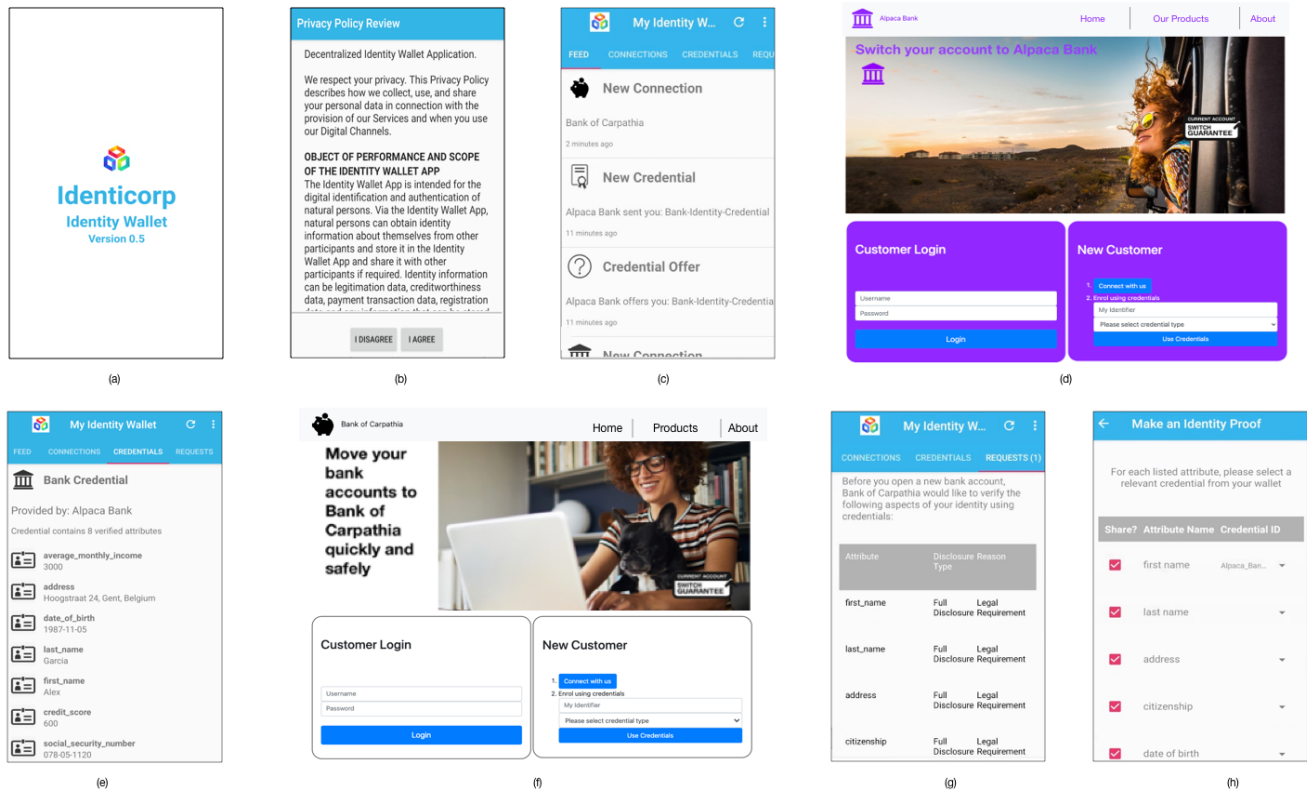


Figure 2: Screenshots from different aspects of our identity wallet prototype: (a) introductory screen, (b) privacy policy, (c) a news feed that illustrates recent and relevant events, (d) the identity provider web page, (e) an example of a verifiable credential (VC), (f) the relying party’s web page, (g) a proof request from the relying party that details the attributes that the bank needs from the end-user, and (h) the process of building an identity proof, by selecting which attributes to share and which VC to use to evidence that attribute.

The full study questionnaire accompanies this paper in the Appendix and has the following sections:

Current forms of identity and identification. Questions focused on how participants currently identify themselves, the techniques they use, how they perceive the process, and if they encounter any challenges (usability, security, and privacy).

Interactions with the identity wallet. Questions following participants’ interaction with the identity wallet app focusing on their perception of the identifier, credentials, and identity proof.

Reflection on identifiers and proofs. Questions focused on how participants perceived the opportunity to generate an identifier for themselves using the identity wallet app, to control the information they share with the Rp, the fact that their transactions are invisible to the IdP, and the security and privacy offered by the identifier, credentials, and proofs.

Usability and user expectations. Questions focused on the usability of the identity wallet app, participants’ trust in the app, and, in general, how they saw the identity wallet app

fitting into their existing practices to identify themselves.

4.3.1 Qualitative Analysis

We extracted audio from the video recording of each participant and performed a complete transcription of all sessions. The process yielded approximately 30 hours of transcribed audio data. Transcripts were then anonymized and subjected to a deductive thematic analysis using the method proposed by Braun and Clarke [9]. Our deductive analysis focused on identifying text that pertained to the sub-research questions that we describe in Section 4. Since we designed our user research sessions to address the research questions, our analysis procedure resembled an inductive analysis. First, we summarized each text extract with an open-ended code. After creating a preliminary codebook, regular meetings were held amongst the research team to understand better and refine the code book and to group codes into themes.

4.4 Participants

We recruited 30 participants for the study with an even split between male and female participants. Seventeen were from the United States and thirteen from the United Kingdom. Participant ages tended to the younger end of the spectrum: two in the range 18-24; 12 in the range 25-34; 10 in the range 35-44; 4 in the range 45-54; and finally, two aged 55+. Participants were generally highly-educated, with 73% educated to at least a bachelor's degree level and 17% with a background in computer science.

Due to restrictions on in-person research as a result of Covid-19, we conducted the study remotely and recruited participants using the research participant recruitment service *user interviews*¹. We paid each participant \$50 for a one-hour session. Given the nature of this online platform, we can assume this platform enabled us to recruit adults who were savvy users of the World Wide Web. The *user interviews* platform has been used in other user-centered research studies [10, 38]. We required that participants have a Google Android phone (minimum version 10) and one additional computing device, e.g. a laptop, to access the websites of our prototype online service providers.

4.5 Ethics

We received research ethics approval from the authors' respective organizations. The study adhered to the principles of the Menlo Report [24]. Participants received an information sheet with details about the study. They were free to participate and could withdraw at any time. There were no disadvantages for those who took part. The study complied with GDPR requirements; for example, we only collected data that was relevant to the study.

5 Results

5.1 Quantitative Results

5.1.1 Task Timings

Using the recordings of the video calls with participants, we measured the time to complete each of the three tasks. Figure 3 illustrates the distribution of task lengths broken down by task. Such data gives a sense of the learnability and efficiency of usage of the identity wallet.

For task one, the median completion time was 225 seconds (Inter-Quartile Range = 160), for task two: 121 seconds (IQR=49.5), and task three: 177 seconds (IQR=84).

5.1.2 User Journey Issues

Four classes of issue contributed to lowering the task completion efficiency: QR codes, security and privacy misunder-

¹<http://www.userinterviews.com>

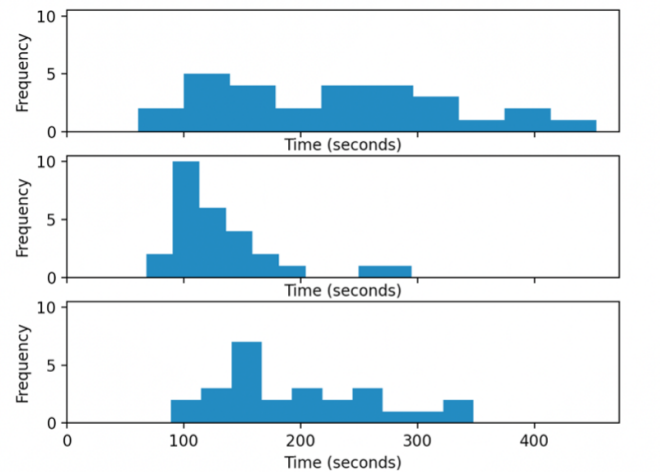


Figure 3: Distribution of task completion times recorded for (top) Task one; (middle) Task two; (bottom) Task three. Task one shows the greatest spread, partially due to the fact that this task requires interaction with QR codes and also user authentication to the IdP.

standings, device switching, and authentication.

QR codes contributed to the most significant proportion of user journey disruptions, some minor and some severe [53]. An example of low severity issues includes difficulty focussing the phone camera on the QR code; an example of a higher severity issue is when the user tries to scan the QR code in the native camera application on the mobile device rather than in the identity wallet app.

Misunderstanding relates to instances where the correct understanding of the identity wallet was not in place, which created a barrier to progress in the task. For example, confusion why the identifier of a newly received credential was different to the recently generated decentralized identifier (DID); the user perceives the credential ID and DID as alphanumeric passwords, and the user was concerned about the memorability of both; concern that a mistake had taken place since the user sent a credential issued by the IdP to the Rp; expectation that the IdP and Rp shared a database, so the authentication material for the IdP should be the same on the Rp website.

Other significant sources of hesitation and confusion included *device switching*, where participants were unsure whether to interact with the IdP, the Rp, or the identity wallet. The *authentication* category relates to issues that emerged from the entry of an alphanumeric password that was required to access the services of the IdP. The password was not a mnemonic and was seemingly randomly composed, which introduced some issues. Example issues relate to matching case sensitivity, copying and pasting, and locating symbols on keyboards with different language layouts.

Question	X=Decentralized Identifier			X=Verifiable Credential			X=Identity Proof		
	Yes	No	Unsure	Yes	No	Unsure	Yes	No	Unsure
X is secure and cannot be forged.	50%	11%	39%	59%	4%	37%	69%	8%	23%
X minimises the personal data that I need to share*	75%	4%	21%	70%	11%	19%	88%	8%	4%
X will be trusted by Alpaca Bank (IdP)	75%	0%	25%	-	-	-	-	-	-
X will be trusted by Bank of Carpathia (Rp)	-	-	-	81%	11%	15%	100%	0%	0%
I trust the X	64%	0%	36%	81%	0%	19%	85%	4%	12%
I need to keep X secret	79%	4%	18%	85%	7%	7%	73%	12%	15%
Alpaca Bank (IdP) can control X	-	-	-	33%	37%	30%	31%	54%	15%
Bank of Carpathia (Rp) can control X	11%	61%	29%	-	-	-	-	-	-
X has the features I require for my task	68%	0%	32%	85%	0%	15%	92%	0%	8%
I would be worried if I lost X	43%	50%	7%	48%	52%	0%	38%	54%	8%

Table 1: The table displays the results of administering the mental model scale to participants. We administered the set of questions associated with each component after the relevant task in the user study. We did not ask specific questions (denoted by a dash above) if a different formulation of the same question was more pertinent to discussing a particular decentralized identity component. (*) indicates the question is paraphrased for brevity. The full text of the question can be found in the appendices.

5.1.3 System Usability Scale (SUS)

We calculated the SUS for each participant using a widely accepted method [11]. The SUS data were normally distributed (Shapiro-Wilks test $p = 0.2$), and the identity wallet received an SUS score of $\mu = 71, \sigma = 16$. The overall SUS score is not a percentage and is graded on a curve. A score of around 71 implies that an identity wallet is ranked slightly above the 50th percentile. However, there is a relatively large standard deviation. A system where users are likely to be net promoters would receive an SUS score of at least 80 [11].

5.1.4 Mental Models

Table 1 illustrates results from our mental model scale. Regarding security perceptions, we learned that 50% had positive intuition about the security of the decentralized identifier (DID); the corresponding result was 59% for verifiable credentials and 69% for identity proofs. These numbers are not particularly high and reflect concerns that participants generally had about how this process could be more secure than processes involving existing paper-based methods.

One question we designed to test participants' understanding of the DID related to whether the relying party (Rp) could control it. The correct answer is no – since the user had not encountered the Rp at that point of the study – yet 11% responded yes, and 29% were not sure. Further questions about whether the IdP could control the verifiable credential or the identity proof were more challenging to answer and resulted in a split of yes, no, and unsure. In reality, Alpaca Bank (the IdP) could revoke the verifiable credential without the user's oversight. Therefore, the IdP has considerable power to control the verifiable credential's utility and the identity proof's verifiability.

At least 50% of participants expressed no concerns about losing access to their decentralized identifier, verifiable credential, or identity proof. The result reflects an expectation

that one of the parties in the scenario would correct the problem and re-establish user access to their data.

In terms of utility, of all three components, participants perceived the decentralized identifier (DID) to be the least helpful wallet component for the completion of their task (68% agreed with its utility). Participants were generally slow to appreciate the merits of DIDs compared to other aspects of the wallet technology.

5.2 Qualitative Results

The 30 participants generated 506 codes which led to four main themes: i) current challenges with identity (20.9% of codes, $n = 106$), e.g. participants highlighting oversharing of data and acknowledging improvements to identity processes; ii) assurances about the identity wallet service (35.6% of codes, $n = 180$), e.g. assuming the presence of trustworthy organizations and expressing concern over bad actors; iii) expectations of the identifier (12.1% of codes, $n = 61$), e.g. contributing to user confidence about the security; and iv) examining stakeholders and their roles (28.4% of codes, $n = 144$). Finally, some participants' responses did not adequately address the research questions and were consequently difficult to code. These were coded as 'other' (2.9% of codes, $n = 15$). We first summarize our findings concerning the current challenges users might face with identity. We then discuss participants' expectations of the identity wallet service and the identifier, followed by their perceptions of the stakeholders and roles. The results discussed in the first section will provide context to support the remaining results.

5.2.1 Challenges Users Have with Current Forms of Identity

We identified one theme from participants' statements relating to current forms of identity, that is, *Status quo is limited*,

convenient, and improving. This theme captured the challenges participants encountered with identity and possible improvements. Passports and driver's licenses were the dominant forms of identity referenced by participants, with driver's licenses dominant for participants from the USA, and a mix of both used by UK-based participants.

The majority of participants' concerns related to *oversharing* of data (20 participants), and we typically observed responses in one of two ways. First, a resignation to oversharing data and users thinking that there was little they could do to share less data or control what happened to their data, e.g. (P22): *"You know I work in information technology already and part of me says the idea that you keep your information secure and people not knowing it is a ship that has probably already sailed"*. Second, some participants drew comfort despite an apparent oversharing of their data for several reasons (where they provided information that was not necessary for a given process). For example, the feeling of comfort due to their having control over access to the identity document, legal structures which protect the use of their data, and their ability to define how the identity document should be used. Four of these participants expressed sentiments related to *both* comfort and resignation.

Other challenges experienced by participants related to the amount of *time* identity and related processes took. This was both online and in-person. Participants' statements referred to the inconvenience of *delays* which did not meet their expectations for more immediate service, e.g. (P18) referred to delays when they needed to replace a lost identity document in person: *"I mean down here in [LOCATION], the process is annoying because there's always a long line outside especially early in the morning to go back to get another license you'll be sitting out there for hours."* However, participants also experienced *convenience* and *ease of use*, as well as improvements to the identity process. In the latter case, the information they needed to provide to identify themselves was reduced, e.g. when opening (bank) accounts (P14): *"It's no longer, oh yeah, like I need a copy of your driver's license, proof of address and utility bill. Here's your account details and that's it. Oh my god that used to take like a week."*

When we asked participants to reflect on the security and privacy of their current forms of ID, they shared *perceptions of forgery*, whereby they were worried about losing their identity document, thought that their identity documents could easily be forged, but were also skeptical whether there was any value for a criminal to forge their ID documents. While current forms of ID received criticism for insecurity, this matter did not seem to be something participants had considered before in detail. Participants focused on *who* they were identifying themselves to so as to address concerns about misuse of the identification document e.g. (P13): *"... I'm only showing it to people who are like from an organization that is like nationally recognized"*, and the length of time the ID document was out of their possession, e.g. (P16): *"it's got my information on,*

the address and everything but it's only a quick look anyway. It's not like it's going to be in their possession quite some time, because then obviously I would question that as you've seen it why do you need to hold on to it."

Participants also made positive statements as they commented on *onboarding improvements* that they had experienced, for example, services using existing information, which the user perceives as minimising their effort (P21): *"There's a thing called government gateway and when you need to renew your passport or your driver's license they're almost interconnected. So I know when I first got my driver's license I didn't really have to do anything they had my information from my passport, so if they could maybe get their checks done through like a government site or a you know post office they also do like an identity service as well."* However, *life changing problems* occurred for participants where identity processes did not work well e.g. (P10): *"And so, when I bought my first house you know 10-12 years ago, they were not able to give me the keys after the closing. [I] had to wait a few days, I think... three or four days, because my name comes up in some kind of watch list or something."*

5.2.2 Assurances about the Identity Wallet Service

The majority of participants (22) questioned *who controls what* as they sought to understand the use of the identity wallet app and the three tasks involving the identifiers, credentials, and proofs. They emphasised their *personal agency* in these three tasks, rather than having the IdP or Rp in control, e.g. (P22): *"I would say no it's not Alpaca bank controlling, it's me controlling it. I mean they can offer to give me the credential but I am the person who is controlling it and allowing them to do so."* On the other hand, participants also thought that the *identity provider controlled the information shared with the Rp*. From their statements, it was clear that they did not perceive that the service was designed to empower them first in decisions on data sharing, e.g. (P23): *"I think Alpaca Bank are deciding what this Bank of Carpathia can know about me, so I would say they are in control because they're the ones that are divulging information to the second party involved. So I would think that they could potentially withdraw your social security number if that is what they chose to do."* Participants were in favour of a *separation of concerns* and did not expect the IdP and Rp to share information with each other, e.g. (P3): *"... I don't think that the Bank needs to have any idea that I'm doing something with a different bank. That's my private business, so I like that it kind of mentions that and I think that's important."* As such, we see this design feature meeting some, but not all of participants' expectations. Concerns about the empowerment of users to truly control their own data-sharing in the face of service demands has parallels to social media platforms, for instance Facebook [50], where Nadon et al. also noticed the potential for users to feel personal failure if they over-shared once given control. This also

contrasts with Farke et al.'s study of online activity data [33], wherein negative consequences of services holding data may not promote action, whereas here in the context of sharing identity information, comparable levels of transparency are evident but it raised questions about where the data was going and to who. Similar concerns have been surfaced when users are confronted with access of data on Google accounts by approved third-party services [4].

Foundations of trust were also highlighted in most of the participants' (23) statements. *Trustworthy entities* were highlighted with a focus on some organizations being better at such service provision than others, therefore they or their products being perceived as trustworthy, e.g. (P10): "I don't know if I trust my device as being as secure as like potentially you know, the bank's devices or network or their security is probably more enhanced than just my phone." This statement highlights the expectation that the user trusts their device and is willing to use it for identity, thereby flagging a critical issue for adoption if this assumption does not hold in practice.

Privacy and security evaluations were carried out with primarily *positive feelings* expressed about using the app for identity and how secure participants thought the process was, e.g. (P10): "Because it seems like, you know, it might be that if, even if I lost a proof, there might be some other part of it, that is needed to you know to complete any kind of functions with the banks."

The user interaction did not seem to match participants' mental models as they flagged what they perceived as *sharing violations* where the focus was on perceptions of either sharing too much information, sharing with unexpected recipients, or, notably, being asked to share what they perceived as too little information, e.g. (P7): "Where they're not asking for driver's license or social security number seems too convenient, because those are two usually two critical pieces of identity..." It would seem, in this case, that participants' expectations have shifted to match practice, and requests to share less information than they expect might be met with suspicion or mistrust. Other issues which did not match participants' mental models include the *language* used, and the novelty of the process since use of the identity wallet app was *unrelated to existing practices*.

Participants were *fearful of bad actors* (11 participants), thinking that use of identity wallet apps would open new avenues for attack (P11): "Oh man, I can see just a whole new breed of hackers. Oh God, as we speak they're breeding."

5.2.3 Expectations of the Identifier

This theme captured participants' (23) *expectations tied to the identifier*. Here, participants were confident about the security of the identity wallet app as a result of the randomness and uniqueness of the identifier, as well as its apparent entropy. This highlights ways in which users can perceive design features to convey assurances about security.

The identifier represented something participants thought only they should know and as a result should be kept *confidential*. From participants' statements, we understood that this expectation of confidentiality was linked to their use of the identifier to open a bank account. While this was a misconception akin to a folk model of security [73], participants notably suggested a behavior which was more rather than less secure. The perceived need to keep the identifier confidential was also linked to its similarity to a password, e.g. (P16): "You don't give out your password, so why would you go on sharing your unique code for your identifier..."

Participants were *confused* about the need to *generate their own identifier* and expressed concern that other users might not understand the procedure, e.g. (P6): "So I will say that I am used to things like these really long string of numbers and letters. But I think that would probably throw off the average user."

5.2.4 Examining Stakeholders and Their Roles

Participants (19) *weighed up the interaction*, assessing it based on the *efficiency, usefulness, and intuitiveness* of the identity wallet app. Issues emphasised included the time and effort participants expected use of the identity wallet app to take especially when considering they would interact with several relying parties (P21): "For me, it feels too time consuming. And the criteria is not the same for every bank or you know every place that you're looking to identify yourself so I'd rather do it on a case by case basis, rather than having something you know, an app on my phone." This then may create a kind of fatigue similar to the 'authentication fatigue', the perceived high effort required to access services to reach personal goals [62]. A knock-on effect here could be, as found with the Passfaces authentication technology [13], that the relative cost of a security technology compared to the primary task may be so high that users would delay important tasks and need more time to access services.

Fifteen participants stated that they found the identity wallet app easy to use, however, they also questioned the value of the app, e.g. (P21) "I mean I'm able to use it. Whether I want to use it is a different thing." Additionally, they did not find the processes intuitive or familiar, (P24): "I think it's still a bit... It's definitely different than a lot of other apps that are used, so there is a learning curve, especially for someone that, I think I'm pretty technologically competent and I think I would still have a bit of a difficulty with this here and there."

When queried about how they expected to recover from failure or loss, nineteen participants expressed *confidence in fallbacks*. They expected that *recovery was a basic feature and automated*. Notably, participants were concerned about the *location* of the backups e.g. (P25) "I wouldn't expect it to be on your phone. I don't think it's very safe to have everything just on your phone. I would expect it to be somewhere and just be able to get it back from a backup place" and

had different perspectives about *who* was responsible for the backup, e.g. the app provider (P13): *"So, because, like, I mean at the end of the day, all these I'm assuming that all these data points are feeding into IdentiCorp's like their, whatever their database for something like that. So I think it's the first. I think they are respon... like they are going to be the, the one that I should reach out to instead of other organizations."* Note that IdentiCorp was the name used in the study to refer to the wallet provider. Participants expected that *backups were kept* but they did not think that they were responsible for this. Additionally, they expected that recovery would be a *hassle*. Similar concerns to these have been raised about FIDO2 authentication technologies (e.g. [47]), wherein users also need to place a great deal of reliance on the service to make meaningful use of it.

While minimisation is a key feature of decentralized identity, eighteen participants had varying *interpretations of minimisation*. *Exercising control* over the information that was shared was a notable feature as they could choose not to share information if they deemed it unnecessary, e.g. (P20): *"I really did like the fact that I could choose like the optional ones and choose not to disclose [those] that weren't necessarily needed by the other bank."* At the same time, participants perceived that there was limited control as they would not be able to refuse to give information which was requested, e.g. (P5): *"If [Bank of] Carpathia wanted to know exactly how much I make a month, they would ask for it. And I wouldn't be able ... to say no."* As such, the environment supporting minimisation also needs to ensure users are not penalised for withholding information. Participants perceived minimisation to be *futile*, questioned *what was being minimised*, and perceived that it was their *effort rather than the data* that was being minimised.

6 Study Limitations

As with any research, our findings are subject to limitations. The pre-screen of participants may introduce bias into the results. For example, we specified that participants must be active users of Google Android. While Google Android is the dominant mobile operating system according to market share [66], we additionally specified that the personal device of the participant supported at least Android 10. Therefore, along with the fact that we recruited participants from *user-interviews.com* we can characterize our participant pool as Internet-savvy 'early adopters' of technology rather than representative of a perfectly random sample.

Identity wallets are a nascent technology, and we cannot be certain that today's design trends will be the same design trends years from now. However, we are confident that the user journeys we replicated capture the core functions of how identity wallets must behave, even if minor details of the user journey change over time.

We did not take steps to evaluate the generalizability of

our results. The quantitative results we report are descriptive statistics, and the qualitative results are innately not generalizable. However, our research method did surface challenges and problems captured using a trustworthy method, and resulting insights are transferable to other contexts with caution.

7 Discussion

7.1 Security Perceptions Enhanced by Trust

We learned that participants had mixed perceptions of the security of the decentralized identity system. We found that 50% felt that decentralized identifiers were secure (concerning forgery), 59% felt verifiable credentials were secure, and 69% felt this way with identity proofs. However, we captured mixed rationales underpinning this argument, highlighting the subjectivity inherent in answering the question. Specific security concerns voiced by participants include the risk of leaking personal information from the user interface of the mobile device and the risk of providing data to the wrong online service during enrolment.

Several threads fed into a positive perception of security – including the identity wallet collating accurate personal information and the acquisition of credentials from financial institutions, but also through misunderstandings of the technology. On the latter, we noted multiple assertions that the randomized alphanumeric composition of the Decentralized Identifier (DID) represents the secure encoding or encryption of personal information. Of course, this was not true, yet the perceived presence of cryptography provided a sense of confidence and comfort. This observation brings to mind ongoing debates around how to create security cues for users of encrypted communications apps (e.g. [67]). Due to the lack of widely used security cues in decentralized identity prototypes, our intuition is that participants derive most security comfort from the trust of the key actors in the study scenario rather than confidence in technical mechanisms.

7.2 User Control and Necessity of Fallbacks

The shared vision of decentralized identity intimates that the user can operate with autonomy from third parties and exercise control (in the purest sense) over the disclosure of verifiable personal attributes. Probing the understanding of user control in our research was particularly interesting concerning the verifiable credential. We observed an almost even 33% split between yes-no-don't know regarding whether the identity provider could control the verifiable credential. In simple cases, the identity proof is a signed wrapper containing verifiable credentials; however, more participants indicated they were in control of the proof than the verifiable credentials. These results show that ascribing control to actors and techniques in a decentralized identity scenario is challenging,

primarily due to the interrelationships between key technologies and the opacity and complexity of the infrastructure.

Participants reporting a positive sense of control over the identity wallet were generally not concerned about losing access to their identity wallet application or stored data. The lack of fear was primarily due to the view that one of the trusted parties in the scenario would be ready to restore access to lost data. Combined with our observation that it is challenging for users to articulate where control lies in the identity network, it presents a challenge in enabling users to pinpoint their locus of control and thus correctly identify risks to the continuity of access to services. There are parallels with the sense of ‘distributed responsibility as noted by Abdelaziz et al. [2] where users may not even know how to recover their identities without the help of a service provider.

7.3 Privacy Appraised More in the Value Exchange, and Less in the Identity Wallet

At the point of the study where the user sent an identity proof to the relying party, we captured the most lively discussion about data minimization. For example, despite leveraging advanced disclosure functions of the identity wallet (e.g. zero knowledge, or optional disclosures), participants predominantly evaluated privacy based on the value provided by the magnitude of the disclosure rather than on the technical tools at their disposal. We also noted that while participants may be surprised if they must disclose a seemingly long list of attributes, once a certain threshold of disclosures is approved, we noticed that participants could become numb to the discomfort of sharing additional attributes. A comparable phenomenon has been noted during e.g. reviewing Google ‘My Activity’ data collection information [33], where users have been seen to lack a sense of action due to the number of data disclosures they would need to process.

That end-users evaluate privacy in an overall transaction allows what we might expect from the privacy in context framework [55]. However, this suggests that innovation in user control cannot exist only in end-user technology and that suitable innovation in data collection practices must also come from service providers.

7.4 Identity Wallet Usability Not in a Vacuum

The identity wallet has inherent complexities that create usability challenges. The identity wallet resides at the intersection of the systems of three principal actors, the identity provider, the relying party, and finally, the wallet provider. Technology that resides at the boundaries between systems can create challenges for adhering to many of the heuristics for good usability [53]. Furthermore, the dominant terminology used in decentralized identity prototypes is esoteric: i.e. "decentralized identifiers", "verifiable credentials", and "identity proofs". While these terms initially create intrigue, they

ultimately form a barrier to the system’s learnability and limit users’ confidence to persevere and resolve problems. Previous endeavors in the identity domain have experimented with optimal names for system components. For example, Microsoft’s CardSpace [5] used the term ‘card’ to refer to a specific credential. Future research can seek to refine this terminology and find consensus between service providers to use terms consistently.

7.4.1 Design Considerations for Identity Wallets

Our research suggests practical ways to improve future (decentralized) identity wallets and related technologies.

- **Minimize reliance on QR codes.** QR codes were at the root of many user journey disruptions, and there may be efficiency gains in minimizing their usage. Rethinking the assumption that the user interacts with a laptop and mobile device simultaneously would open new avenues of interaction — e.g. mobile inter-app communication.
- **Provide meaningful errors in blockchain SDKs.** At times, we found it impossible to explain to participants why a cryptographic signature check might sporadically fail when expected to succeed (e.g. in an identity proof). We also had difficulty quickly understanding and explaining blockchain-specific errors.
- **Deploy Decentralized Identifiers (DIDs) only if essential.** DIDs are more complex than traditional email-based usernames. If DIDs replace traditional usernames, there will be undesirable scenarios where users must type a DID string, or where a service cannot authenticate a DID due to system problems elsewhere. Therefore designers must deploy DIDs only if essential to a use case.

8 Conclusion

There is a growing expectation that political and technical initiatives towards digital identity will gather pace in the foreseeable future. However, user perspectives have not been a driving force in shaping those ongoing initiatives. The findings of this study point to the dominance of paper/card-based identity methods for online identity verification and a large gap between identity verification today and what it might be in the foreseeable future. Our results suggest that technical narratives might not be a compelling driving force for future uptake and that, as previous work in identity management has highlighted [20], the user proposition should receive further thought. What seems most salient to drive adoption is the existence of supporting (infra)structures, the appeal of the list of available verifiers, and the low complexity of using a new identity wallet tool. Future work might evaluate identity wallet apps in the wild to identify opportunities to close these gaps between technical idealism and everyday reality.

References

- [1] OAuth 2.0. <https://oauth.net/2/>, Last accessed on 6th Jan 2022.
- [2] Yomna Abdelaziz, Daniela Napoli, and Sonia Chiasson. End-users and service providers: trust and distributed responsibility for account security. In *2019 17th International Conference on Privacy, Security and Trust (PST)*, pages 1–6. IEEE, 2019.
- [3] Christopher Allen. The Path to Self-Sovereign Identity, 2016. <http://www.lifewithalacrity.com/2016/04/the-path-to-self-sovereign-identity.html>, Last accessed on 6th Jan 2022.
- [4] David G Balash, Xiaoyuan Wu, Miles Grant, Irwin Reyes, and Adam J Aviv. Security and Privacy Perceptions of Third-Party Application Access for Google Accounts. *USENIX Security '22*, 2022.
- [5] Vittorio Bertocci, Garrett Serack, and Caleb Baker. *Understanding Windows CardSpace: An Introduction to the Concepts and Challenges of Digital Identities*. Addison Wesley, 2007.
- [6] Patrick Beuth. Verantwortungslos und gefährlich, 2021.
- [7] Abhilasha Bhargav-Spantzely, Jan Camenisch, Thomas Gross, and Dieter Sommer. User centricity: a taxonomy and open issues. In *Proceedings of the Second ACM Workshop on Digital Identity Management - DIM '06*, page 1, New York, New York, USA, 2006. ACM Press.
- [8] Luís T. A. N. Brandão, Nicolas Christin, George Danezis, and Anonymous. Toward Mending Two Nation-Scale Brokered Identification Systems , no.2, 2015, pp.135-155. In *Proceedings on Privacy Enhancing Technologies (PETS)*, pages 135–155, 2015.
- [9] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
- [10] Deanna G Brockman, Lia Petronio, Jacqueline S Dron, Bum Chul Kwon, Trish Vosburg, Lisa Nip, Andrew Tang, Mary O'Reilly, Niall Lennon, Bang Wong, et al. Design and user experience testing of a polygenic score report: a qualitative study of prospective users. *BMC Medical Genomics*, 14(1):1–20, 2021.
- [11] John Brooke. SUS: A retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.
- [12] Sacha Brostoff, Charlene Jennett, Miguel Malheiros, and M. Angela Sasse. Federated identity to access e-government services - Are citizens ready for this? In *Proceedings of the 2013 ACM Workshop on Digital Identity Management*, pages 97–107, New York, New York, USA, 2013. ACM Press.
- [13] Sacha Brostoff and M Angela Sasse. Are passfaces more usable than passwords? A field trial investigation. In *People and computers XIV—usability or else!*, pages 405–424. Springer, 2000.
- [14] Kim Cameron. The Laws of Identity. Technical report, Microsoft, 2005.
- [15] Kim Cameron and Michael B. Jones. Design Rationale behind the Identity Metasystem Architecture. In *ISSE/SECURE 2007 Securing Electronic Business Processes*, pages 117–129. Vieweg, Wiesbaden, 2007.
- [16] Stéphane Ciolino, Simon Parkin, and Paul Dunphy. Of Two Minds about Two-Factor: Understanding Everyday FIDO U2F Usability through Device Comparison and Experience Sampling. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security, SOUPS'19*, pages 339–356, USA, 2019. USENIX Association.
- [17] Congressman Bill Foster. US Congressmen reintroduce sweeping digital ID bill, 2021. <https://foster.house.gov/media/in-the-news/us-congressmen-reintroduce-sweeping-digital-id-bill>, Last accessed on 6th Jan 2022.
- [18] Kevin Corre, Olivier Barais, Gerson Sunyé, Vincent Frey, and Jean-Michel Crom. Why Can't Users Choose Their Identity Providers On The Web? In *Proceedings on Privacy Enhancing Technologies*, pages 75–89, Minneapolis, Aug 2017.
- [19] Albese Demjaha, Jonathan M Spring, Ingolf Becker, Simon Parkin, and M Angela Sasse. Metaphors considered harmful? An exploratory study of the effectiveness of functional metaphors for end-to-end encryption. In *Proc. USEC*, volume 2018. Internet Society, 2018.
- [20] Rachna Dhamija and Lisa Dusseault. The Seven Flaws of Identity Management: Usability and Security Challenges. *IEEE Security and Privacy*, 6(2):24–29, mar 2008.
- [21] DIACC. Pan-Canadian Trust Framework, 2021. <https://diacc.ca/trust-framework/>, Last accessed on 6th Jan 2022.
- [22] Die Deutsche Kreditwirtschaft. DK begrüßt Experimentierklausel zur Kundenidentifizierung, mit der Banken und Sparkassen innovative digitale Initiativen erproben werden, 2021.
- [23] Andrea diSessa. Models of computation. *User centered system design*, pages 201–218, 1986.

- [24] David Dittrich and Erin Kenneally. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, 2012.
- [25] Paul Dunphy, Luke Garratt, and Fabien Petitcolas. Decentralizing Digital Identity: Open Challenges for Distributed Ledgers. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 75–78, 2018.
- [26] Paul Dunphy and Fabien A.P. Petitcolas. A First Look at Identity Management Schemes on the Blockchain. *IEEE Security & Privacy*, 16(4):20–29, Jul 2018.
- [27] Serge Egelman. My profile is my password, verify me! The privacy/convenience tradeoff of Facebook Connect. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 2369–2378, New York, NY, USA, apr 2013. ACM.
- [28] N. Sakimura et al. E. Garber, M. Haine, V. Knobloch, G. Liebbrandt, T. Lodderstedt, D. Lycklama. Gain Digital Trust: How Financial Institutions are taking a leadership role in the Digital Economy by establishing a Global Assured Network. In *European Identity and Cloud Conference*,, Munich, 2021.
- [29] European Commission. Commission proposes a trusted and secure Digital Identity, 2021. https://ec.europa.eu/commission/presscorner/detail/en/IP_21_2663, Last accessed on 6th Jan 2022.
- [30] Evernym. Connect.me. <https://www.connect.me/>, Last accessed on 6th Jan 2022.
- [31] Facebook. What names are allowed on Facebook?, 2020. <https://www.facebook.com/help/112146705538576>, Last accessed on 6th Jan 2022.
- [32] Sascha Fahl, Marian Harbach, Thomas Muders, Matthew Smith, and Uwe Sander. Helping Johnny 2.0 to encrypt his Facebook conversations. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pages 1–17, 2012.
- [33] Florian M Farke, David G Balash, Maximilian Golla, Markus Dürmuth, and Adam J Aviv. Are privacy dashboards good for end users? Evaluating user perceptions and reactions to Google’s My Activity. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 483–500, 2021.
- [34] U. Feige, A. Fiat, and A. Shamir. Zero-knowledge proofs of identity. *Journal of cryptology*, 1(2):77–94, 1988.
- [35] FIDO Alliance. The FIDO Alliance, 2020. <https://fidoalliance.org>, Last accessed on 6th Jan 2022.
- [36] Financial Conduct Authority. FCA fines Deutsche Bank £163 million for serious anti-money laundering controls failings, 2017.
- [37] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 268–285. IEEE, May 2020.
- [38] Cal Halvorsen and Sylvia Brown. In their own words: Small-and mid-level donors express their views on charitable giving. *SSRN 3916288*, 2021.
- [39] Hyperledger Foundation. Hyperledger Indy SDK, 2019. <https://github.com/hyperledger/indy-sdk>, Last accessed on 6th Jan 2022.
- [40] IDUnion. IDUnion, 2021. <https://idunion.org/>, Last accessed on 6th Jan 2022.
- [41] Kat Krol and Sören Preibusch. Control versus effort in privacy warnings for webforms. In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, pages 13–23, 2016.
- [42] Susan Landau and Tyler Moore. Economic Tussles in Federated Identity Management. *First Monday*, 17(10), oct 2012.
- [43] Leona Lassak, Annika Hildebrandt, Maximilian Golla, and Blase Ur. "It’s stored, hopefully, on an encrypted server": Mitigating users’ misconceptions about FIDO2 biometric WebAuthn. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 91–108, 2021.
- [44] Let’s Encrypt. Let’s Encrypt. <https://letsencrypt.org/>, Last accessed on 6th Jan 2022.
- [45] LISSI. The new solution for identities: Digital. Decentralized and Self-sovereign., 2020. <https://lissi.id/>, Last accessed on 6th Jan 2022.
- [46] Christian Lundkvist, Rouven Heck, Joel Torstensson, Zac Mitton, and Michael Sena. uPort: A Platform for Self-Sovereign Identity. Technical report, 2017.
- [47] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *IEEE Symposium on Security and Privacy*, pages 268–285, 2020.

- [48] Deepak Maram, Harjasleen Malvai, Fan Zhang, Nerla Jean-Louis, Alexander Frolov, Tyler Kell, Tyrone Lobban, Christine Moy, Ari Juels, and Andrew Miller. Can-DID: Can-do decentralized identity with legacy compatibility, sybil-resistance, and accountability. *Proceedings - IEEE Symposium on Security and Privacy*, 2021-May:1348–1366, May 2021.
- [49] Florian Mathis, Kami Vaniea, and Mohamed Khamis. Prototyping usable privacy and security systems: Insights from experts. *International Journal of Human-Computer Interaction*, 38(5):468–490, 2022.
- [50] Guillaume Nadon, Marcus Feilberg, Mathias Johansen, and Irina Shklovski. In the user we trust: Unrealistic expectations of facebook’s privacy mechanisms. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 138–149, 2018.
- [51] Satoshi Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System. Technical report, 2008.
- [52] Lily Hay Newman. Microsoft’s Dream of Decentralized IDs Enters the Real World, 2021. <https://www.wired.com/story/microsoft-decentralized-id-blockchain/>, Last accessed on 6th Jan 2022.
- [53] Jakob Nielsen. 10 Usability Heuristics for User Interface Design, 1994.
- [54] Jakob Nielsen. Usability inspection methods. In *Conference Companion on Human Factors in Computing Systems*, CHI ’94, page 413–414. Association for Computing Machinery, 1994.
- [55] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books, 2009.
- [56] Darrell O’Donnell. The Current and Future State of Digital Wallets. Technical report, Continuum Loop Inc, 2019.
- [57] Onfido. Onfido Identity Fraud Report 2022. Technical report, Onfido, 2022.
- [58] Ping Identity. ShoCard | Identity for a Mobile World. <https://www.shocard.com/en.html>, Last accessed on 6th Jan 2022.
- [59] Alex Preukschat and Drummond Reed. *Self-Sovereign Identity*. Manning Publications Co, 2021.
- [60] Lee Rainie. Americans’ complicated feelings about social media in an era of privacy concerns. Technical report, Pew Research Center, 2018.
- [61] Y Rogers, H Sharp, and J Preece. *Interaction design: Beyond human-computer interaction*. John Wiley and Sons, 2 edition, 2011.
- [62] M Angela Sasse, Michelle Steves, Kat Krol, and Dana Chisnell. The great authentication fatigue—and how to overcome it. In *International Conference on Cross-Cultural Design*, pages 228–239. Springer, 2014.
- [63] M.A. Sasse, Sacha Brostoff, and D Weirich. Transforming the ’Weakest Link’ – a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, 19(3):122–131, Jul 2001.
- [64] SelfKey. SelfKey. <https://selfkey.org/>, Last accessed on 6th Jan 2022.
- [65] Manu Sporny, Dave Longley, Markus Sabadello, Drummond Reed, Ori Steele, and Christopher Allen. Decentralized Identifiers (DIDs) v1.0. Technical report, W3C, 2021.
- [66] Statcounter. Mobile Operating System Market Share Worldwide: Jan 2021 - Jan 2022, 2022. <https://gs.statcounter.com/os-market-share/mobile/worldwide>, Last accessed on 6th Jan 2022.
- [67] Christian Stransky, Dominik Wermke, Johanna Schrader, Nicolas Huaman, Yasemin Acar, Anna Lena Fehlhaber, Miranda Wei, Blase Ur, and Sascha Fahl. On the limited impact of visualizing encryption: Perceptions of E2E messaging security. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 437–454, 2021.
- [68] San Tsai Sun, Eric Pospisil, Ildar Muslukhov, Nuray Dindar, Kirstie Hawkey, and Konstantin Beznosov. What makes users refuse web single sign-on?: An empirical investigation of OpenID. In *SOUPS 2011 - Proceedings of the 7th Symposium on Usable Privacy and Security*, page 1, New York, New York, USA, 2011. ACM Press.
- [69] UK Government. UK digital identity & attributes trust framework: Alpha version 2. Technical report, UK Government, 2021.
- [70] U.S. Department of Justice - Office of the Inspector General - Audit Division. The Department of Justice’s Efforts to Combat Identity Theft. Technical report, 2010.
- [71] Jennifer van Grove. Each Month 250 Million People Use Facebook Connect on the Web, 2010.
- [72] W3C. Verifiable Credentials Data Model 1.1, 2021.
- [73] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS ’10, New York, NY, USA, 2010. Association for Computing Machinery.

- [74] Rick Wash and Emilee Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325, 2015.
- [75] Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. Technical report, 2014.
- [76] World Economic Forum. A Blueprint for Digital Identity - The Role of Financial Institutions in Building Digital Identity. aug 2016.

A Information Sheet

What is the purpose of the study?

We are inviting you to take part in a research study to help us investigate and understand users’ experience of a digital identity wallet app. The app gives users a way to identify themselves (that is, who they are) in order to access various services over the Internet.

Why have I been approached?

We need to recruit several adult participants to take part in the study. You have been approached in an effort to recruit people who have a Google Android phone (running at least Android 10), some experience with Internet banking and banking apps, and who have installed (or can install) the Zoom application on a laptop/computer and a mobile phone.

Do I have to take part?

Participation is entirely voluntary. If you change your mind about taking part in the research study, you can withdraw at any point during the study.

What happens during the study?

The research study will take place remotely over Zoom. You will be asked to install the digital identity wallet app on your Google Android phone. You will join the Zoom call both from your computer and your mobile phone and will share the screen on the mobile phone so we can see how you use the app. The digital identity wallet app will not collect any information from your phone, and at the end of the study, you can uninstall it. You will be given three tasks to carry out using the digital identity wallet app and you will be asked to share your thoughts of the experience as you carry out the tasks. After you complete the tasks, you will be asked to fill in a questionnaire about the app and your experience and to answer a few questions about the process. The study will last approximately 45 minutes. The research study will be audio- and video-recorded for review and analysis in order to gain insights into users’ experience using the digital identity wallet app. No identifying information will be shared outside the research team.

What are the possible disadvantages and risks of taking part?

We do not anticipate any disadvantages or risks associated with participation in this study.

What are the possible benefits of taking part?

While individual benefits may be limited, your participation will help us to build an understanding of users’ experience of digital identity wallet apps. It is hoped that the results of this research study will contribute to the development of such apps in the future.

Who is organising and funding the research?

The research is conducted by Maina Korir and Dr. Paul Dunphy from the OneSpan Innovation Centre.

Will my participation be confidential?

Yes. We will not share personally identifying information outside of the research team.

What happens if something goes wrong?

In the unlikely case of concern or complaint, please contact Dr. Paul Dunphy, Principal Research Scientist at the OneSpan Innovation Center (paul.dunphy@onespan.com).

Where can I get more information?

If you would like more information, please contact the researcher: Maina Korir (maina.korir@onespan.com).

Data protection

All collected data will be de-identified soon after the research study and before the data is analysed. Participants will be given a pseudonym to refer to their data during the data analysis process meaning it will not be possible to link this data back to any of the participants.

B Consent Form

Participants could indicate yes or no in response to the following:

- I have read and understood the information sheet and have had the opportunity to ask questions about the study.
- I consent voluntarily to be a participant in this study and understand that I can withdraw from the study at any time if I so choose.
- I understand that taking part in the study involves joining a Zoom call, installing and using a mobile app, and taking part in an audio- and video-recorded interview to discuss my experience of using the app.

- I agree to the interview being audio- and video-recorded and to the interview being transcribed and personal identifiers removed.
- I understand that information I provide, which cannot identify me, may be published in journals, conference proceedings and reports.
- I understand that personal information collected about me that can identify me, such as my name will not be shared beyond the research team.
- I understand that my data will be stripped of personal identifiers during the transcription process. I understand that data that cannot identify me will be encrypted and stored for the duration of the project.
- After the data has been stripped of all personal identifiers and has been anonymized I agree that the information I provide during the interviews can be quoted in research outputs.

C Interview Script

We asked participants the following questions, touching on the issues identified in the different categories:

Current Forms of Identity and Identification

- For the purposes of this interview, what name can I use to refer to you?
- *Name selected*, if I asked you to prove that you are *name*, what would you do?
- Have you been in a situation where you've been asked to prove that you are *name*?
- Could you tell me about the experience?
- How do you feel about using a *item indicated by participant* as a means of ID?
- Are there ways that a *participant's ID document* is a secure form of ID?
- Are there ways that a *participant's ID document* as a form of ID offers you privacy?
- How do you feel about an opportunity to decide what piece or pieces of information to share to identify yourself?
- If you could change one thing about the process of identifying yourself online, which one would you pick?
- Why would you choose to focus on that?

Interactions with the Identity Wallet

Scenario

Imagine that you are Alex. Alex is a customer at Alpaca Bank. Alex opened a bank account in person at the bank branch closest to where they live. Alex now wants to open another account with a second bank - Bank of Carpathia. You will carry out three tasks to achieve this goal using the digital identity wallet app. The building blocks of a new privacy-respectful identity wallet app are: identifiers which you will interact with in the first task, credentials, which you will interact with in the second task, and proofs, which you will interact with in the third task. You will carry out each task in turn and I will ask you a few questions about the experience between each task.

Reflection on Identifiers and Proofs

Participants were given instructions to carry out the steps for the three tasks: identifiers, credentials, and proofs. They then answered the following questions:

- MyIdentifier is secure, that is, it cannot be forged
- My Identifier will minimise the information about me that I have to share to identify myself
- My Identifier will be trusted by Alpaca Bank
- I trust My Identifier
- I need to keep My Identifier secret
- Bank of Carpathia can control My Identifier
- My Identifier has the features I require for my tasks
- I would be worried if I lost My Identifier

We replaced 'MyIdentifier' with 'verifiable credential' and 'proof' for the second and third tasks.

Usability and User Expectations

- Based on your experience using the digital identity wallet app today, what would you say is the best thing about the app?
- What would you say are the limitations of the app?
- Are there any needs or challenges you have faced with identity that the digital identity wallet app addresses?
- Are there any needs or challenges you have faced with identity that the digital identity wallet app does not address?
- In what ways do you see the digital identity wallet app fitting into your regular practice of identifying yourself?

Usability and Security of Trusted Platform Module (TPM) Library APIs

Siddharth Prakash Rao
Nokia Bell Labs, Finland

Gabriela Limonta
Nokia Bell Labs, Finland

Janne Lindqvist
Aalto University, Finland

Abstract

Trusted Platform Modules (TPMs) provide a hardware-based root of trust and secure storage and help verify their host's integrity. Software developers can interact with a TPM and utilize its functionalities using standardized APIs that various libraries have implemented. We present a qualitative study (n=9) involving task analysis and cognitive interviews that uncovered several usability and security issues with `tpm2-tools`, one of the widely used TPM library APIs. Towards this end, we implemented a study environment that we will release as open source to support further studies.

Our results support two major conclusions: 1) `tpm2-tools` APIs, as designed, are not designed to be developer-friendly, and 2) One of the major causes for these usability issues is in the TPM specifications. Since other libraries also mirror the specifications and provide no significant usability improvements, our results are likely to indicate similar issues with all current TPM library APIs. We provide recommendations for improving the TPM library APIs documentation and software, and we highlight the need for HCI experts to review TPM specifications to preemptively address usability pitfalls.

1 Introduction

A Trusted Platform Module (TPM) [51] is a tamper-resistant chip that is used as a hardware-based root of trust in many modern applications [34, 61]. TPMs can carry out common cryptographic operations, such as secure key generation, encryption, hashing, and signing. Furthermore, since the TPM is physically isolated from the processing system of its host,

it can be used for securely storing a small amount of sensitive data (e.g., keys and certificates), which can further be utilized for verifying the integrity of its host. TPMs also provide various non-cryptographic security features for imposing access control restrictions on the objects created or stored in the TPM. Such restrictions play a crucial role in hardening the security of applications built using TPMs. The Trusted Computing Group (TCG) defines standard specifications that cover TPM architecture and implementation [56], and several high-level Application Programming Interfaces (APIs) to interact with the TPM hardware [52, 54, 55]. The latter is implemented by various software libraries, and is the scope of our study.

APIs play a crucial role in modern software development because they provide reusable components for developers to build applications efficiently and in less time. Nevertheless, APIs tend to be complex, and making them usable (or developer-friendly) has been an ongoing research theme. Previous works have analyzed various APIs that offer security using cryptographic features to understand and improve their usability [13, 40]. In this work, we extend this research theme for TPM library APIs, which offer additional non-cryptographic features (e.g., access control). We believe that the combination of these security features adds more to the complexity of TPM concepts and hinders the APIs' usability and security. Our work explores them by systematically analyzing the implementation of `tpm2-tools`, a widely used TPM library API with 85830 downloads (refer to Appendix A).

Our main goals are to understand the usability and security pitfalls of TPM developers and review the current API implementation to provide concrete design guidelines for usable secure API development. In this realm, we conduct a qualitative study with TPM developers using mixed methods (task analysis and cognitive interviews). We identify common use cases of TPM, conduct a thorough review of the available APIs, literature survey of the prior art, and combine them to design tasks and questionnaires for our participants. We also involve the participants in a follow-up interview to understand their experiences, perceptions, and opinions about the APIs. We conduct thematic analysis and code analysis to identify

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

themes and common coding patterns that give an overview of the usability and security pitfalls of `tpm2-tools` library APIs. Based on these results, we provide concrete recommendations for the library documentation and software.

Contributions: First, we built a study environment that supports all major TPM libraries and works right out of a browser [6]. It is the only available platform for studying TPM-related tasks to our best knowledge.

Second, we show that the `tpm2-tools` APIs are not user-friendly based on a systematic study involving an analysis of developers' inputs collected through tasks and interviews. Although various guidelines are available to design usable secure APIs, we find that the `tpm2-tools` library have not seriously considered or implemented them. Consequently, developers struggle to use the APIs efficiently and are prone to make trivial mistakes that undermine security. The complexity of the topics and lack of developer-friendly APIs and supporting materials could pose a major barrier for developers to fully utilize TPM's capabilities. Our work identifies and highlights various usability pitfalls that impact security and provides concrete recommendations to address them.

Third, we highlight that standard specifications are a potential venue to influence the usability of technologies. We observed that the TPM API implementation strictly follows the standards and found many instances where the pitfalls could be traced back to them. Thus, we believe that there is a need for HCI experts to be involved in the design and review of standards to preempt any possible usability pitfalls that otherwise would be propagated to the software implementation.

2 Background

This section covers the background of TPM security features. We explain the following components: cryptographic operations, hierarchies, TPM-specific restrictions, platform configuration registers, authorizations, and sessions. The TPM uses these components to provide security-related functions, such as key generation, the hardware-based root of trust, device identity, remote attestation, and secure storage.

The TPM supports common cryptographic operations such as **encryption**, **signing**, and **hashing**. Additionally, it provides secure key generation functionalities, where the TPM-generated keys can be used internally for cryptographic operations, or they can be exported for external applications. The keys that reside within the TPM are protected by a logical abstraction (i.e., collection of objects) called a **hierarchy**. The TPM provides four hierarchies: *owner*, *endorsement*, *platform*, and *NULL*. The first three are meant to be used by the TPM's owner, manufacturer, and host platform. The *NULL* hierarchy is reserved for short-lived objects that are expected to be lost on reboot. In addition to keys, a hierarchy can include another kind of object: a sealed data blob, a structure for protecting small amounts of user data in the TPM.

Each hierarchy is associated with a random seed, which is used to generate primary keys that can serve as the root in a tree of other (child) objects inside the hierarchy. Primary keys are stored inside the TPM and cannot be exported or read externally, whereas child objects (such as sealed data blobs and non-primary keys) are generated on the TPM but stored in the disk until loaded back for actual use. Since the child objects leave the trust boundary of the TPM when exported, they can be misused. In such cases, the child's security is guaranteed by *wrapping*, a mechanism where the child's sensitive part is encrypted by its parent key and can only be decrypted upon loading into the TPM.

The TPM gives users control over the usage of TPM-generated keys by allowing them to impose different **restrictions** over the key's *purpose*, *duplication*, and *usage*. The *purpose restriction* implies restricting the key for encryption/signing or only for decryption by setting the attributes `sign` and `decrypt`, respectively. The term *duplication* in the TPM context refers to the possibility of a key having more than one parent. The *duplication restriction* includes setting the `fixedParent` attribute to allow the key's parent to change within the same TPM or `fixedTPM` attribute to allow the key duplication for using it on a different TPM. Finally, the *usage restriction* refers to restricting the key, by setting the `restricted` attribute, to sign/decrypt only TPM-internal data. Please note that these attributes can only be set during the key generation; hence, the restrictions cannot be updated later.

TPM manufacturers utilize and predefine the key restrictions to provide two special keys: the endorsement key (EK) and the attestation key (AK). The EK is a restricted encryption primary key generated from the endorsement hierarchy's seed; it is certified by the TPM manufacturer as proof of its authenticity. The AK is a restricted signing key protected by the EK as its parent. The TPM can use the AK to prove its unique identity, e.g., during remote attestation. It guarantees that a specific, legitimate TPM has produced the signed message.

TPMs provide **Platform Configuration Registers (PCRs)** as an option for secure storage. PCRs are a set of 24 registers used to store hashes of different system components (e.g., firmware, kernel, hypervisor, operating system and files in the filesystem) usually during the boot sequence. PCRs are considered secure storage because a user cannot directly write hash values into the PCRs. Instead, they can provide the hash, and the TPM will extend one of the PCRs by concatenating it with the preexisting PCR content. The concatenated result is fed to the hashing algorithm to compute the digest, which is then stored in the PCR as the new value (referred to as *measurements*) to represent the state of a TPM's host platform at that given time. The hashing algorithm used determines the size of the measurement. The TPM supports multiple hashing algorithms and the group of PCRs associated with the same algorithm are referred to as a *PCR bank*.

PCR measurements can be used in remote attestation to verify that the integrity of a device has not been tampered

with. TPM uses *quotes* – the measurements and other information (e.g., clock and number of reboots and suspends)– and signs it with a restricted key (e.g., AK). This process of obtaining a TPM-equivalent of message authentication code is called *quoting*. A verifier uses the quotes to detect tampering by checking the contents against reference known values. Furthermore, since an AK signs the quote, it guarantees the verifier that the measurements are trustworthy and generated by a legitimate TPM. If tampering (e.g., changes to firmware or kernel) is detected, the verifier can also identify the exact component that is tampered with because each PCR contains a measurement that represents a specific host component.

The TPM provides a limited amount of Non-Volatile Random Access Memory (NVRAM) that can be used to store persistent data. Users may use NVRAM for secure storage because access to NVRAM can be restricted with TPM security mechanisms such as *sealing*. When an NVRAM area is sealed against a particular state of the host, its content can only be read if the state is unaltered.

Similar to sealing, TPM provides other types of restrictions for TPM objects (e.g., keys, sealed data blobs, or NVRAM areas). For example, a user can create policies to define restrictions on how an object can be used. Furthermore, it can also impose **restrictions against TPM-internal and TPM-external states**. The TPM-internal states include the PCR state, NVRAM contents, and TPM counters, and the TPM-external states may be passwords, state of external hardware (e.g., biometrics or GPS information), and signatures from smart cards. These restrictions are set when an object is created, and they will be checked before the object is used.

A user may need to use **session-based authorizations** to comply with the restrictions imposed on objects in the TPM or to authorize commands. Sessions are a way to communicate authorizations to the TPM since they carry the information needed to prove that the user can perform the intended action. The authorization contained in a session may be reused to execute several commands repeatedly, as sessions preserve the state between commands. For example, HMAC sessions may be used to communicate a password more securely, as the password only needs to be specified once during the creation time of an object (e.g., a key). In such sessions, the password is used as an input to calculate the HMAC of each command and response from the TPM, which allows authorizing an action without actually sending the password.

3 Related work

TPMs have been used in a wide range of security-critical applications [26, 27, 37, 38, 48]. Security flaws in the TPM could undermine the security of the applications that use it, and finding such flaws is attractive to researchers. Prior research works have analyzed the TPM architecture and specifications to verify the security guarantees of TPM [22–24, 49, 62]. These existing studies mainly on formal methods for analyzing TPM

specifications. To our best knowledge, analysis of TPM software libraries and their API implementation has not been studied. In our study, we explore this topic by involving the TPM developers to understand how they use the APIs and investigate how the API implementation could pose security and usability barriers.

We build upon the prior research that identified various human factors of secure application development. In particular, we draw insights from the studies that evaluated cryptographic libraries [13, 17, 28, 35, 40, 44] due to an overlap of cryptographic features offered by TPM library APIs. These studies have found that lack of usable APIs and documentation are major barriers to developers. Our study evaluates whether these issues are replicated in the TPM ecosystem.

Similarly, previous studies have identified that developer’s background [19, 43, 63], information sources [14–16], and workplaces [18, 32] are some of the other factors that affect secure application development. We are particularly interested in the information sources referred to by TPM developers as supporting materials. They may have to refer to a broader set of sources, e.g., library documentation, TPM specifications, and other security guidelines, and their effect on the way developers use TPM library APIs is yet unknown. In our study, we explore the role of such sources in TPM development.

4 Research Methodology

We first created a pool of potential participants and sent them a preliminary survey (refer to Appendix B) inviting them to participate in the study. This survey helped us filter suitable participants and identify the common use cases for designing the tasks. In parallel, we conducted an independent analysis of the TPM ecosystem to scrutinize the intrinsic details of concepts, software libraries, and supporting resources (e.g., standards and documentation) available for the developers. This helped us define evaluation criteria for the tasks. Similarly, we conducted a literature review of previous usable security research works (e.g., [13, 47]), that served as a reference when preparing our questionnaires. Next, we designed TPM-related tasks and questionnaires to collect practical and conceptual barriers (mental models) while coding with TPM libraries. We built our study environment (refer to §4.1) using open source modules and hosted it on our servers. After participants completed the tasks, we conducted cognitive interviews to understand their experiences while dealing with TPMs (refer to §4.3). During the interview, we allowed them to introspect about their experience with the TPM ecosystem and provide suggestions for improvement. Finally, we conducted a two-stage analysis (refer to §4.4) of the data we gathered from the study environment and the interviews.

Participant demography. We targeted experienced developers with TPM experience such that we could design realistic tasks leveraging TPM-specific security features. We aimed to

evaluate usability and security pitfalls closer to real-life TPM usage, as opposed to struggles with its steep learning curve. We used purposive sampling (refer to Appendix C for more details). Our results are based on the responses from the 9 participants who matched the criteria for our target population and went through the entire study protocol. We offered them a compensation worth 100 €. Participants were male (aged 18–59) with a security background, <1–5 years of TPM, >=2 years of coding experience and a bachelor’s degree.

Ethical and privacy considerations: We followed best practice guidelines throughout our study [47]. Our institution’s review team approved our study and confirmed that it meets the ethics and privacy standards. After a careful review of our methodology, the review team drafted a GDPR-compliant *privacy policy* and *participation consent form*. We presented both of them to the participants before the study began. Accordingly, the collected personally identifiable information (e.g., email and names) was solely used for contacting the participants. We excluded such data from the analysis and discarded it immediately by the end of the study. Furthermore, we used open-source components to build our study environment, hosted it on our servers, and ensured that the participants’ data was entirely under our control.

4.1 Task and questionnaire design

Tasks. We identified four common use cases of TPM from our preliminary survey as follows: *encryption* (symmetric and asymmetric), *storing measurements* of files on PCRs, *securing secrets* on the TPM, and *remote attestation*. We also identified various cryptographic and non-cryptographic security features. We designed simple tasks around the use cases and added conditions with a combination of security features to evaluate functional correctness and the participant’s security choices. Such conditions allowed us to understand whether a knowledgeable developer (from a coding and security point of view) would be able to choose suitable cryptographic parameters and impose TPM-based security restrictions while dealing with common use cases. The tasks required the participant to use well-known TPM commands. Each task was divided into simple steps (refer to Appendix D) to ensure better understanding and obtain higher completion rates. We ran a pilot study to evaluate this. We improved the text and adjusted the tasks’ complexity based on the feedback.

Table 1 summarizes the mapping of use cases and security features of our tasks. We assigned four tasks to the participants and did not impose time restrictions for completion. As the first task, five of them got asymmetric encryption and the rest got symmetric encryption. The remaining three tasks were common to all. We assigned a specific library to each participant based on their preliminary survey responses: one was assigned IBMTSS and the rest `tpm2-tools`. However, switching libraries was allowed, and the participant whom we

assigned IBMTSS switched to `tpm2-tools`.

Questionnaires. We collected data about the participants’ backgrounds and understand their perceptions and opinions while using TPMs (refer to Appendix F). After showing the general instructions, we asked basic questions about the participant’s *demographics*, *TPM background* and *contact details*.

After each task, we presented *task-specific* questions where the participants had a chance to report their perceptions and opinions about the task. In particular, we wanted to understand their familiarity and complexity perception about the task; also their security and correctness perception about their response. We also asked questions about the type of resources they used, the reason for not completing the task (if applicable), and their opinion on the usefulness of the error messages in fixing their mistakes or making secure choices. The questionnaire used a 5-point Likert scale.

After attempting all the tasks, we presented an exit questionnaire to ask about their use of supporting materials for TPM-related activities. This question was asked to compare with a similar question asked in the *task-specific* questionnaire and to understand whether they had to refer to new types of materials for our study. We also asked them why they referred to supporting materials outside the library documentation.

4.2 Study environment

We built our study environment as an online Integrated Development Environment (IDE) and hosted it on our servers. Refer to Appendix G for the technical details. We designed the study environment using our personal experiences to mimic the real-world TPM development conditions with minimal participant effort, and our participants confirmed its ecological validity during the interviews. Each participant got a dedicated server accessible over a unique URL and only needed a browser to participate. The IDE allowed interacting with the TPM emulator using function calls provided by each supported library (see Appendix A for details of supported libraries). The participants were free to switch between libraries at any point during the study by choosing from a dropdown menu in the IDE. We also added two additional features to the IDE to *reboot* and *reset* the TPM to its initial state.

Figure G.1 shows the interface of our study environment. The environment started with a *Welcome* page, which contained the study’s purpose and logistics, an instructional video that covered basic features and navigation of the environment, and links to FAQs. This page was followed by the *Demographics* page. Then, every task was shown on two pages: one for the task description and IDE and the other for the task-specific questionnaire. In the end, the *Final questions* page included the exit questionnaire. Participants were free to move between tasks and attempt them any number of times. But we stored everything they executed in the IDE.

Task	Security features		Description
	Crypto	Non-Crypto	
Asymmetric encryption	C2	NC1, NC2	<ul style="list-style-type: none"> • Create secure encryption keys using the TPM • Perform asymmetric encryption • Impose restrictions on key attributes to allow duplication of the key, which makes it exportable to other devices
Symmetric encryption	C1	NC4, NC5	<ul style="list-style-type: none"> • Create password-protected symmetric encryption keys using the TPM • Perform symmetric encryption • (Optional) Use TPM sessions to authorize the use of a password protected key
Storing measurements	C4	NC6	<ul style="list-style-type: none"> • Perform secure hashing • Identify suitable PCRs for storing measurements • Extend measurements to all available PCR banks in the TPM
Securing secrets	—	NC1, NC3	<ul style="list-style-type: none"> • Create NVRAM index to securely store information in the TPM • Use the correct parameters for NVRAM index creation • Seal the reading operation against a PCR state • Lock the NVRAM index against future writes
Remote attestation	C3	NC1, NC2, NC6	<ul style="list-style-type: none"> • Create secure keys using the TPM • Impose restrictions on the key used for signing a quote • Create a quote with the TPM including the state of the kernel (stored in PCR2) • Verify if a given quote is valid for remote attestation

Table 1: Mapping of use cases and security features into tasks (also refer to Appendix E and D)

4.3 Interview

We conducted semi-structured interviews with 9 participants to extract in-depth qualitative insights beyond the task analysis. In particular, we were interested in understanding participants’ mental models and experiences while working with TPMs. We also wanted to know how they would resolve the usability and secure programming barriers that we identified. Two of the authors facilitated the interviews, one for leading the conversation and the other for observing and note-taking. We prepared the main questions we wanted to cover with our participants. We analyzed each participant’s responses to pick the most suitable version of their code snippets and relevant observations from their questionnaire responses. During the interviews, we showed these responses and used appropriate probes to obtain in-depth information about our main question. We invited the participants for a one-hour online interview, and asked for consent to record and auto-transcribe the call. The interviews were loosely structured into three segments. Refer to Appendix H for more details.

(1) Introduction. We first reminded the participants about the study details and explained the purpose and structure of the interview. Next, we probed them with open-ended questions about the working mechanism of TPMs, their use cases, and their relevance to participants’ work. We also confirmed that the participants were comfortable with the study environment and had a positive overall experience with the logistics.

(2) Task- and questionnaire-specific observations. We were interested in understanding the participant’s motivation behind specific choices, as well as their mental model and problem-solving approach when completing the tasks. We showed them their responses and asked them to outline their approach. We used necessary probes (e.g., “*how did you pick the cryptographic algorithms for creating the keys?*”) to un-

derstand further their thoughts on the security and correctness of their approach. We also probed to check their awareness of alternative approaches and discussed their advantages and disadvantages. In addition, we tried to understand how they search for relevant information about TPM-related topics.

(3) General discussions. We probed the participants on general topics outside the tasks and questionnaires. For example, we asked the participants for their usual approach to solving TPM programming tasks. Also, we collected their suggestions on improving the TPM ecosystem and offered them a chance to address anything that was not directly part of our questions.

4.4 Overview of analysis

Our analysis included two phases. In the first phase, we analyzed the data collected from the study environment, such as the code snippets executed in the IDE and the responses to the questionnaires. While we analyzed the code snippets for correctness and security, our primary goal was to understand the typical solutions and coding patterns of TPM developers. Furthermore, we identified interesting code snippets and responses to use as probes for the interview. We present our observations about common coding patterns in §5.2.

In the second phase, we conducted independent and iterative analysis of the interview transcripts for thematic coding (refer to §5.1). First, we identified three broad categories under which we could explore themes. Then, two of the authors independently coded the transcripts (using an inductive approach) to generate a list of all potential codes that would suit these categories. We noticed that no new concepts emerged from the last two interviews, which indicates saturation. This process was repeated several times, until all concepts mentioned by participants were reflected, to refine the codebook. We consolidated the codebooks with refined code names and

guidelines. The codes represented concrete roadblocks or influencing factors for the TPM developers. We do not report any measure of inter-rater reliability (IRR), because we report no quantitative results and the iterative review of the codes was the process to yield the themes we identify in our work [36]. Furthermore, the codes themselves are not the product of our work. We identified eighteen codes under seven themes from our thematic analysis (refer to Table 2).

5 Results

In this section, we present our findings of themes from the thematic analysis and common coding patterns of TPM developers. The results from thematic analysis (in §5.1) are presented loosely as a tuple of *observations* and *evidences* along with a brief *discussion* pertaining to each theme. Each theme’s main observation is aggregated from analyzing the developer’s perceptions and opinions of the questionnaire data and thematic analysis of the interview transcripts. Relevant snippets from the participant’s code submissions or quotes from interviews are used for justifying the observations. The discussion reflects our observation from the literature analysis and conversation with the participants about the potential treatment to some of the roadblocks they usually face while working with TPMs. We have followed a similar approach for presenting the common coding patterns in §5.2.

5.1 Themes emerged

Our thematic analysis yielded three categories: library, supporting materials and user themes, as summarized in Table 2.

5.1.1 Library Themes

This category captures the reasons why something went wrong when completing the study, which were related to the library itself. We identified three themes as follows:

Naming conventions and usage. Participants P1–P3, P5, P8, and P9 expressed difficulties with the naming conventions used in the libraries and deciding when to use each command. This theme includes codes around three areas: *confusion between two available options* with overlapping functionality, *names do not convey the functionality or cause confusion*, and *inconsistency* in syntax when specifying similar parameters but for different commands.

For example, P1 was confused between the commands `tpm2_createprimary` and `tpm2_create`. The former is used to create primary objects (such as *primary keys*), and the latter is used to create child objects (such as *child keys* and *sealed data blobs*) that are protected by a primary key in the hierarchy. P1 was confused because the difference in functionality was not clear from the names of the commands.

While the library implementation follows the specifications [52], they can alleviate the confusion by providing

abstraction functions with more appropriate names. We suggest that the functionalities of the original commands can be split into two new commands: e.g., `tpm2_createkey` for creating parent or child (with `-P` and `-C` as flags, respectively) and `tpm2_createblob` for creating sealed data blobs. Nevertheless, such abstractions must retain the parameters used in the original commands and implement secure defaults.

Output formats. P1–P4, P6, P8, and P9 had difficulties interpreting the outputs of different commands they used for completing the tasks. The problems include *insufficient information* and a *lack of clarity on how to interpret* the output.

In the remote attestation task, the participants were asked to verify a quote, which required using `tpm2_checkquote` to verify the contents and signature of a quote. The output lacks a success message, making developers rely on ad hoc methods to confirm the verification. For instance, P9 said that “I think I was doing just a minimal verification (...) I was executing this command twice, one with the correct files and one with false files to see if the end result was different”. The above example is also the case where the output lacks information about what is verified. Altogether, developers may wrongly interpret the output and assume that the verification succeeded as the command did not return any errors.

We also noticed a lack of details in the meta-information shown to developers. For example, when a key is created, the attributes of the key are printed as `fixedtpm|fixedparent|restricted|decrypt`. This meta-information works as a reminder of the capabilities and restrictions of the key. Although our participants found this meta-information helpful, and it seemed important, they admitted their reluctance to check each meta-information even if they did not fully understand it. While the brevity of such meta-information suffices for seasoned developers, the less-experienced ones still have to refer to the documentation to understand their meaning.

Clear description, success, and error messages are important for the developers [29]. Hence, a possible treatment for this situation is to return more verbose outputs that clearly confirm a successful command or indicate what went wrong.

Error handling. All participants except P6 expressed difficulties when interpreting error and warning messages provided by the library. They found the messages to be unclear and hence did not fully understand them. Also, the messages did not provide any valuable feedback (i.e., lack pointers on resolving the errors), and the participants did not have enough domain-specific knowledge to fix it by themselves.

For example, all participants of the symmetric encryption task failed to specify an initialization vector (IV). Although a warning message indicated that the IV was weak, everyone ignored it because there was no feedback on how to specify the IV. P5 commented that,

I didn’t really find a way how I could specify a better IV and, I don’t know, **I find it’s kind of destructive criticism when the**

program just tells me “well, you used the wrong IV”, but doesn’t make any comments on how to do it better. So, if it finds out hey, you are using a weak IV, it could suggest the use of the appropriate flag to specify a better one.

We traced the cause of this problem to the example in the documentation, which lacks an IV [10]. Therefore, we believe that participants copied this example and ignored the importance of an IV. One solution to this problem is to update the code in the documentation. Also, the command could return a descriptive error message that suggests how to specify the IV and the correct flag to use. The library developers can refer to the RUST compiler [60] as a good example of suggestions to include in the error messages.

	Themes	Codes	Participants	
Library	Naming conventions and usage	Confusion between two available options	6/9	
		Names do not convey functionality or cause confusion		
		Inconsistency		
Output formats	Error handling	Insufficient information	7/9	
		Lack of clarity on how to interpret		
Supporting materials	Documentation shortcomings	Error message is unclear	8/9	
		Lack of pointers on how to resolve		
		Lack of examples		7/9
		Lack of background information		
		Incorrect or missing explanations		
Complicated presentation				
User	Mental models	Not easily accessible	5/9	
		Misunderstanding by user (documentation is correct)		
		Misunderstanding due to missing or incorrect documentation		
	Trust factors	Misunderstanding due to unknown sources		7/9
		Past experience or code		
		Defaults and documentation examples are secure		
		TPM’s ability to handle security		

Table 2: Identified Themes

5.1.2 Supporting Materials Themes

We asked participants what supporting materials they used to complete our study’s tasks and in general when doing TPM-related coding. Supporting materials refer to the library documentation, TCG standards, and additional resources (e.g., blogs, personal notes, forums). Although all participants reported using the library documentation as the primary resource, we found that participants with less TPM experience reported relying on the TCG standards to get background information and on third-party forums when they faced issues.

On the other hand, experienced TPM developers said using only the library documentation to obtain ready-made code examples that they can tweak. They know what to look for based on previous experiences.

We now present the common themes around supporting materials, especially their shortcomings. We limit the results in this section to library documentation because the participants did not provide details about the exact resource they used, and also, we could not trace them. Nevertheless, we believe that the shortcomings we report in this category may also exist in all types of supporting materials.

Documentation shortcomings. Participants P1–P3, P5, P7 and P9 indicated a *lack of examples*, as well as a *lack of background information* about related TPM or security concepts. They emphasized the need for customizable examples and descriptive background information, primarily aimed at beginners. In particular, P5 stated that

What I would really love would be example code. (...) I mean, there is example code for the simple problems, **but as soon as you want to do something that goes away from the simple problems, it gets a little difficult.**

Participants P1, P5, and P7 highlighted that the documentation had *incorrect or missing explanations* and that they would prefer clear explanations for using one approach over another. An example of this theme arose in the securing secrets task, where we asked participants to store a secret string “in the TPM” and to impose restrictions on reading and writing that secret. We expected them to create an area in the NVRAM with the appropriate security controls for accessing the memory area. Instead, most participants (6/9) created a sealed data blob, which is stored outside of the TPM. During the interviews, we learned that the participants were aware of both approaches and the latter being a less secure one. However, most participants mentioned that they avoided the NVRAM approach because the documentation lacked a good explanation. Also, it would be too time-consuming for them to figure out how to complete our task using NVRAM. Hence, they settled for the less-secure but better-documented approach.

Finally, P1, P3, P7–P9 indicated that supporting materials had a *complicated presentation* or were *not easily accessible*. In particular, P9 said it was time-consuming to find the command they would need from the library documentation:

The first task was taking a lot of time because I couldn’t find the correct command. It was like, I know what I’m going to do, but none of these commands seems to be relevant. (...) And then it was almost by accident, that I found out the correct option.

All the aspects mentioned in this section could be tackled by improving the documentation quality and adding more background information about TPMs. Additionally, the documentation could include use case-based example tutorials. The `tpm2-tools` library started adding these kinds of comprehensive guides [58] with code examples and background information, but it has been defunct since February 2021.

5.1.3 User Themes

Mental Models of TPM Concepts. We identified different mental models formed by the participants that led to misconceptions of TPM concepts. We observed an example of a *misunderstanding by the user* in the asymmetric encryption task, where the participants had to create a key (using `tpm2_create`) that is exportable to another TPM. P3 misunderstood the documentation and formed an incorrect mental model about the trust boundaries of the TPM, believing that any key created with a TPM can be loaded into other TPMs. They stated that,

When I used the `tpm2_create` command, you can see that I have given two files with `-u` and `-r`. So the keys are already outside. So, I know that if I can load this key in another TPM, it should work because normally any key can be loadable on TPM as long as it is cryptographically valid.

It is true that the flags `-u` and `-r` return the public and private keys in two separate files. However, the private part remains protected by its parent key in the TPM, and it cannot be used outside unless its restrictions are relaxed. This can only be achieved by setting the `fixedTPM` and `fixedParent` attributes of the key to `False`, which P3 missed. The man page for `tpm2_create` [9] covers all flags and includes a link to a separate page [4] that covers the flag `-r` in detail. The latter page mentions the private key (i.e., the child key) being protected by its parent and further links to the TPM standard [51], which explains these concepts in depth. Although the documentation is correct, we believe that its nested presentation could have caused P3 to form a wrong mental model about the trust boundaries of the TPM. A potential solution would be to simplify the documentation by covering all necessary details (e.g., command usage, TPM concepts, and concrete examples) in a single page.

Another example from P3 is an incorrect mental model about remote attestation, due to a lack of background information in the documentation. They had not done any remote attestation tasks before our study and only read the `tpm2_checkquote` [8] man page. This man page does not include a high level of detail about all the requirements that need to be verified to trust a quote; it only discusses the quote's signature and reference PCR values. This, along with the insufficient information in the output of the command, implied to the participant that verifying a quote meant just checking the signature. However, this is only a part of the verification, since the verifier must also check the signing key's properties, e.g. that it is `restricted` and comes from a legitimate TPM. Such incorrect mental models may lead a developer to implement remote attestation insecurely. The library can address this issue by improving the already existing tutorial for remote attestation, which lives outside the documentation [57]. The tutorial can be improved by adding concrete code examples and moving it to the `tpm2-tools` documentation.

We also found several cases of incorrect mental models and misconceptions, but we could not trace the exact source that confused the participants. In such cases, we could only confirm that the confusion did not originate from the documentation and speculate or attribute it to *unknown sources*. An example of such a case emerged in the discussions about the asymmetric encryption task. Many participants were unsure of why they had to create parent keys before creating child keys, but they blindly followed the documentation examples. On the other hand, one participant (P8) had selected cryptographic parameters only for the parent key but not for the child key, as they believed the parameters would be inherited. We believe that P8's security background and the use of the words parent/child appealed to them that the child key inherits security properties from its parent. However, in reality, the child key will only be as secure as the defaults defined for the library, with no impact from its parent's cryptographic parameters. A possible treatment would be to simplify the key creation process, e.g., by offering abstraction functions to create both the parent and the child key.

Trust factors. We observed that the participants relied on their trust in different factors when making secure choices. For example, P1, P3–P5, P7, and P8 trusted and relied on their *past experience* of completing similar tasks. As a result, they preferred to use their old code snippets instead of figuring out how to approach our study's tasks from scratch. However, one pitfall of this method is that outdated code snippets could lead to trivial errors, e.g., due to software version mismatch. One participant encountered such a situation and struggled to complete a task before realizing that they wrote their code using an outdated API version.

Similarly, we observed that developers tend to trust that *the library defaults and documentation examples are secure*. This was prevalent in situations where the lack of a relevant security cryptographic background prevented the participants (e.g., P1–P3) from feeling confident in making an informed choice. Therefore, the library needs to provide secure default values and examples to support less experienced developers.

Finally, participants P3 and P8 had implicit trust in the *TPM's ability to handle security* details, so they did not have to make any explicit choices. This subtheme was prevalent in the tasks where participants had to create parent and child keys, where they would trust the TPM to take care of security aspects of those keys. P3 expected that, when using the endorsement hierarchy, the TPM would prevent the user from creating signing keys that would allow a third party to correlate a set of signatures and determine that they came from the same TPM; whereas P8 expected the TPM to assign child keys the same cryptographic parameters as their parent key. We suggest that the library should clarify in the documentation which security aspects are covered by the TPM to inform users about the security guarantees it can provide.

5.2 Common coding patterns

This section highlights the common coding patterns followed by participants. We limit our discussion to the features for which we found interesting patterns.

5.2.1 While using non-cryptographic security features

Here, we cover the common patterns observed when using the following non-cryptographic features of the TPM: *use of hierarchies* (NC1), *key restrictions* (NC2), *session-based authorizations* (NC5) and *PCR usage* (NC6). We evaluated feature NC1 when creating TPM objects in the asymmetric encryption, securing secrets, and remote attestation tasks. We wanted participants to avoid using the NULL hierarchy since objects in this hierarchy are lost upon reboot. Most participants selected the correct hierarchy for the use case but relied on the defaults. This pattern highlights the importance of providing relevant and secure defaults, as discussed in the trust factors theme in §5.1.3. The exception to this pattern occurred in the remote attestation task, where most participants explicitly selected the endorsement hierarchy due to its privacy protections. We noticed that some participants were reluctant to rely on defaults if the use case demanded a specific type of protection from the hierarchy and would instead explicitly specify the hierarchy. We speculate that developers hesitate to rely on defaults and tend to be extra cautious when specifying parameters if the use case's security requirements are well understood. Hence, the library must create awareness by highlighting such requirements in the documentation.

Then, feature NC2 was evaluated when participants set restrictions on the keys created for the remote attestation task, where they were asked to create a `restricted` key for signing a quote. A signing key without the `restricted` attribute could be misused during remote attestation to sign any data, including a forged quote. Although the library offers a convenience function `tpm2_createak`, which takes care of setting the `restricted` attribute automatically, only two participants used this function. Moreover, only 3/9 participants created restricted signing keys, whereas the rest relied on the default key attributes set at creation time. Again, we found that most developers rely on key's default attributes instead of setting them explicitly and that convenience functions are rarely used.

We evaluated NC5 in the symmetric encryption task, where participants had to encrypt and decrypt both a file and a string multiple times using a password-protected key. When password authorization is used, the password is sent in plaintext between the user and the TPM every time the participants perform an operation. A local attacker can eavesdrop on such communication to capture the password. A secure way to use password protection is to utilize TPM session utilities, e.g., HMAC or policy sessions (refer to §2); however, none of the participants used them. During the interviews, we found that many participants had theoretical knowledge of sessions but lacked the hands-on experience to use them.

Finally, feature NC6 was evaluated in the storing measurements tasks, where participants had to pick suitable PCRs for storing the measurements of a configuration file. We expected them to avoid PCRs 16 or 23, which can be reset and repopulated with arbitrary measurements during run-time by an attacker. We found that 2/9 participants used PCR 23. During discussions, they blamed the TPM specification [53], which states that PCR 23 is meant for “Application support” and deceived them into thinking PCR 23 is reserved for any application needing to store measurements. However, the specification also mentions that the operating system dictates its usage, and it may be reset and used at any time. Unfortunately, both participants had ignored the latter part and formed incorrect mental models. This highlights the importance of the specifications in shaping users' mental models and misconceptions.

5.2.2 While using cryptographic security features

We now report the common patterns observed while performing *symmetric encryption* (C1), *asymmetric encryption* (C2), *signing* (C3) and *hashing* (C4). For cryptographic security features C1, C2, and C3, we noticed that when participants were asked to create keys for encryption and signing, they first created a parent key in one of the hierarchies and then a child key. Also, most participants specified the cryptographic attributes (e.g., algorithm, key length, purpose, and duplication restrictions) only for the child key, whereas they relied on the library defaults for the parent. We traced back the origin of such patterns to the documentation examples in the `tpm2-tools` library that miss out on specifying the attributes for the parent. Child keys are stored outside of the TPM (e.g., on the disk) until they are loaded onto the TPM for use; therefore, it is good that developers are cautious and explicitly specify their attributes. Nevertheless, attributes of the parent are equally crucial because child keys may be compromised if their parent has weak or insecure attributes. We can confirm from our analysis that the defaults of the parent keys are secure. Nevertheless, the library needs to be aware that the developers have a high degree of trust in defaults, and it has a responsibility always to keep the defaults secure and updated.

Another common mistake appeared in the symmetric encryption task (feature C1). All four participants asked to complete the task failed to specify an IV, which lowers the security of the encryption process.

C4 was evaluated in the storing measurements task. We noticed that 8/9 participants relied on the hashing functions provided by the TPM to obtain the hash of the file they would later extend to a PCR, where one participant used an external library for hashing. TPM library developers should be aware of such patterns and restrict the use of external libraries that lie outside their control and may contain vulnerabilities.

On top of the above features, we observed an interesting pattern in the remote attestation task where the participants were asked to generate a quote. Including a random nonce

with the quote allows the attestation server to defend against replay attacks. Unfortunately, only 2/9 participants' quotes included a nonce, where one used a random nonce, and the other had blindly copied the nonce from the man page example of the `tpm2_checkquote` command. During the interviews, we learned that most participants ignored using nonce because the documentation is inconsistent and does not emphasize its importance. Another reason for not using a nonce was the lack of a threat model explicitly indicating a replay attack. Some participants also noted that the library uses a relatively unfamiliar term (`-qualification`) to refer to a nonce, which does not immediately evoke the concept of a nonce.

The library can address the concerns presented in this section by revising the documentation to be consistent, providing secure defaults, clear explanations and examples, and using familiar terminologies that could remind the developers about common threats they should consider.

6 Recommendations

This section provides explicit recommendations, in the form of concrete action points, for improving the TPM library API documentation and software. Some of our recommendations overlap with Green and Smith's usable and secure API design principles [30]. These include abstracted and readable API design that does not go against developers' habits and mental models, non-ambiguous and safe defaults, and detailed and visible errors and outputs. We observed that `tpm2-tools` does not follow these existing guidelines; therefore, we reiterate and emphasize some of these principles in the form of actionable recommendations for the library developers. The emphasis of [30] is on making APIs easy to learn and use such that the developers have no need to understand the complexities of cryptography and minimal reliance on the documentation. While this standpoint is valid for mature ecosystems, it is crucial to understand complex concepts in niche ecosystems like TPM, as the knowledge is not widespread. We argue that it is still the library's responsibility to educate and assist the developers. Thus, our recommendations also focus on improving the documentation with missing details of the TPM background as the first step towards more useably secure APIs.

6.1 For library documentation

Technical specifications and standards are meant to be detailed and comprehensive. Despite that, many developers refer to software documentation or tutorials written by other developers. Our study confirms this phenomenon. In particular, we found that the participants' primary information source was the `tpm2-tools` documentation, and it influenced the developers' decisions. Although questionnaire responses deemed the `tpm2-tools` documentation satisfactory, our participants expressed frustration during the interviews while referring to them. Similarly, our analysis of the ecosystem revealed

several shortcomings. We now present them collectively as concrete pointers to improve the documentation.

Include background information. Developers need to clearly understand TPM concepts and how they can be leveraged for security functionalities. We argue that the current documentation lacks background information about TPM concepts. As some of the experienced participants in our study pointed out, their theoretical knowledge and experience can only give them a sense of familiarity. However, given the complexity of concepts and the abundance of options available, providing more information would help them confidently utilize the needed options. Furthermore, we found several examples of misconceptions and incorrect mental models by less-experienced developers, which could also be addressed by providing additional background information.

Provide code snippets for common use cases. Most developers use code examples from the documentation as their starting point and repurpose them as per their needs. Although the `tpm2-tools` documentation contains simple code snippets, it lacks concrete examples for common use cases that require the use of multiple commands (e.g., remote attestation). Our participants also confirmed that they could not benefit much from the simple snippets. In this realm, we recommend that the TPM library uses the common use cases that we have identified (refer to §4.1) as a starting point and include comprehensive examples around them in the documentation.

Improve entry-level documentation. Despite their prior coding and security experience, all our participants shared similar struggles when they started developing with TPMs. They found the TPM documentation to not be beginner-friendly. Many beginners seem to have faced difficulties setting up their development environment on their local machine and setting up a communication interface with a TPM. We speculate that these difficulties, along with the complexity of TPM concepts, would discourage the developers and may be one of the reasons why there are few TPM developers. To this end, our recommendation would be to improve documentation with carefully curated content for entry-level developers. Our study environment provides an online coding experience that could be leveraged to teach how to code with TPM without installing anything on a local machine.

Include guidelines for picking security attributes. We found that developers selected security attributes and cryptographic primitives based on their prior experience rather than explicitly looking for existing guidelines (e.g., [41]). In the case of TPMs, developers have additional non-cryptographic security attributes to choose from, and a strong cryptographic background would not be enough to help with their decision. We recommend including brief guidelines within the documentation about both cryptographic and TPM-specific security attributes. Having all required information in the documentation would help developers formulate a security

rationale and encourage secure choices. Any such guidelines should cover the various attributes available but explicitly promote the most secure ones (e.g., by using them in the code examples). At the very least, the documentation should include external links to any relevant guidelines.

Improve the documentation to address incoherence. We observed that developers were confused by incoherent aspects of the documentation on four occasions. The first occasion is with the confusing naming conventions (refer to §5.1.1) — i.e. when referring to commands or functions that share a common prefix (such as `tpm2_create` vs. `tpm2_createprimary`). The documentation could address this by alerting and reminding the developer about the other commands. Secondly, many commands are to be used in conjunction with suitable flags and parameters, and the developers doubted of their choices. Instead of just listing the available options, the documentation should provide verbose descriptions and concrete examples.

The third occasion is when there are multiple approaches for a task. For example, when to choose NVRAM over sealed blobs for storage is trivial for experienced TPM developers, whereas newcomers struggle to make the distinction. In such cases, the documentation could compare the approaches and help developers make the correct choice for their use case. Finally, several instances of confusion exist due to inconsistencies in the naming conventions across software versions. For example, a participant complained about discrepancies in flag usage between different versions of `tpm2-tools`. This inconsistency breaks Green and Smith’s “ensure that APIs are easy to read and update” principle, confuse the participant and make them gullible to commit trivial mistakes. Along with clear documentation, these confusions can be prevented by using common naming conventions that align with developers’ habits and mental models.

All our recommendations above provide essential components that should reflect in the documentation. However, the libraries must pay attention to balancing the depth and presentation of information such that it neither overwhelms the new developers nor bores the seasoned developers. One potential solution would be to use documentation tools with rich text features to improve readability and presentation.

6.2 For library software

Provide developer-friendly error messages. Designing effective error messages to provide feedback and assist developers is a longstanding research theme [33, 50]. However, many libraries, including `tpm2-tools`, produce incomprehensible error messages that are least useful and frustrate the developers. We have highlighted some of the examples in §5.1.3. A common concern among our participants was that the error messages lacked clarity and did not help resolve the problem. Despite that, seasoned developers have adapted their mental models of associating the ambiguous error messages with

something more concrete based on their experience. Since new developers do not have this advantage, we suggest the library review existing error messages carefully. There is a wide range of design guidelines on developing human-centric error messages [20, 25, 46, 59], and we recommend the TPM community to adopt them. Similar to Green and Smith’s principles “APIs should interact with the end user” and “APIs should be hard to misuse”, we suggest revising the error messages to be more specific and provide constructive feedback, e.g., that suggests resolution or guides towards the right resources.

Provide concise output messages. Similar to error messages, our participants were unsatisfied with the output messages of `tpm2-tools` as they lacked clear feedback. Our recommendation is to review the output message formats to include the following necessary components. The output should show a *success message* that not only assures the developer that the commands have been executed without errors but also provides them *feedback* on whether the command has achieved its goal. Also, any interpretation and obvious additional steps (e.g., executing another command) to be taken must be clearly communicated. Additionally, the output should include a *concise description of meta-information* (e.g., of the objects created) if applicable. This recommendation is one way to make the API self-explanatory and ensure that principle “APIs should be easy to use” from Green and Smith is followed.

Utilize abstractions for sequential command execution. There are several occasions where multiple commands have to be executed sequentially. For example, for storing measurements of a file in a PCR (in task 2), one can take the hash of a file and then extend it to a PCR using `tpm2_hash` and `tpm2_pcrextend` commands, respectively. Alternatively, `tpm2-tools` offers an abstraction function called `tpm2_pcrevent` which combines hashing and extending in one go. We strongly believe that such abstractions provide convenience to the developers and make their code less error-prone by triggering a sequential execution of functions that might be missed otherwise. Unfortunately, while there are several abstraction functions available in the `tpm2-tools` library, they seem to be underutilized. We recommend that libraries promote and highlight the advantages of abstraction functions whenever available. Also, they should identify occasions where the order of command execution has to be preserved and provide abstractions for them. Even better would be to provide abstraction functions for common use cases.

Promote secure cryptographic primitives. We recommended in §6.1 that the documentation should include guidelines for picking security attributes. While security-conscious developers may benefit from that, it is common for developers to rely on the default options, especially while picking cryptographic primitives provided by the library. This finding is also reflected in Green and Smith’s “Make defaults safe and unambiguous” principle. Thus, libraries must avoid supporting algorithms with known security vulnerabilities and

set the defaults to the most secure primitive. If the insecure algorithms have to be supported for legacy reasons, their use should be discouraged, e.g., via warnings. When we examined cryptographic algorithms supported by the `tpm2-tools` library, we found that the library does not support insecure algorithms. However, the default options are set to the primitive with bare minimum security in most cases. We suggest that the library should consider updating defaults to the most secure option available. We also remind that support for cryptographic algorithms should be regularly audited, and support for insecure ones (if found) should be discontinued.

7 Discussion

Our results hint at the importance of threat models in the API ecosystem. We found that developers do not always feel forced to make security choices unless a threat model is given or they understand the threats to defend against. Documentation could bridge this gap by including common threat models, extensive background topics and secure coding examples built around common use cases. We believe that including threat models in the documentation of security-critical technologies would help developers cultivate intuitive security thinking and form correct mental models.

Similarly, naming conventions, along with text and format of errors and outputs, are crucial in invoking security thinking among the developers. For example, using familiar terminologies (such as `nonce`) as part of the commands could do the trick. However, in practice, software implementations often blindly borrow the names and texts defined in the standard specifications. This leaves minimal scope for improvement in the later stages. In this realm, we recommend considering usability and human factors already in the creation of the specifications. In particular, HCI usability experts should review the function names, text, and error and output formats. We believe API design principles [30] that are typically recommended at the implementation stage can also be applied while designing the specification.

Roadblocks faced by developers could easily make them gullible to commit mistakes, which could have serious consequences in the context of security-critical technologies. Libraries can learn from the common mistakes and address them in the software development life cycle. Modern code repositories provide an easy way of tracking end-user issues and creating automated workflows for integrating their solutions (e.g., as a feature) in software updates.

Our interaction with the developers drew attention to the importance of communities, as they often seek help from sources outside the software documentation. In particular, they rely on peer support and prefer immediate help (e.g., ready-made code and error resolution tips). Unfortunately, the generic venues for such help (e.g., Stack Overflow) lack useful content for niche technologies, such as TPM. In such cases, community forums of these technologies play a vital role. For

example, we observed that *Tpm.dev* is one such community forum for TPM developers to discuss concerns and learn from each other, irrespective of experience or library preference. *Tpm.dev* has also recently started to share beginner-friendly resources [11], written by seasoned developers. We have discussed our results with *Tpm.dev* to encourage their initiative and hope they benefit from this study.

Limitations. One of the limitations of this work is the small number of participants. Despite that, we believe our results are generalizable due to the participants' expertise. Similar to Nielsen-like heuristic evaluations [42], we argue that the usability issue discovered for one participant is likely to indicate a general usability issue [31, 39, 45]. The other limitation is that our results are drawn only by observing the `tpm2-tools` library. Despite being allowed to choose a different TPM library supported by our study environment, all participants, including those familiar with multiple libraries, used `tpm2-tools`. However, we consider our results generalizable because the other libraries, similar to `tpm2-tools`, closely mirror the specifications. A comparison is provided in Appendix A. Our independent analysis confirmed that they do not provide more user-friendly function abstractions or naming. While the author of the `IBMTSS` library claims to provide a simpler interface, there is no empirical evidence to support their usability claims. Future studies can utilize our study environment and insights from this study to validate such claims or, even better, to conduct a quantitative and comparative analysis between different TPM libraries.

8 Conclusions

We conducted the first qualitative study targeting TPM library APIs and found that they are not developer-friendly. In particular, we identified specific areas where the TPM library APIs contain usability and security pitfalls and provided recommendations to fix them. Our contributions also include an open-source environment for TPM usability studies.

Our findings support those of past API usability and security studies. Additionally, we found new insights by studying the interesting combination of cryptographic and non-cryptographic features of TPM that is rarely seen in previously studied security APIs. Some of the identified pitfalls can be traced back to the TPM specification that forms the design basis for software implementation. Based on this observation, we highlight an important issue: any technology that follows standard specifications tends to accumulate usability pitfalls, well before its implementation, in the standards design. This is an opportunity for standardization bodies to prioritize usability by involving HCI experts in the design process. Previous studies have not highlighted this issue, and no usability frameworks have been explicitly created for the standards. We hope that our work inspires and steers future research in this direction.

Acknowledgments

We thank our anonymous reviewers for their insightful reviews and feedback that helped us improve the paper. We also thank our participants, without whom this study would not have been possible. We are grateful to Yoan Miche from Nokia Bell Labs for his support and discussions throughout this research project.

References

- [1] Go-TPM. [Online]. <https://github.com/google/go-tpm>.
- [2] IBM's TPM 2.0 TSS. [Online]. <https://sourceforge.net/projects/ibmtpm20tss/>.
- [3] Judge0. [Online]. <https://github.com/judge0/judge0>.
- [4] Protection details. [Online]. <https://github.com/tpm2-software/tpm2-tools/blob/5.0/man/common/protection-details.md>.
- [5] SurveyJS. [Online]. <https://github.com/surveyjs/survey-library>.
- [6] TPM study environment. [Online]. <https://github.com/nokia/tpm-study-environment>.
- [7] tpm2-tools. [Online]. <https://github.com/tpm2-software/tpm2-tools>.
- [8] tpm2_checkquote(1) tpm2-tools | General Commands Manual. [Online]. https://github.com/tpm2-software/tpm2-tools/blob/5.0/man/tpm2_checkquote.1.md.
- [9] tpm2_create(1) tpm2-tools | General Commands Manual. [Online]. https://github.com/tpm2-software/tpm2-tools/blob/5.0/man/tpm2_create.1.md.
- [10] tpm2_encryptdecrypt(1) tpm2-tools | General Commands Manual. [Online]. https://github.com/tpm2-software/tpm2-tools/blob/5.0/man/tpm2_encryptdecrypt.1.md.
- [11] TPM.dev tutorials. [Online]. <https://github.com/tpm2dev/tpm.dev.tutorials>.
- [12] wolfTPM. [Online]. <https://github.com/wolfSSL/wolfTPM>.
- [13] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. Comparing the usability of cryptographic apis. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171. IEEE, 2017.
- [14] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305. IEEE, 2016.
- [15] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. How internet resources might be helping you develop faster but less securely. *IEEE Security & Privacy*, 15(2):50–60, 2017.
- [16] Yasemin Acar, Christian Stransky, Dominik Wermke, Charles Weir, Michelle L Mazurek, and Sascha Fahl. Developers need support, too: A survey of security advice for software developers. In *2017 IEEE Cybersecurity Development (SecDev)*, pages 22–26. IEEE, 2017.
- [17] Steven Arzt, Sarah Nadi, Karim Ali, Eric Bodden, Sebastian Erdweg, and Mira Mezini. Towards secure integration of cryptographic software. In *2015 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward!)*, pages 1–13, 2015.
- [18] Hala Assal and Sonia Chiasson. 'think secure from the beginning' a survey with software developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.
- [19] Dejan Baca, Kai Petersen, Bengt Carlsson, and Lars Lundberg. Static code analysis to detect software security vulnerabilities-does experience matter? In *2009 International Conference on Availability, Reliability and Security*, pages 804–810. IEEE, 2009.
- [20] Brett A Becker, Paul Denny, Raymond Pettit, Durell Bouchard, Dennis J Bouvier, Brian Harrington, Amir Kamil, Amey Karkare, Chris McDonald, Peter-Michael Osera, et al. Compiler error messages considered unhelpful: The landscape of text-based programming error message research. In *Proceedings of the working group reports on innovation and technology in computer science education*, pages 177–210. 2019.
- [21] Stefan Berger and David Safford. SWTPM - software TPM emulator. [Online]. <https://github.com/stefanberger/swtpm>.
- [22] Jan Camenisch, Liqun Chen, Manu Drijvers, Anja Lehmann, David Novick, and Rainer Urian. One tpm to bind them all: Fixing tpm 2.0 for provably secure anonymous attestation. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 901–920. IEEE, 2017.
- [23] Stéphanie Delaune, Steve Kremer, Mark D Ryan, and Graham Steel. A formal analysis of authentication in

- the tpm. In *International Workshop on Formal Aspects in Security and Trust*, pages 111–125. Springer, 2010.
- [24] Stéphanie Delaune, Steve Kremer, Mark D Ryan, and Graham Steel. Formal analysis of protocols based on tpm state registers. In *2011 IEEE 24th Computer Security Foundations Symposium*, pages 66–80. IEEE, 2011.
- [25] Paul Denny, James Prather, Brett A Becker, Catherine Mooney, John Homer, Zachary C Albrecht, and Garrett B Powell. On designing programming error messages for novices: Readability and its constituent factors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.
- [26] Andreas Fuchs, Christoph Krauß, and Jürgen Repp. Advanced Remote Firmware Upgrades Using TPM 2.0. In *ICT Systems Security and Privacy Protection*, pages 276–289. Springer, Cham, May 2016.
- [27] William Futral and James Greene. *Intel® Trusted Execution Technology for Server Platforms: A Guide to More Secure Datacenters*. Apress, Berkeley, CA, 2013.
- [28] Peter Leo Gorski and Luigi Lo Iacono. Towards the usability evaluation of security apis. In *HAISA*, pages 252–265, 2016.
- [29] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Moeller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, Too: On the Effect of Integrated Security Advice on Cryptographic API Misuse. In *Proceedings of the Fourteenth Symposium on Usable Privacy and Security*, page 17, Baltimore, MD, USA, August 2018.
- [30] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security & Privacy*, 14(5):40–46, 2016.
- [31] Thomas Grill, Ondrej Polacek, and Manfred Tscheligi. Methods towards api usability: A structural analysis of usability problem categories. In *International conference on human-centred software engineering*, pages 164–180. Springer, 2012.
- [32] Julie M Haney, Mary Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "we make it a big deal in the company": Security mindsets in organizations that develop cryptographic products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 357–373, 2018.
- [33] James J Horning. What the compiler should tell the user. In *Compiler Construction*, pages 525–548. Springer, 1974.
- [34] Shohreh Hosseinzadeh, Bernardo Sequeiros, Pedro RM Inácio, and Ville Leppänen. Recent trends in applying tpm to cloud computing. *Security and Privacy*, 3(1):e93, 2020.
- [35] Luigi Lo Iacono and Peter Leo Gorski. I do and i understand. not yet true for security apis. so sad. In *Proc. of the 2nd European Workshop on Usable Security, ser. EuroUSEC*, volume 17, 2017.
- [36] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of ACM Human-Computer Interaction*, 3(CSCW), November 2019.
- [37] Microsoft. BitLocker (Windows 10) - Windows security. [Online]. <https://docs.microsoft.com/en-us/windows/security/information-protection/bitlocker/bitlocker-overview>, December 2021.
- [38] Microsoft. Secure the Windows boot process - Windows security. [Online]. <https://docs.microsoft.com/en-us/windows/security/information-protection/secure-the-windows-10-boot-process>, December 2021.
- [39] Eduardo Mosqueira-Rey, David Alonso-Ríos, Vicente Moret-Bonillo, Isaac Fernández-Varela, and Diego Álvarez-Estévez. A systematic approach to api usability: Taxonomy-derived criteria and a case study. *Information and Software Technology*, 97:46–63, 2018.
- [40] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. Jumping through hoops: Why do java developers struggle with cryptography apis? In *Proceedings of the 38th International Conference on Software Engineering*, pages 935–946, 2016.
- [41] National Institute for Standards and Technology (NIST). Cryptographic standards and guidelines. [Online]. <https://csrc.nist.gov/Projects/cryptographic-standards-and-guidelines>, 2017. Accessed on: Nov 3, 2021.
- [42] Jakob Nielsen. How to conduct a heuristic evaluation. [Online]. <https://www.ingenieriasimple.com/usabilidad/HeuristicEvaluation.pdf>, 1995.
- [43] Daniela Oliveira, Marissa Rosenthal, Nicole Morin, Kuo-Chuan Yeh, Justin Cappos, and Yanyan Zhuang. It's the psychology stupid: how heuristics explain software vulnerabilities and how priming can illuminate developer's blind spots. In *Proceedings of the 30th Annual Computer Security Applications Conference*, pages 296–305, 2014.

- [44] Nikhil Patnaik, Joseph Hallett, and Awais Rashid. Usability smells: An analysis of developers' struggle with crypto libraries. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 245–257, 2019.
- [45] Helen Petrie and Christopher Power. What do users really care about? a comparison of usability problems found by users and experts on highly interactive websites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2107–2116, 2012.
- [46] James Prather, Raymond Pettit, Kayla Holcomb McMurry, Alani Peters, John Homer, Nevan Simone, and Maxine Cohen. On novices' interaction with compiler error messages: A human factors approach. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*, pages 74–82, 2017.
- [47] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. A summary of survey methodology best practices for security and privacy researchers. Technical report, 2017.
- [48] Nabil Schear, Patrick T. Cable, Thomas M. Moyer, Bryan Richard, and Robert Rudd. Bootstrapping and maintaining trust in the cloud. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pages 65–77, Los Angeles California USA, December 2016. ACM.
- [49] Jianxiong Shao, Yu Qin, Dengguo Feng, and Weijin Wang. Formal analysis of enhanced authorization in the tpm 2.0. In *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, pages 273–284, 2015.
- [50] Ben Shneiderman. Designing computer system messages. *Communications of the ACM*, 25(9):610–611, 1982.
- [51] TCG. Trusted Platform Module Library Part 1: Architecture. Trusted Platform Module Library Specification, Family 2.0 Level 00 Revision 01.59, The Trusted Computing Group, November 2019.
- [52] TCG. TCG Feature API (FAPI) Specification. Technical Report Version 0.94 Revision 09, June 2020.
- [53] TCG. TCG PC Client Platform Firmware Profile Specification. TCG PC Client Platform Firmware Profile Specification, Family 2.0 Level 00, Version 1.05 Revision 23, May 2021.
- [54] TCG. TCG TSS 2.0 Enhanced System API (ESAPI) Specification. Technical Report Version 1.00 Revision 14, October 2021.
- [55] TCG. TCG TSS 2.0 System Level API (SAPI) Specification. Technical Report Version 1.1 Revision 36, October 2021.
- [56] The Trusted Computing Group (TCG). Tpm 2.0 library specification. [Online]. <https://trustedcomputinggroup.org/resource/tpm-library-specification/>, 2019. Accessed on: Nov 27, 2021.
- [57] tpm2-software community. Remote attestation. [Online]. <https://tpm2-software.github.io/tpm2-tss/getting-started/2019/12/18/Remote-Attestation.html>.
- [58] tpm2-software community. Tutorials. [Online]. <https://tpm2-software.github.io/tutorials/>.
- [59] V Javier Traver. On compiler error messages: what they say and what they mean. *Advances in Human-Computer Interaction*, 2010, 2010.
- [60] Jonathan Turner. Shape of errors to come. [Online]. <https://blog.rust-lang.org/2016/08/10/Shape-of-errors-to-come.html>.
- [61] Juan Wang, Yuan Shi, Guojun Peng, Huanguo Zhang, Bo Zhao, Fei Yan, Fajiang Yu, and Liqiang Zhang. Survey on key technology development and application in trusted computing. *China Communications*, 13(11):70–90, 2016.
- [62] Stephan Wesemeyer, Christopher JP Newton, Helen Treharne, Liqun Chen, Ralf Sasse, and Jordan Whitefield. Formal analysis and implementation of a tpm 2.0-based direct anonymous attestation scheme. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, pages 784–798, 2020.
- [63] Glenn Wurster and Paul C Van Oorschot. The developer is the enemy. In *Proceedings of the 2008 New Security Paradigms Workshop*, pages 89–97, 2008.

Appendices

A TPM library comparison

Table 3: Comparison between TPM Library APIs

	tpm2-tools [7]	IBMTSS [2]	go-tpm [1]	wolfTPM [12]
Programming language	Shell	C/Shell	Go	C
Date of creation	August 2015	May 2015	February 2018	January 2018
Usage statistics ⁺	507 GitHub stars 85830 downloads	11 Source-forge reviews 22136 downloads	391 GitHub stars dependency of 58 Go projects	124 GitHub Stars 75 downloads
Version used in our study [*]	v5.0	v1.5.0	v0.3.2	v2.0.0
Supports functions needed for study tasks	Yes	Yes	Yes	Yes
Follows the standards closely	Yes	Yes	Yes	Yes
Usability claims	No	Yes	No	No

^{*} This was the latest release of the library at the time of the study design.

⁺ As of May 19, 2022.

B Preliminary survey

- Have you worked on hardware-based security (trusted computing)?
- Have you developed using the Trusted Platform Module (TPM)? Development of any kind that requires hands-on skills
- What are you using/have you used TPM for? (As part of work; Hobby projects; School/University projects; Other (please specify))
- How many years of experience do you have with TPM development (Less than a year; 1-2 years; More than 2 years)
- Which version of TPM have you used? (TPM v1.x; TPM v2.0)
- Have you used any of the TPM simulators?
- Which TPM software stack (TSS) have you used or use? [All that apply] (tpm2-software (tpm2-tss/tpm2-tools); TPM 2.0 TSS by IBM, TPM TSS by Microsoft; go-tpm by Google; WolfTPM; Other (please specify))
- What have you used TPM for? Feel free to describe it in detail. If you have any open source projects, blog posts, or products, we would love to have a look at them. (free text)
- Are you interested in taking part in our study about the usability of TPM libraries? The study will include some

simple tasks that involve using TPM libraries. Your efforts will be fairly compensated.

- Please provide us your email. We will use your email only for contacting you for the next phases of our study. If not, your email will be deleted immediately and permanently. The email provided by you is not used for any other purposes. (free text)
- Please tell us your name. We will use it only for addressing you when we send further communication about our study. (free text)

C Participant sampling and recruitment

First, we created a participant pool of the target population from personal contacts, mailing lists, forums, and code repositories. Then, we contacted them to take part in our preliminary survey via emails and social media (Twitter and LinkedIn). This survey collected participants' contact and demographic details, information about prior experience, TPM software libraries, their use of TPM and willingness to participate in the next part of the study. 36 out of 48 people expressed their interest in participating in the second part. We prioritized 34 participants who matched the criteria for our target population. We then invited them to participate via an email that contained: a reminder of their preliminary survey, a brief description of the goals, a link to their study environment, assigned library, compensation details (worth 100 €), and the approximate time needed to complete the study. 13 participants completed all the tasks, and the majority (N=11) used the `tpm2-tools` library. Our results are from a qualitative study of 9 of them whom we interviewed.

D Task descriptions

Task 1 (Track A): Asymmetric Encryption

Step 1a: Create Key. Your task is to create a secure asymmetric key of your choice for encryption purposes (e.g., for encrypting a file on your disk) using the TPM.

While creating the key, make sure the following conditions are met:

- The key should be exportable to other devices
- The key should be available across system reboots

Step 1b: Encrypt. Use the key you created in **Step 1a** to encrypt `path/to/file`

Step 1c: Reboot and decrypt

- First, reboot the environment by clicking the “*Reboot TPM*” button in the IDE.
- Then, decrypt the file you encrypted in step 1b.

Task 1 (Track B): Symmetric Encryption

Step 1a: Create Key. Your task is to create a password-protected symmetric key of your choice using the TPM.

While creating the key, keep in mind the following:

- You may have to repeatedly use this key for encryption and decryption

Step 1b: Encrypt. Use the key you created in **Step 1a** to encrypt the following:

- The string “TPMMakesMeFeelGreat”
- `path/to/file`

Step 1c: Decrypt

- Now, using the same key, decrypt the string `<enc.string>` from step 1b.
- Also, decrypt the file `<enc.file>`.

Step 1d: Cleaning the environment (Optional step) Other users may be using this environment in the future. If there is any code cleanup you would like to add, you can do so now.

Task 2: Storing Measurements

Step 2a: Measure and store in PCRs. Your task is to measure the file `path/to/file` and store the measurements using suitable PCRs.

The measurements stored in the PCRs are used in remote attestation to validate that the host machine has the correct configuration (based on `config.json`).

When storing the measurements, keep in mind the following:

- The attestation server may request the measurements from any PCR bank.

Step 2b: Read measurements. Read the contents of the PCRs (extended in step 2a). This is done to ensure that the measurements are recorded correctly.

Please make a note if you encounter any error(s).

Task 3: Securing Secrets

Step 3a: Store secret in the TPM. Your task is to store the secret string “workingWithTPMisAwesome” securely in the TPM.

The PCR allocation is as follows.

PCR 0: Core Root of Trust for Measurement,
PCR 1: Firmware, PCR 2: Kernel,
PCR 3: Config, and PCR 4–23: Unused.

While storing in the TPM, make sure that the following conditions are met:

- The secret should only be readable when the firmware has not been modified.

- The secret should not be modifiable after it has been written into the TPM.

Step 3b: Read secret. Read the secret (stored in step 3a) from the TPM. This is done to ensure the secret is stored correctly.

Please make a note if you encounter any error(s).

Task 4: Remote Attestation

Step 4a: Get quote. Your task is to get a quote of a machine for remote attestation.

The PCR allocation is as follows.

PCR 0: Core Root of Trust for Measurement,
PCR 1: Firmware, PCR 2: Kernel,
PCR 3: Config, and PCR 4–23: Unused.

While getting the quote, make sure that the following conditions are met:

- The quote should include the state of the kernel.
- The quote should be signed using an appropriate key to prove that the quote contents were generated by a legitimate TPM.

Step 4b: Verify quote. You are provided with the following files in your environment under the directory `path/to/directory`.

`q1.msg`: Quote file (signed by `key1`),
`q1.sig`: Signature file for `q1.msg`,
`key1.pub.pem`: Public part of `key1` in PEM format,
`key1.pub.tss`: Public part of `key1` in TSS format.

Your task is to verify that the quote contents of `q1.msg` can be used for remote attestation.

While verifying the quote, make sure that the following conditions are met:

- The content of the quote was generated by a TPM.
- The content of the quote has not been tampered with.

E TPM security features

Table E.4: List of cryptographic security features

Code	Description
C1	Symmetric → Encryption
C2	Asymmetric → Encryption
C3	Asymmetric → Signing
C4	Hashing

Table E.5: List of non-cryptographic security features

Code	Description
NC1	Use of the TPM hierarchies
NC2	TPM key restrictions
NC3	Restrictions against TPM-internal states (e.g. PCRs, NVRAM, counters)
NC4	Restrictions against TPM-external states (e.g. password, signature, smart cards)
NC5	Session-based command or object authorization
NC6	PCR usage

F Questionnaires

Demographics

- How long have you been programming? (Less than a year; 1-2 years; 2-5 years; More than 5 years)
- How long have you been programming with TPMs? (Less than a year; 1-2 years; 2-5 years; More than 5 years)
- In which context do you usually deal with TPM related topics? (Big company (>250 employees); Small and medium enterprise (including startups); Academic institution; On my own free time after work; Other (Please specify))
- What is your occupation?
- Are you associated with *<library>* in any of the following capacities? [All that apply] (Creator; Maintainer; Regular contributor; I might have contributed something minor; End user; Other (Please specify))
- Do you have a computer security background?
- What is your highest level of education? (No formal education; Some high school; High school or equivalent; Technical or occupational certification; Some college course work completed/Associate degree; Bachelor's (or undergraduate) degree; Master's degree; Doctorate degree)
- Please tell us your gender. (Female; Male; I prefer not to say; Other (Please specify))
- Where are you from? (dropdown)
- What is your age (in years)? (<18; 18-29; 30-39; 40-49; 50-59; >60)
- We will contact you again with respect to compensation once the survey is over. Please leave your email address below. (free text)

Task-specific questionnaire

- How familiar are you with the task that you have just attempted? (Not at all familiar; Slightly familiar; Somewhat familiar; Moderately familiar; Extremely familiar)
- How frequently have you done tasks like this one? (Never; Rarely; Sometimes; Often; Frequently)

- How difficult was this task? (Very difficult; Difficult; Neutral; Easy; Very easy)
- Did you encounter any error messages? If yes: Please rate your agreement to the following statements on a scale from 'strongly disagree' to 'strongly agree.' (Strongly disagree; Disagree; Neutral; Agree; Strongly agree)
 - The error or warning messages were helpful in improving my answers.
 - The error or warning messages were helpful in making secure choices, e.g. while selecting parameters for specific library functions.
- Did you manage to complete all the steps in this task? If yes:
 - I think my code snippet for this task is correct. (Strongly disagree; Disagree; Neutral; Agree; Strongly agree)
 - I think my code snippet for this task is secure. (Strongly disagree; Disagree; Neutral; Agree; Strongly agree)
 - Did you refer to any of the following resources while completing the task? [All that apply]
 - * Official resources (Official library documentation; TCG Technical standards; I did not use official resources)
 - * Additional resources (Mailing lists or community forums of the library that you used; Third-party/generic TPM forums (e.g., stack overflow, social media groups); Blogs, walk-through and hands-on guides; Training and workshop materials; Personal notes; I did not use additional resources; Others (please specify))
 - Did you observe anything interesting when completing the task? If yes, please describe it. (free text)
- If no:
 - Why do you think you could not complete all the steps? [All that apply] (I did not know how to do it; I could not find suitable resources to help me complete the task; I tried and gave up midway because the steps were too difficult or time-consuming; The description was not understandable; The task did not interest me; Other (please describe))
 - If "I did not know how to do it" or "I could not find suitable resources to help me complete the task:"
 - Did you refer to any of the following resources while completing the task? [All that apply]
 - * Official resources (Official library documentation; TCG Technical standards; I did not use official resources)
 - * Additional resources (Mailing lists or community forums of the library that you used; Third-party/generic TPM forums (e.g., stack

overflow, social media groups); Blogs, walk-through and hands-on guides; Training and workshop materials; Personal notes; I did not use additional resources; Others (please specify))

Exit questionnaire

- In general, which of the following do you refer to for your regular TPM-related activities? (Official resources only; Additional resources only; Mostly official resources, but sometimes additional resources; Mostly additional resources, but sometimes official resources)
- Overall, I would rate the user-friendliness of *<library>* as (Worst imaginable; Awful; Poor; Fair; Good; Excellent; Best imaginable)
- How satisfied are you with the *<library>* documentation? (Not at all satisfied; Slightly satisfied; Moderately satisfied; Very satisfied; Extremely satisfied)
- How do you rate the quality of the *<library>* documentation? (Very poor; Poor; Acceptable; Good; Very good)
- How frequently do you refer to additional resources? (Never; Rarely; Sometimes; Often; Frequently)
- If you refer to additional resources, what do you think is the reason? [All that apply] (*<library>* documentation is not clear; *<library>* documentation is incomplete/work-in-progress; *<library>* documentation does not add much beyond what is already there in the standards; There are no examples (code snippets or pseudo-code) of common use cases; Background information (e.g., TPM or programming concepts) is missing)
- Is there anything else you want to tell us about *<library>*? (free text)

G Technical details of the study environment

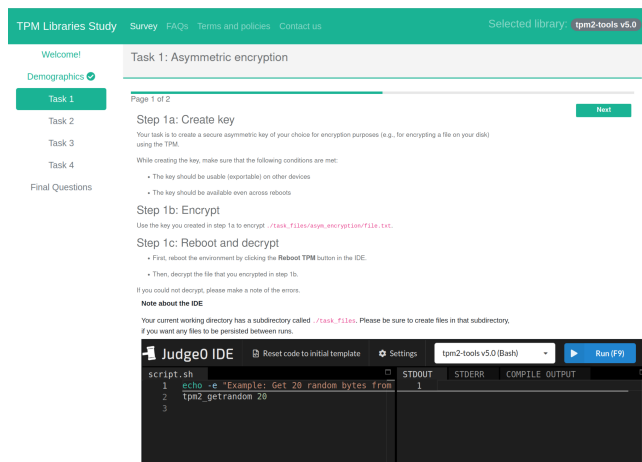


Figure G.1: Interface of the study environment

Our main goal for the study environment was to present tasks and questionnaires in a single platform, as an online Integrated Development Environment (IDE), to give a seamless user experience. Also, we wanted to control both the frontend and backend to tweak the user interfaces and capture crucial details, such as partial and intermediate submissions. The off-the-shelf solutions neither satisfied all our requirements nor supported TPM functionalities. Hence, we built the study environment from scratch using open-source components.

We used the SurveyJS library [5] to build a survey app for the questionnaires. We then integrated the self-hosted version of the Judge0 online IDE [3] with a TPM emulator [21] in the backend. Please note that, unlike real TPMs, the emulators do not include manufacturer-certified keys in the endorsement hierarchy. However, they are a helpful utility to test TPM functionality without having the hardware and in remote studies. We supported coding with widely used TPM libraries: tpm2-tools [7], IBMTSS [2], go-tpm [1], and wolfTPM [12].

The backend also contained a MongoDB database that collected responses to the tasks and questionnaires. We bundled the survey app, IDE, TPM emulator, and the database into a docker image and hosted it on our servers.

H Interview script

We present a skeletal structure of our interviews along with some of the questions in this section. Please note that, for each participant, we had created prompts using their code snippets and questionnaire responses that we found to be worth discussing in-depth. We used probes based on them, which differ for each participant and cannot be generalized; hence, we excluded them from the script presented below.

H.1 Introduction

Mutual introduction and reminder about the study

Ice breaker

- What do you use TPMs for?
- What are the common use cases of TPMs for you?
- How technology like TPM is contributing the field of security?
- Where is TPM useful and where is TPM not useful?

Confirming ecological validity

- How did you feel about the study environment and logistics?
- What about your prior experience or familiarity with such studies involving coding tasks, mainly the IDE?
- Are there any troubling components in our study that you want to highlight?
- Do you have any suggestions for improvement?

H.2 Task- and questionnaire-specific observations

Task-specific

General approach to TPM programming:

- How did you figure out what commands to use?
- Describe your process when starting a task?
- How did you search for relevant information?
- What resources helped you get started with the tasks?

Task 1: Symmetric/Asymmetric encryption

- How did you pick the cryptographic algorithms for creating the keys?
- Could you describe any examples that might have helped you when choosing other parameters? e.g., key length.

Task 2 :Storing measurements

- How did you select which PCR to extend with the measurements of the configuration file?

Task 3: Securing secrets

- Could you describe how and where is this secret stored?
- Could you describe any other ways to complete this task?
- Knowing now that there are other approaches, do you see any advantages or disadvantages of your approach over others?

Task 4: Remote attestation

- How do you verify the quotes are good to be used for remote attestation?

- Can you describe quote verification process?
- How do you verify the quote was generated by a TPM and it has not been tampered with?
- How do you confirm that the quote is valid?
- How do you think the verification could be simplified?

Questionnaire-specific

Correctness and security

- How did you verify that the task conditions were met?
- Why do you think your answer is secure?
- How did you verify you answer is secure?
- How do you know the defaults (or chosen parameters) are secure?
- Did you do any extra checks or referred somewhere?

H.3 General discussion

- Why do you think you had to look for help outside the official documentation?
- Why do you think the non-official resources are more reliable and useful than official resources?
- How do you think TPM library documentation and TPM standards could be written to compliment each other?
- How do you think the library can be improved?
- How do you think developers can contribute to improve the library further?
- Is there anything that you want to tell us regarding your experience about the library?

Increasing security without decreasing usability: A comparison of various verifiable voting systems

Melanie Volkamer
Karlsruhe Institute of Technology
melanie.volkamer@kit.edu

Jonas Ludwig
Karlsruhe Institute of Technology
jonas.ludwig@student.kit.edu

Oksana Kulyk
IT University of Copenhagen
okku@itu.dk

Niklas Fuhrberg
Karlsruhe Institute of Technology
niklas.fuhrberg@student.kit.edu

Abstract

Electronic voting researchers advocate for verifiable voting schemes to maximise election integrity. In order to maximise vote secrecy, so-called code-voting approaches were proposed. Both verifiability and code voting require voters to expend additional effort during vote casting. Verifiability has been used in actual elections, but this is not the case for code voting due to usability concerns. There is little evidence from empirical studies attesting to its usability. Our main contribution is to extend an existing verifiable voting system (used for real world elections) with a code-voting approach to improve the system's security properties. We minimise voter effort as corresponding QR codes are scanned instead of requiring manual code entry. We conducted a user study to evaluate the general usability of this proposal as well as its manipulation-detection efficacy. In particular, we found that extending the considered verifiable voting systems with code-voting approaches to enhance vote secrecy is feasible because we could not observe a significant decrease in general usability while manipulation detection improved significantly.

1 Introduction

The pandemic caused an increasing number of organisations to contemplate vote casting over the Internet for secret polls, and governments are also considering online elections. However, this requires deployment of various cryptographic techniques to ensure vote secrecy and election integrity. One of these techniques is *individual verifiability*. It allows voters to verify that their vote is cast correctly and also that their vote has been

recorded as cast. Another one is universal verifiability which provides strong cryptographic proofs, that enable independent third parties to verify that the final tally correctly reflects all recorded votes. Together, they facilitated detection of attacks on the election's integrity and, in the absence of detected manipulations, permits strong guarantees of election integrity. The level of achieved vote secrecy depends on the verifiable voting scheme being in place. However, most verifiable voting schemes do not defend against a compromised voting client (i.e., the voter's laptop, smartphone or vote-casting application) that can violate vote secrecy despite the presence of verifiability.

From a usability perspective, individual verifiability is particularly challenging, as this requires voters themselves to undertake extra steps in addition to casting their vote. Therefore, it is not surprising that a range of usability and manipulation-detection efficacy studies have been carried out, e.g. [1–17]. Yet, the studied verifiable voting schemes rely on the trustworthiness of voting clients with respect to vote secrecy.

Our paper focuses on the verifiable voting system used in Switzerland¹ for elections and referenda. The usability of this system and corresponding voting materials — in general, as well as with respect to its efficacy to enable voters to detect manipulations — was evaluated and improved by [16, 17]. The Swiss system assumes that attackers cannot manipulate voters' devices nor the vote-casting application. If this assumption does not hold, vote secrecy could be violated.

One way to address the question of voting client trustworthiness is *code voting*. This means that voters cast their vote by entering so-called voting codes, which are uniquely assigned and delivered to each individual voter: one per voting option. Because the vote casting device cannot map the voting code entered to any of the options, the assumption of trustworthy voting clients is no longer required to ensure vote secrecy. Code-voting schemes have already enjoyed attention from

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022, August 7–9, 2022, Boston, MA, United States.

¹Switzerland is one of the few countries to allow Internet Voting for political elections and referenda. As a direct democracy, it holds several elections/referenda per year. Most of their elections and referenda are based on simple ballots (m out of a small number of n options).

security researchers, e.g. by [18–22]. However, these schemes have not been used for real elections. Usability concerns are often voiced by election officials due to the additional complexity of handling these codes. To the best of our knowledge, only two user studies were conducted in which code-voting approaches were included. In [23], three different code-voting approaches were evaluated in a within-subjects study. None of these approaches provided any means of verifiability. In [24], the authors compared three vote casting approaches (each with a different security level), including one approach with voting codes and a confirmation code to enable individual verifiability. The within-subject study mainly evaluated acceptance of the three approaches. While the authors also report on System Usability Scales (SUS) for each approach, the code-voting and verifying approach responses might have been influenced by the fact that participants used two less secure and also less complicated approaches beforehand. Note, the efficacy in enabling voters to detect manipulations for code-voting based verifiable schemes has not been evaluated by any user studies, yet.

In this paper, we make two proposals to improve the security of one concrete implementation of a verifiable voting scheme: The Swiss voting system. In essence, first, we extend the Swiss verifiable system by code-voting, thus improving it towards vote secrecy (*proposal-code-voting-with-QR-codes*). To address potential usability shortcomings, we propose that voters are issued with QR codes, so that they can use a camera-equipped device, most likely their smartphone, to cast their vote. These QR codes also contain the so called initialisation code. This allows us to use longer initialisation codes compared to those voters need to manually entered in the original system. Thereby, we also improve the security of voter authentication compared to the original Swiss voting system. Second, we propose a variant of the Swiss system (*proposal-standard-voting-with-QR-codes*) that does not rely on code voting but still uses QR codes for the so called initialisation code. This second proposal is equally vulnerable to vote secrecy violations from malicious voting clients as the original system. However, it provides more evidence with respect to voter authentication as compared to the original system. Thus, from a security perspective, our first proposal out-performs the second, and both out-perform the original system and thereby also its improved versions from [16, 17]. Note, from a security perspective, our first proposal also out-performs the approaches evaluated in [23] (which uses code voting, but does not provide any means to verify).

Besides proposing these two improvements to the Swiss verifiable voting system, the goal of this paper is to report on the evaluation of both systems in terms of general usability as well as in terms of manipulation-detection efficacy. We compared our results with those of the original system reported in [16] and to relevant related work. We developed a study protocol which facilitated remote participation, i.e., study materials incl. the voting materials was sent through the post, because Covid-regulations did not permit face-to-face user studies. Using this protocol, we conducted a user study

– consisting of two between-subjects experiments – with 139 participants. The first experiment evaluated the general usability of both proposals. The second experiment evaluated their manipulation-detection efficacy.

In summary, our contributions are as follows²: (1) We make *two proposals* to improve security of the Swiss voting system. (2) We evaluate the *general usability* of these two proposals. We compare our results in comparison to those reported by Kulyk et al. in [16] for the original system. We did not detect a decrease in general usability of our proposals as compared to the original system. The average SUS scores are compared to those reported in relevant related work. (3) We evaluated the *efficacy* of our two proposals with respect to manipulation detection. Both performed significantly better in this as compared to the original system (using the data from [16]). We also compare the manipulation-detection efficacy of our proposals with those reported in relevant related work. (4) We propose a study protocol which allows user studies to be conducted *remotely* in the context of electronic voting. We also discuss lessons learned.

With our research, we are in particular the first to study the efficacy in enabling voters to detect manipulations for a code-voting based verifiable schemes. While there is also room for further improvements to enable even more voters to detect manipulations, our research shows that the Swiss election officials should consider extending their system as proposed in this paper. In general, our research indicates that it is worth considering more QR-code-enabled code-voting for verifiable voting systems, because they can be implemented in such a way that their usability is comparable to the usability of extant systems while it is not required anymore to trust that voting clients are trustworthy.

2 Related Work

Several security analyses of electronic voting systems have found serious vulnerabilities, e.g., [26–28]. These results show the importance of verifiability. Verifiable voting schemes enable voters, candidates, and election officials to check whether or not the voting system has been manipulated. Unsurprisingly, the research community focuses on such verifiable voting schemes. Several were already being used for real elections and secret polls.

Several studies have evaluated the usability of verifiable electronic voting systems, e.g. [1–16, 24]. While most of these studies focus on one of five approaches that enable voters to verify, Marky et al. compared approaches with each other [15]. Most of the user studies reported good results for the general usability. The manipulation-detection efficacy results were mixed. In particular those studying actual systems in use have shown less optimistic results with regards to manipulation

²One of the two proposals (i.e. the proposal-code-voting-with-QR-codes) was presented in a work-in-progress paper together with its general usability evaluation [25] – but without the evaluation on its manipulation detection efficacy.

detection. This is the case for the Norwegian Internet voting system in [11], and for the Swiss voting system in [16, 17].

Some researchers studied user perceptions and mental models of verifiable voting systems. Distler et al. [14], for example, found that users felt less secure after having verified. Other works, e.g. [1, 9, 10, 13, 29], identified misconceptions which prevented study participants from verifying their vote. These studies have concluded not only that improvements need to be made in the usability of verifiable voting systems needs to be improved. They also stressed the importance of properly communicating the 'extra' steps needed to verify.

While verifiability is used to maximise the election integrity guarantees, an important building block for maximising vote secrecy is the use of voting codes. Here, voters receive an individual code for each voting option. These codes are usually sent via the national postal service. Instead of selecting their option on the screen, they enter the corresponding voting code. The security of code-voting schemes have already been studied, e.g., in [18–22]. The usability of code-voting schemes, however, has been the subject of only two studies, i. e. [23, 24].

The usability of three different approaches to implement a non-verifiable code-voting scheme have been evaluated in [23]. In a between-subject study with 18 participants, they used the SUS items to report on the usability of three approaches to enter voting codes: (1) manually entering the code, (2) scanning a corresponding QR code from a booklet of QR codes, and (3) tangible objects. The focus of the study was on the process of casting a vote, while other steps, such as voter authentication, were not part of the participants' task. Furthermore, there was also no verifiability in place. The mean SUS performance for the manual approach was 61.25, for the QR-code approach 84.02, and for the tangible objects 78.61. While the study has several limitations, the authors found that code voting as such can be usable. In particular they report that the QR-code approach significantly out-performs the manual approach in terms of usability. We were inspired by their work to further consider code voting based on QR codes. However, we wanted to propose an entire system that provides both voter authentication and verifiability – and not just consider vote casting alone. We thereby explored using QR codes instead of having voters manually entering codes, because QR codes have advantages not only in terms of usability but also in terms of security: QR codes enable voters to enter large codes much more efficiently and effectively. Larger codes can significantly improve the security of verifiable voting schemes both with code-voting (see our proposal-code-voting-with-QR-codes) and without code-voting (see our proposal-standard-voting-with-QR-codes).

Three voting schemes on different security levels were evaluated using a between-subjects study in [24]. One of their schemes employed a verifiable code-voting. The authors mainly studied the impact of explaining the need for the various additional security-related steps (besides clicking on the voter's preferred candidate) on user acceptance. Because of the need to explain each scheme, all participants interacted

with the three schemes in the same order. The authors conclude that although the verifiable code-voting scheme obtained a SUS value of only 67, the participants tended to prefer this less usable system to a more usable but less secure system. The authors also report that the SUS performance for the code-voting approach was significantly lower than the none-code-voting approaches. However, participants may have been biased when judging the verifiable code-voting scheme, having already interacted with two more usable approaches. The research focus of this study differed from ours (acceptance versus usability/manipulation-detection efficacy). In particular, the authors did not study manipulation-detection efficacy. Furthermore, the verifiable code-voting approach in [24] requires voters to manually enter voting codes while in our proposal, codes are entered by scanning corresponding QR codes.

3 Background

We begin by first explaining the Swiss online voting system, as we propose security improvements for it (see Section 4). We then summarise the usability improvements that have been proposed for this system by Kulyk et al. [16] for this system.

3.1 The Swiss Electronic Voting System

In the Swiss voting system, the process of casting a vote proceeds as follows (see also Figure 4 in the Appendix, which shows the underlying voting scheme): Voters receive an individual code sheet (usually called a polling sheet) via the postal service, containing one initialisation code, check codes for each voting option, one confirmation code, and one finalisation code – all being unique for each voter. It should be noted that Switzerland has no electronic ID system by which voters could be authenticated online. Therefore, the system generates an election-specific key pair for all voters. The private key is derived from the initialisation code.

To start the vote casting process, voters manually enter their initialisation code (which is provided to them on their polling sheet) by typing the corresponding characters in the corresponding field of the election webpage and then selecting their voting option using the election webpage. The election webpage then displays a check code with which voters are supposed to compare the code next to their voting option on their polling sheet. If the check codes match, the voter confirms the correct code by (again manually) entering the confirmation code. If the check code is incorrect (or no check code is displayed), the voter is supposed to complain. Finally, voters receive a finalisation code that should match the code on their polling sheet, as a confirmation, that their vote has been recorded. If this is not the case, again they are supposed to submit a complain. It should be noted here that the check code would be sufficient to verify. From an organisational perspective, however, it is recommended that there are two additional steps and codes, respectively, in order to allow the

system to be able to react to complaining voters; in this way, voters reporting issues can be offered an alternative voting channel (postal or in person). Furthermore, the communication between the voters' devices and the election infrastructure is secured on the transport layer using TLS.

Usability and Security Considerations. According to the requirements in [30], the initialisation must have at least 20 characters, each check code must have at least four digits, the confirmation code must have at least nine digits, and the finalisation code eight digits. Thus, in terms of general usability, the most error-prone task is to manually enter the initialisation code as well as the confirmation code. More information on the underlying scheme, e.g. how the election infrastructure³ computes the codes to be sent back in distributed manner and how the votes are tallied in a verifiable way, is provided at [31]. The Swiss voting system relies on the trustworthiness of the voting client (i.e., the voters' laptops or smartphones, and the vote-casting application) for ensuring vote secrecy. The voting system also assumes that the printers in charge of printing the polling sheets do not maliciously cooperate with the voting client to violate election integrity. For the second assumption to be realistic, printing presses are operated offline. We refer to this system including the voting materials and user interfaces as the '**original system**'.

3.2 Improvements and Study Reported in [16]

Two papers have proposed and evaluated usability improvements for the Swiss system [16, 17]⁴. Both approaches propose changes in the design of the polling sheet to a more step-by-step instruction and a reduction in the information provided on the election webpage. We base our research on the improvement from Kulyk et al. from [16] for the following reasons. First, Kulyk et al. studied manipulation-detection efficacy with respect to two different manipulation approaches, while Marky et al. studied only one. Second, the study design in [17] is less reliable with respect to manipulation-detection efficacy. The low reliability is due to: (a) In [17], manipulation had to be reported using a corresponding button on the election webpage, which is contrary to the adversary model that assumes a potentially malicious voting client; and (b) participants used the system twice, once without manipulation and then with manipulation, potentially making it much easier to detect a difference from the previous election, than if the last election is some months or even years ago. Third, the authors of [17] did not specify the details of their manipulations. Most importantly, they failed to describe the changes introduced to the user interfaces; thus it is not clear from the paper, how easy it was for participants to detect the manipulation. Fourth,

³The election infrastructure is a composition of several services conducted by independent parties. The details of their interaction are not relevant for the usability of the cast as intended verifiability functionality.

⁴The authors of [15] also studied an improved version of the Swiss system. However, the improved version is identical with the version studied in [17].

the data from Kulyk et al. [16] is available on the Internet⁵.

In [16], Kulyk et al. analysed the voting materials and the election webpage of the Swiss system through several brainstorming and feedback sessions with lay persons and various experts. Based on the issues raised in these discussions – particularly those that may prevent voters from detecting manipulations – Kulyk et al. proposed a revision of both the voting materials (see Fig. 2 in [16]) and the election webpage (see Fig. 3 in [16]). The voting scheme is as described in Figure 4 in the Appendix. It should be noted that, while their focus was on usability improvements to the Swiss system⁶, our proposals focus on security improvements and their potential implications for both general usability and the manipulation-detection efficacy.

Both the original system and the improved system were evaluated on two parameters: (1) the general usability based on the System Usability Scale (SUS) and (2) their manipulation-detection efficacy, i.e., whether voters could detect manipulations of their cast votes. The evaluation was based on a lab study of a total of 128 participants. Their study evaluated the general usability of the original system compared to the improved system ('general usability groups'). In addition, the study tested the manipulation-detection efficacy ('manipulation groups') by manipulating the votes cast during the study. Their research included two types of manipulations, both of which would enable attackers to change the intended vote to a vote preferred by the attacker, if undetected. The difference between the manipulations lay in the changes to the user interface: In the first manipulation type (called '*replace-manipulation*'), the attacker would show the check code which the attacker would obtain from the election infrastructure after having cast their own vote (which is different from the vote cast by the actual voter). This check code differs from the code that voters would expect. Thus, the attacker would need to hope that voters do not check whether the displayed check code is correct, because the interface says 'Continue by scanning the confirmation code'. In the second type of manipulation ('*remove-manipulation*'), no check code would be shown; instead, the voter would see a message confirming that their cast vote had been accepted by the voting system. Afterwards, the vote casting would continue – for both manipulation types – as described in the polling sheet.

Thus, Kulyk et al.'s user study consisted of six groups – three groups using the original system and three using their improved system. Participants were randomly assigned to one of the groups. All groups were told that the study goal was to evaluate the system's usability, and all participants were given the task of casting a vote. All participants carried out the following steps: First, they were given an information packet containing an informed consent form, general information about the study, role card (including which option to select) and actual voting materials (including the election letter and the polling sheet). After the participants had read this information, they

⁵https://secuso.aifb.kit.edu/downloads/voting_manipulations_2020.xlsx

⁶This is also the case for [17].

could use the study laptop to proceed with vote casting. The election letter explained to participants that they should contact the (study) support in case of any problems. After these initial orientation steps, the following steps differed depending on the situation. Those participants assigned to the 'general usability-groups' and those from the other groups who did not report the manipulation were asked to fill out a questionnaire. Those in the 'manipulation groups' were debriefed afterwards, i.e. they were informed about the manipulation and that the actual goal of the experiment was to study the manipulation-detection efficacy. Those who reported the manipulation were first debriefed and then asked whether they would be willing to continue the study and to complete the survey questionnaire. The study found that the detection rates for both types of manipulation were significantly higher for their improved system compared to the original system (76% versus 100% and 10% versus 43%). The SUS scores were very similar for both schemes with 79.9 for the original system and 80.9 for the improved scheme.

We refer to this system including the voting materials and user interfaces as 'system from [16]'.

4 Proposed Voting Systems

In this section, we describe the two extensions which we propose for improving the security of the Swiss voting system. Both extensions have been evaluated (see Sections 5 and 6).

4.1 Security Improvements with Voting Codes (proposal-code-voting-with-QR-codes)

The first proposed improvement is to extend the Swiss voting system using code voting. With this extension, the individual polling sheet from the Swiss system also contains one individual voting code per voting option on the ballot. The voting codes are different for each voter. Thus, voters must enter the voting code corresponding to their chosen option. The different independent parties building the election infrastructure together deduce the actual option from each cast voting code during the tallying. As the voting client cannot map the voting code to any of the options, there is no need to operate with the assumption of a trustworthy voting client⁷.

While this proposal increases the security level compared to the original Swiss system and to the improved system studied in [16, 17], the actual usage of voting codes made the voting process more complicated (and potentially less intuitive for voters). Furthermore, in order to achieve an adequate level of security, the codes need to be complex enough to prevent the adversary from guessing valid voting codes. However, the longer the voting codes are the more error prone and less usable it is to enter these codes manually. To address this shortcoming,

⁷Note, however, that one needs to ensure that the mapping of the voting codes to options for each voter remains secret to the adversary. Therefore it is important that the printers are operated offline.

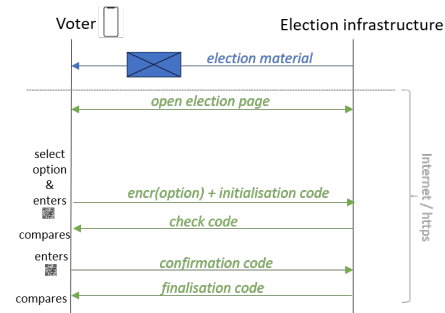


Figure 1: Vote casting with the proposal-code-voting-with-QR-codes

we propose that voters use their camera-equipped computer device, i.e. most likely their smartphones, to cast a vote by scanning a corresponding QR code (containing the complex voting code), as smartphones are now capable of scanning QR codes. While the use of QR codes as voting codes was already proposed by Marky et al. [23], the authors did not consider any means of voter authentication and verifiability.

With this proposal, entering the initialisation code is not needed as a separate step anymore: Given that QR codes can encode a lot more characters than needed for the voting code, they can contain both the initialisation code as well as the corresponding voting code. The QR code can actually contain an even longer initialisation code than the one used in the Swiss voting system as it does not need to be entered manually anymore. Moreover, the initialisation QR code can potentially encode the voters' actual cryptographic private key⁸. Hence, the security level for voter authentication is also increased. We also propose that QR codes are used to provide confirmation codes on the polling sheets. With this improvement, voters avoid having to enter any codes manually. The corresponding (simplified) scheme is depicted in Figure 1.

To integrate these ideas we revised the voting materials from [16] accordingly. The polling sheet is not one sheet anymore. It is a leaflet (see Figure 10 in the Appendix) with one voting card per each voting option (see Figure 5 in the Appendix). As we do not trust the voting client, scanning the QR code for the selected option does not require having the check code(s), nor the actual voting option present, nor the confirmation code, nor another option's QR Code. For similar reasons, the finalisation code must not be visible when scanning the confirmation code. Therefore, voting codes are presented on voting cards, the confirmation code is on a different page. The finalisation code is covered by a scratch field. Furthermore, the voting card should be placed on the inner page so as to ensure that the remaining voting cards are not too close by when scanning. The election webpage was revised as well (see

⁸The Swiss system is design for contexts in which voters do not possess any electronic ID. Instead an election specific key pair is generated. In the original system, the voter receives the initialisation code from which the actual private key is derived. With the QR code, the actual private key can be sent to voters.

Figure 6 in the Appendix). The entire voting system was developed and improved through feedback from participants. We refer to this first proposal, including the voting materials and user interfaces as ‘proposal-code-voting-with-QR-codes’.

4.2 Second Security Improvement (proposal-standard-voting-with-QR-codes)

The security of the original Swiss voting system as well as the system studied in [16, 17] can also benefit from using QR codes and using the camera-equipped smartphone in the vote casting process even in cases where there is no switching to a code-voting scheme. We propose to use QR codes for both the initialisation code and the confirmation code. In the Swiss voting system, both codes need to be entered manually. By using QR codes, more information can be transferred without decreasing the usability. As explained in the previous subsection using QR codes containing the initialisation code would increase the security for voter authentication as the actual private key can be included. To illustrate the changes from the original Swiss voting scheme, we provide a description of the corresponding scheme in Figure 7 in the Appendix. We revised the polling sheet (see Figure 14 in the Appendix) and the election webpage (see Figure 8 in the Appendix), accordingly. We refer to this system, including the voting materials and user interfaces as ‘proposal-standard-voting-with-QR-codes’.

4.3 Considered Manipulation-Types

As mentioned in Section 3.2, Kulyk et. al [16] examined two different types of manipulations in [16], both of which simulated an attack where the adversary attempts to cast a different vote on behalf of the voter. However, such an attack would not be possible in our proposal-code-voting-with-QR-codes, as adversaries would need to know the voting code for the option for which they want to cast a vote for – which is not the case by design of code-voting schemes.

Nonetheless, adversaries can still attempt to nullify votes by blocking the transmission of the voting code to the election infrastructure and manipulating the voting client so that the voter believes that their vote has been cast successfully. This type of attack is less attractive than replacing the vote, as the attacker would need to know how the voter intended to vote, in order to block only those votes considered “undesirable”. Otherwise the attacker might accidentally nullify the votes in favour of their preferred candidate. While this might be the case with high degree of certainty if the attacker knew enough about the voter (e.g. geography, age), removing too many votes would trigger suspicion if there was, for example, an unexpectedly low turnout of a certain demographic (e.g. voters living in an area historically known to support a particular political party). This is another, albeit small, advantage of our proposal-code-voting-with-QR-codes.

From a voters point of view such an attack would resemble the *remove-manipulation* investigated in [16] (see Section 3.2): after entering the voting-code, the election webpage would confirm that the check code entered was correct. Figure 11 (a) and (b) in the Appendix shows how the manipulated interfaces could appear given such a manipulation. As such, Step 4 confirms that the vote was cast. Furthermore, it indicates that the check code is correct and that the voter can continue their vote casting process. For proposal-code-voting-with-QR-codes, it is not possible for the adversary to show the finalisation code, as they are unable to send a valid voting code to the election infrastructure. Therefore, in order to avoid alerting the voter, the adversary would need to change these steps as well: Instead of asking voters to compare the displayed finalisation code with the code listed on the polling sheet, the manipulated voting client could ask voters to enter the finalisation code.

For the proposal-standard-voting-with-QR-codes, we assume –similar to the remove-manipulation described in [16] (see Section 3.2) – that the adversary forwards their altered vote to the election infrastructure i.e. the adversary would try to change the vote. Figure 11 in the Appendix (c) shows how the Step 4 interface would appear if such a manipulation were carried out. The remaining steps would be the same. If voters did not notice that no check code had been displayed and as such continued with the process, their altered vote would actually be stored and tallied. Manipulation of voters under the proposal-code-voting-with-QR-codes is potentially less obvious than for the proposal-standard-voting-with-QR-codes, as only one step is changed instead of two. As mentioned above, this type of attack would be more preferable for an attacker, but it cannot be applied under a code-voting-based scheme.

For the voter who experiences the voting process, their comparative perceptions for the proposal-code-voting-with-QR-codes and the proposal-standard-voting-with-QR-codes is shown in Figures 2 and 9 (see Appendix), respectively.

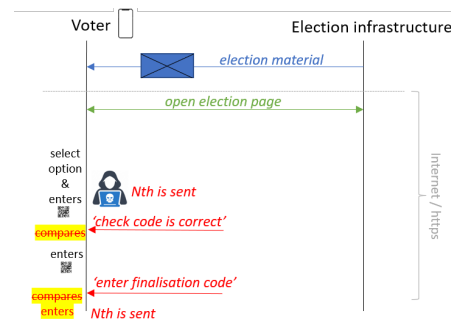


Figure 2: Manipulation for the proposal-code-voting-with-QR-codes.

5 Methodology

We first introduce our research questions and corresponding hypotheses. This is followed by a description of our study procedure. We then discuss ethical issues, how we meet data protection regulations, and how we recruited our participants.

5.1 Research Questions, Hypotheses

Our proposals improve the security level of the original scheme. We aim to answer the question, how these proposals perform with respect to the general usability as well as in terms of the manipulation-detection efficacy. Correspondingly, we define the following research questions:

RQ1 *How does each of our proposals perform in terms of general usability (measured as the System Usability Scale)?*

The study comparing the system from [16] and the original system did not find any significant difference between these two systems with respect to their SUS scores. However, both our new proposals include steps that might be less familiar to the users (i.e. scanning QR codes as compared to manual input of codes) and less comparable with traditional paper-based voting (i.e. using codes to enter a vote instead of choosing among the voting options presented in plain text on the screen). These new steps may have a negative effect on the general usability. We therefore define the following hypotheses:

$H_{1,1}$: The proposal-standard-voting-with-QR-codes has a significantly lower general usability than the one from [16].

$H_{1,2}$: The proposal-code-voting-with-QR-codes has a significantly lower general usability than the system from [16].

$H_{2,1}$: The proposal-standard-voting-with-QR-codes has a significantly lower general usability than the original one.

$H_{2,2}$: The proposal-code-voting-with-QR-codes has a significantly lower general usability than the original system.

Note, we decided against conducting statistical tests comparing to other relevant related work. As these studies have not made their data publicly available, the validity of such tests performed using only reported aggregate data (i.e. average SUS value) would be limited. Instead, we compare the descriptive data from related work in Section 7.

RQ2 *How do both proposals perform in terms of manipulation-detection efficacy (measured as the rate of participants detecting and reporting the manipulation)?*

We base our voting materials and election webpage for both of our proposals on the revision from [16], which had a significantly higher manipulation detection rate than the original system. Therefore, we expect that our proposals will also outperform the original system with respect to manipulation-detection efficacy. We define the following hypotheses:

$H_{3,1}$: The proposal-standard-voting-with-QR-codes has a significantly higher manipulation-detection efficacy than the original system.

$H_{3,2}$: The proposal-code-voting-with-QR-codes has a significantly higher manipulation-detection efficacy than the original system.

5.2 Study Procedure

In this subsection, we describe how we conducted the study⁹, i.e. the two experiments consisting of two groups each. The study was conducted in German, as were the voting materials and the election webpage. The text has been translated into English for the purpose of this paper.

First, in order to address RQ1, we conducted an experiment with two groups (i.e. for the proposal-standard-voting-with-QR-codes and the proposal-code-voting-with-QR-codes) to evaluate the general usability of the corresponding schemes. We then conducted the second experiment in order to evaluate the manipulation-detection efficacy – addressing RQ2. The study protocol was very similar for these two experiments. In the following paragraph, we describe the overall study procedure and explain where it differs for the two experiments.

Both experiments were announced as a remote study to evaluate the usability of an online voting system. The ballot of the simulated election contained four options. After having agreed to participate in the study, participants received the study materials in an envelope, either via postal service or from someone whom they knew. They received the study materials some days prior to the start of the experiment. Both experiments ran for two weeks each. The envelope had the following content:

- An information letter describing the study, the time frame, the other materials included, the procedures, and information that they can withdraw their participation at any time. In a footnote of the information letter, the link to the post-survey was included.
- A role card explaining who they were supposed to be in terms of the experiment and which option they were to vote for as part of the user study.
- An inner envelope with the actual voting materials, i.e.: (i) the official election letter from the election officials recommending them to first read the polling sheet before casting the vote. Furthermore, it mentions that in case of problems or questions they should call the (study) support; (ii) the polling sheet (and for the proposal-code-voting-with-QR-codes group, the cards with the voting-code); see in the Appendix Figures 10 and 14 for the polling sheets of both groups and Figure 5 for the voting cards with the voting codes.

⁹Figure 15 in the Appendix provides an overview of the study procedure. For a description of the different groups, see Figure 12 in the Appendix.

Participants were instructed to open the envelope and to read the information letter and the role card. Afterwards, they were supposed to open the inner envelope with the voting materials, then, to read the polling sheet before commencing their actual vote casting. So far, the process was the same for all four groups. The following steps differ and are therefore explained in separate paragraphs:

The **two groups studied to assess the general usability** could cast their vote according to the polling sheet. After having completed the vote-casting process, the election webpage displayed the link to the post-survey. This survey begins with information about the study and data collection. It contained the informed consent. Then, this survey asked the questions from the System Usability Scale (SUS) questionnaire and collected feedback on the system. The survey also included demographic questions. Finally, we asked participants to refrain from speaking with each other about this study until after they had completed their participation tasks.

The **two groups to study the manipulation-detection efficacy** received the manipulated interfaces as described in Section 4.3. *In the case that participants did not notice the manipulation or had noticed it but did not call the (study) support*, they could just finish casting their vote. After they had completed the voting process, the election page displayed the link to the post-survey. This survey, first, provided information about the study and data collection. It included an informed consent form. Participants then received a debriefing. If they decided to continue with the survey, they were asked whether they detected the manipulation that they had read about in the debriefing text. Afterwards, feedback on the scheme was collected, and the survey included demographic questions. The question regarding detection of manipulation offered the participant three options: (1) I noticed it and I called on the phone the (study) support; (2) I noticed it but I did not call the (study) support; and (3) I did not notice the manipulation. In case the first option was selected in the survey, participants had to confirm this option by entering the number 22. Those who called the (study) support received this number on the phone after they had reported the manipulation they had observed. In case the second option was selected, participants were asked an additional open text question on why they did not call the (study) support. *In the case that study participants did notice the manipulation and called the (study) support*, the support person first asked about details of the problems they observed. The goal was to first ensure they actually observed the manipulation and to determine whether the person calling was in the proposal-code-voting-with-QR-codes group or in the proposal-standard-voting-with-QR-codes group. Afterwards participants were debriefed. If they decided to continue participating in the study, they were provided with instructions on where to find the link to the post-survey¹⁰ and with the number 22. The support person thanked the participant for taking part in the study and took note about the group

¹⁰The post-survey was the same as for those who did not call the (study) support.

membership and the time. In case the support person could not answer the call, this participants was called back as soon as possible. All telephone numbers were subsequently deleted.

5.3 Ethics, Data Protection, Recruitment

The study protocol was approved by the ethics committee of our university. Their requirements also include various legal issues such as compliance with data protection laws. The study materials given to the participants contained a telephone number and an email address allowing them to get in touch with us in case of general questions regarding the study or if they had any other unresolved issues. As we could not guarantee a 24/7 service, the two participants who were unable to not reach us were subsequently called back.

The study was announced as a user study intended to evaluate the usability of an online voting system. Participants in the experiment to test the manipulation-detection efficacy who called the (study) support in order to report the manipulation were debriefed on the phone, and it was explained to them that they could withdraw from the study if they desired and that if they withdrew their data would be deleted. Furthermore, every participant in this experiment who filled out the survey received the debriefing text (see Appendix 16 for the corresponding text) regardless of whether or not they had contacted the (study) support beforehand. The telephone numbers from study participants who called were deleted after the call.

Participants received a role card describing who they are for the study and which option on the ballot they should select. This approach was used to ensure that we did not gather information about participants' actual vote. Furthermore, all participants received the same credentials (per group), i.e. the same voting materials, and the study's election server did not store participants' IP addresses.

Participants were recruited in two different ways: Through public channels announcing the study as well as through a snowball method, asking those who agreed to participate to announce it to their friends and family. In the first case, we usually sent the study materials via postal mail to those who had agreed to participate. In the second case, we usually send the first contact person the study materials to distribute it further.

The study announcement included information about the study, an explanation why we need their postal address if they want to participate, how we treat their postal address in terms of confidentiality, as well as confirmation that by sending us their postal address they agreed that we send them the study materials. In particular, the information related to the participant's postal address was important in fulfilling the data protection requirements. Potential participants were also informed that even after having received the materials that they could withdraw from the study at any time without any negative consequences.

We first prepared the study materials. When all the envelopes had been sealed, we mixed them up so as not to know which person would be assigned to which group.

Experiments	Age	Gender
Data from [16]	34.34/15.54	66F, 62M
Our proposals, no manip.	40.45/16.43	40F, 40M
Our proposals, manip.	42/15.85	29F, 26M

Table 1: Demographic data for participants for age Mean/SD. In their study, Kulyk et al. [16], authors report demographic data for both groups together.

Afterwards, the addresses were added. The addresses were deleted once they were put on the envelopes. For the survey, we used SocSciSurvey which is GDPR compliant¹¹.

Following discussions with our data protection officer, we decided not to offer financial reimbursement to participants. This was also mentioned in the recruiting brochure. As we did not know our participants personally, and since they came from all over the country, the only way to collect payment details would have been via email or postal mail. The first option would have been questionable in terms of data protection (e.g., many people do not know how to encrypt emails), while the second option would have imposed extra burdens on the participants. We also regarded that the participants’ total time and effort in participating in our study to be much less compared to Kulyk et al. [16], where the participants had to come to the lab on a particular day. In our case, participants could participate from home and where flexible, could participate anytime within a two-week period. Participants in our pre-study used an average of 20 minutes, including phoning the (study) support to report the manipulation.

6 Results

While recruiting the participants for our study, we also sent out the voting instructions to 200 participants, 100 of which for the first study (which involved no manipulations) and 100 for the second study (involving manipulations).

Eventually, a total of 135 people completed their respective study, 80 of which in the non-manipulation study (evaluating RQ1) and 55 in the study where their vote has been manipulated (evaluating RQ2). Table 1 shows the demographics of the participants of both of our studies alongside a comparison of the participants in Kulyk et al. [16]. The results for the two research questions are explained in the following two subsections¹² before briefly summarising the feedback we received. All statistical calculations for our hypotheses are performed using *R* packages “stats” and “rstatix”.

¹¹For their data protection policy see <https://www.sosciurvey.de/en/data-protection>.

¹²Figure 13 in the Appendix shows the overview of the results.

6.1 RQ1 - general usability

The mean values of the SUS score were 84.1 for proposal-standard-voting-with-QR-codes and 82.2 for proposal-code-voting-with-QR-codes, which corresponds to the grade between ‘good’ and ‘excellent’ according to Bangor et al. [32]. This is comparable to the scores of the original system and the system from [16] (mean values of 79.1 and 80.9 respectively). Figure 3 furthermore depicts the distribution of the SUS scores (as boxplot) for all the four systems. The Mann-Whitney tests¹³ failed to confirm the hypotheses $H_{1,1}$, $H_{1,2}$, $H_{2,1}$, $H_{2,2}$ (p-values of .658, .739, .932 and .975 respectively)¹⁴. We also calculated the effect size of the differences between the systems studied by Kulyk et al. [16] and our proposals, showing small effect sizes of these differences (see Appendix). Thus, for the general usability, we were not able to detect any difference between our proposals and those evaluated by Kulyk et al. [16]. While we acknowledge that this finding by itself is not a proof that there is no such decrease, we can conclude that it is at least unlikely that such a decrease, if at all present, is significant.

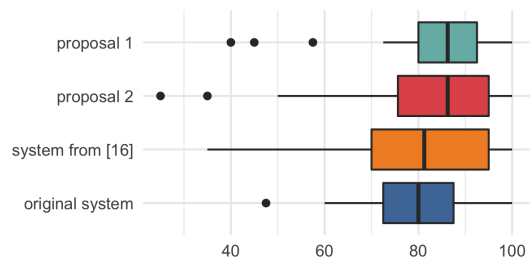


Figure 3: Boxplots of SUS scores. Proposal 1 is proposal-code-voting-with-QR-codes. Proposal 2 is proposal-standard-voting-with-QR-codes.

6.2 RQ2 - manipulation-detection efficacy

In our study, 22 participants called the (study) support to report the manipulation (11 in the proposal-standard-voting-with-QR-codes group and 11 in the proposal-code-voting-with-QR-codes group). This number is deduced from the survey. Table 2 shows the distribution of participants who reported the manipulation. While only 10% of the participants noticed the manipulation using the original system, the detection rate was at the same level for the system from [16], the proposal-standard-voting-with-QR-codes and the proposal-code-voting-with-QR-codes. Fisher’s test shows a significant difference between the original system and the proposal-standard-voting-with-QR-codes ($p = .049$ ¹⁵, $OR = 0.178$, 95% CI [0, 0.807]) as well as between the original system

¹³We chose to use a non-parametric test because the distribution of scores did not resemble a normal distribution.

¹⁴For a complete statistical overview, see Table 4 in the Appendix.

¹⁵The p-values are reported after a Bonferroni adjustment.

and the proposal-code-voting-with-QR-codes ($p = .0404$, $OR = 0.168$, 95% CI [0, 0.765]), confirming both $H_{3,1}$ and $H_{3,2}$.

	not detected	detected
Original System	18 (90%)	2 (10%)
System from [16]	12 (57%)	9 (43%)
Proposal 1	16 (59%)	11 (41%)
Proposal 2	17 (61%)	11 (39%)

Table 2: Manipulation detection rates. Proposal 1 is proposal-code-voting-with-QR-codes. Proposal 2 is proposal-standard-voting-with-QR-codes.

In the survey, 17 participants (8 in the proposal-standard-voting-with-QR-codes group and 9 in the proposal-code-voting-with-QR-codes group) reported that they had detected the manipulation but decided not to contact the (study) support. In an open text question field, they were asked to explain why they had decided not to call the (study) support. Their responses were analysed via open coding that was done by two of the paper authors. The answers were first coded independently and then discussed between the two authors. We identified three different types of answers (i.e. three different codes): Plausible reason, not critical, and mistake. The mapping of quotes to answers is provided in Table 5 in the Appendix¹⁶. Our findings can be summarised as follows: Five of the participants named a *plausible reason* for not contacting the (study) support, such as not wanting to disturb the support person by calling them at a late hour, or being outside of the country with high fees for international calls. Eight of the participants reported feeling that the error they noticed was *not critical*, e.g. they believed that their vote had been cast successfully despite of a check code because the voting website told them so. Three of the participants believed that the error was due to either their own *mistake* or a *mistake* made by those running the study (e.g. error in the voting materials).

We concluded that the five participants with *plausible reason* did not contact the study examiner because of the study setting but that they would probably have done so in a real election. As such, we assume that voters are more motivated to report discrepancies if the integrity of their real vote depended on it (hence, they would be prepared to exert more effort than in the study setting), and that ideally they would be aware about the availability of a reporting hotline (thus avoiding the situations where the voters are reluctant to call because of the late hour). Hence, if we count these ‘plausible reason’ participants as having detected the manipulation, the numbers would be higher with 13 out of 27 (48%) for the proposal-code-voting-with-QR-codes and 14 out of 28 (50%) for the proposal-standard-voting-with-QR-codes. As we are aware that our interpretations are subjective and in so far as these higher numbers are based on self-reported data, we only considered those participants who actually called in when conducting the above hypothesis test.

¹⁶We translated them using forward-backward translation

6.3 Feedback on the proposed schemes

We received a lot of positive feedback, including feedback from participants in the manipulation-experiment. In particular the participants stated that the instructions they received were clear. There were several suggestions for small improvements: i.e. removing the check-list icons on the right side, as this was not necessary to complete vote casting; using a larger piece of paper to have everything on a single page / a larger font size; and rendering the election URL as a QR code. In particular, several participants who did not detect the manipulation recommended that the instructions state more clearly the need to follow each step precisely, some also recommending that more explanation be provided as to why this was important.

7 Discussion

Our study shows that it is possible to improve existing (verifiable) voting systems (in particular the system used in Swiss elections) to provide enhanced security guarantees while we did not detect a decrease with respect to the general usability. The security advantages of the proposal-code-voting-with-QR-codes compared to the original system are three-fold: (1) security improvements on the scheme level (i.e., better guarantees with regards to vote secrecy), (2) fewer incentives for the adversary to attempt vote manipulation by targeting the voting client (as even if the attack were successful and were not detected by the voter, the adversary would have only managed to block votes as opposed to replacing them with a vote for another option – as it would be possible using the original system¹⁷, and (3) significant better manipulation-detection efficacy.

The proposal-code-voting-with-QR-codes also shows a high average SUS score (with 84.1 being considered between ‘good’ and ‘excellent’ usability according to [32, 33]), which is inline with the results reported by Kulyk et al. [16, 17] (although their schemes did not use voting codes). Furthermore, these findings are inline with the results of [23]. The authors of [23] evaluated *non-verifiable* code-voting schemes. They report a mean SUS score of 84 when using QR codes as voting codes. The SUS score drops to 61 when entering voting codes manually. Similarly, in [24], the authors report for their code-voting verifiable system – in which voters had to enter voting codes manually – a SUS score of 67. Thus, it appears that code voting decreases the usability when codes need to be manually entered while the usability level is not affected when QR codes are used instead. This is also supported by the fact that our second proposal (proposal-standard-voting-with-QR-codes) in which QR codes were used to enter the initialisation code and the confirmation code obtained a SUS score of 82.2. Another possibility is that the pandemic indirectly enabled such high SUS scores, as there were many instances where scanning QR codes proved convenient. Our results for manipulation-detection efficacy confirms that the proposed improvements described by Kulyk

¹⁷(1) and (2) also hold for [16, 17].

		Code Voting	Verifiable	SUS (mean)	Efficacy	Voter Authentication	Study Type
Our paper	proposal 1	QR Codes	yes	82	41%	included	between
	proposal 2	no	yes	82	39%	included	between
[24]	approach 1	no	no	88	no	not included	within
	approach 2	no	yes	83	no	not included	within
	approach 3	manual	yes	67	no	not included	within
[23]	approach 1	manual	no	61	no	not included	within
	approach 2	QR Codes	no	84	no	not included	within
	approach 3	tangibles	no	79	no	not included	within
[17]	original scheme	no	yes	81	33%*	included	within
	improvement	no	yes	85	100%*	included	within
[16]	original scheme	no	yes	80	76%/10%**	included	between
	improvement	no	yes	81	100%/43%**	included	between
[15]	code-sheet***	no	yes	85	100%*	not included	between

Table 3: Overview of the properties of our own proposals and approaches from related work. Remarks: * Results are based on a unrealistic setting (see Section 3.2). ** Two different types of manipulations were evaluated, the second one is similar to the one tested in our paper; the other one is easier to detect. *** The authors evaluated five different types of verifiability techniques tested, only their code-sheet has similar properties to the verifiability techniques in place for our proposals.

et al. in [16] actually enabled voters to detect manipulations significantly more often compared to the original system – as we based our proposed systems on the improvements from [16]. For a security and usability comparison between our proposals and the most relevant related work, see Table 3.

While our findings indicate that our proposals perform significantly better than the original system, an attacker controlling the voting client or the vote-casting application still has a high success rate of manipulating the vote without being detected. While the detection rate is comparable to the rate reported for the corresponding manipulation type in [16], manipulation effort for the proposal-code-voting-with-QR-codes is less attractive for adversaries, as they can only delete votes but not replace/alter them. Whether the detection rate is sufficiently high is a question to be decided upon by the election officials on a case-by-case basis. This decision depends – besides other facts like the importance of the election – on how complaints are treated. This kind of risk assessment, in particular deciding whether the risks are acceptable as compared to paper-based voting, is necessary for any kind of technology used in elections [34].

In order to increase the detection rate as well as the usability, as future work, the received feedback should be applied. One improvement would be to ensure that the voting materials have clearer and more salient statements, alerting the voter about the importance of following the process from the polling sheet in detail and to stop their voting if they receive a response that is not as described in the polling sheet. Such statements

could be supplemented with videos demonstrating the process of vote casting. On top of that, further measures might be needed to both increase awareness of the importance of verifiability and to explain why the polling sheet can be trusted but not necessarily the information displayed on the screen. The development and evaluation of such measures, and their effect on voters’ trust in the election system (using, e.g., a questionnaire developed in [35]) is also an important direction of future work.

Study limitations: Our study has limitations similar to other user studies evaluating the general usability of electronic voting systems and the manipulation-detection efficacy in verifiable electronic voting (e.g., the user studies in [1–14, 24]). First, in these studies, participants cast a vote they were asked to cast in a mock election. Compared to actual elections, this vote is not as personally engaging for them. Second, participating in a study and, thus, agreeing to take time for it, may thus lead the participants to spend more time in reading the instructions compared to casting a vote in an actual election. Both of these aspects may have an effect – in particular on the manipulation-detection efficacy. However, introducing vote manipulations in an actual election in order to measure manipulation-detection efficacy would pose critical ethical and legal issues. Hence, some kind of mock election process will remain. Another limitations of all these studies testing manipulation-detection efficacy (including ours) is that we need to trust that those few participants who know each other have not informed others about the manipulation element. This is in particular important for our setting (using the snowball

method for the recruitment and conducting a remote study).

We studied one implementation of adversaries' attempt to make voters believe that their vote was cast as intended while their vote will not be considered in the tally. The details could vary, i.e., the text displayed to convince voters that their vote was submitted correctly, although the steps did not conform to the steps in the polling sheet. As future work, one could study the attack using a different text. Adapting existing formal methods for modelling security-critical processes in human-computer interactions [36, 37] could help towards developing a more systematic approach to identifying different implementations of such attacks.

We compared our results with those from [16]. They conducted a lab study while our study was conducted. The study from [16] has therefore a higher internal validity and a lower external validity than our remote study. We could not control for various factors (incl. whether and how long they spend on reading the material, whether they were alone, and whether they tried several times before calling the (study) support). One may argue that this is actually the same when enabling enabling online voting.

Limitations of the proposed systems: Despite its security advantages, code voting is limited with regards to the type of elections for which it can be used. As such, it is most suitable for approval voting, that is, elections with 1 out of n options. Even then, however, it is yet to be studied whether the system remains usable when the number of available voting options increases. Applying code voting to other voting rules, such as m out of n or ranked voting, is much less feasible. The decision on whether to apply our proposal-code-voting-with-QR-codes or any other system based on code voting, can only be made on the basis of a particular election. A further issue that needs to be addressed is adapting our proposals, as well as the original system, to meet the needs of visually impaired voters, which might be particularly challenging due to the reliance of these systems on paper-based materials that have to be distributed and read by the voters. Finally, aside from the risks addressed by our proposal-code-voting-with-QR-codes (namely, threats towards vote secrecy and vote integrity resulting from compromised voting clients), several other issues need to be addressed in order to make Internet voting feasible in practice. Such threats, including but not limited to voter coercion or vote buying, or general social engineering attacks, are well-known and acknowledged by both the academic community and election practitioners. Identifying ways to mitigate these threats is an important research topic. As such, one important recommendation (see also [38]) is to offer Internet voting as a secondary voting channel, encouraging voters to cast their vote on paper if they experience problems with the Internet voting system or if Internet voting is not easily accessible to them.

Lessons learned for remote studies: Although this is the first time that we have conducted a remote study, we had good experiences, particularly with respect to recruiting participants. However, there are also a few lessons learned that we believe

should be shared with the community: (1) Make it more clear to participants when the (study) support is available and that the contact person is affiliated with the research team. (2) Explain what to do, if the (study) support cannot be reached, e.g., providing an alternative channel. (3) Test the prototypes on a variety of devices, operating systems and web browsers. Participation should be restricted to settings in which such tests have been thoroughly conducted. (4) Carefully select the sending out of the materials and the study period. For the second experiment, which was to commence on January 2nd, we sent out the materials just before the Christmas holidays. We believe that predictable delays in the postal system may have caused lower turnout in our second experiment.

8 Conclusion

Verifiable voting schemes are the de-facto standard when considering online voting for political elections. At the same time, the verifiable voting systems in place can provide adequate vote secrecy only where the voting client is trustworthy. While this shortcoming can be addressed with code voting, such approaches are currently not considered, as the community and election officials are concerned about the usability implications. Prior to undertaking our own study, there was little evidence from empirical studies that could demonstrate general usability or a lack thereof. The effect of code voting on the manipulation-detection efficacy was also not known. Our study has shown that code-voting verifiable voting schemes are worth considering, as the cumbersome steps of entering voting codes manually can be replaced by easy-enough steps – i.e., scanning QR codes – without significantly reducing the usability, while enabling systems with higher security guarantees. In the concrete instance of the Swiss verifiable system, our first proposal (the proposal-code-voting-with-QR-codes) has the following advantages compared to the original system: the trust assumption regarding the voting client is not needed anymore, manipulating the election outcome is less attractive as votes can only be removed but not replaced/changed, and the tested manipulation was detected significantly more often. We also used the QR code scanning solution for the second proposal, the proposal-standard-voting-with-QR-codes. While this proposal is less attractive than our first proposal from a security point of view, the fact that this systems did also not significantly reduce the usability underscores the value of QR code scanning as a useful element to be integrated in a vote casting process – in particular if it increases the overall security of the voting system. Thus, our findings should encourage further research on combining QR-code-enabled code voting with verifiable schemes.

Acknowledgements

We would like to thank Reto Koenig and Philipp Locher for their participation in the discussion on technical aspects of

how code voting can fit into the cryptographic protocol of the Swiss system as well as on how to design the polling sheet of the proposal-code-voting-with-QR-codes in a way that the security properties of the scheme are not violated. This research was further supported by funding from the topic Engineering Secure Systems, subtopic 46.23.01 Methods for Engineering Secure Systems, of the Helmholtz Association (HGF) and by KASTEL Security Research Labs.

References

- [1] M. Bär, C. Henrich, J. Müller-Quade, S. Röhrich, and C. Stüber, “Real world experiences with bingo voting and a comparison of usability,” in *EVT/WOTE*, 2008.
- [2] J.-L. Weber and U. Hengartner, “Usability Study of the Open Audit Voting System Helios.” <https://www.jannaweber.com/wp-content/uploads/2009/09/858Helios.pdf>, 2009. [Online, February 16th 2022].
- [3] A.-M. Oostveen and P. Van den Besselaar, “Users’ experiences with e-voting: A comparative case study,” *Journal of Electronic Governance*, vol. 2, no. 4, 2009.
- [4] M. Winckler, R. Bernhaupt, P. Palanque, D. Lundin, K. Leach, P. Ryan, E. Alberdi, and L. Strigini, “Assessing the Usability of Open Verifiable E-Voting Systems: a Trial with the System Prêt à Voter,” in *ICE-GOV*, pp. 281–296, 2009.
- [5] F. Karayumak, M. M. Olembo, M. Kauer, and M. Volkamer, “Usability Analysis of Helios-An Open Source Verifiable Remote Electronic Voting System,” in *EVT/WOTE*, USENIX, 2011.
- [6] D. MacNamara, T. Scully, and P. Gibson, “Dualvote addressing usability and verifiability issues in electronic voting systems,” 2011. <http://www-public.it-sudparis.eu/~gibson/Research/Publications/E-Copies/MacNamaraSGCOQ11.pdf>, [Online, February 16th 2022].
- [7] D. MacNamara, P. Gibson, and K. Oakley, “A preliminary study on a DualVote and Prêt à Voter hybrid system,” in *CeDEM*, p. 77, 2012.
- [8] K. S. Fuglerud and T. H. Røssvoll, “An evaluation of web-based voting usability and accessibility,” *Universal Access in the Information Society*, vol. 11, no. 4, pp. 359–373, 2012.
- [9] C. Z. Acemyan, P. Kortum, M. D. Byrne, and D. S. Wallach, “Usability of voter verifiable, end-to-end voting systems: Baseline data for Helios, Prêt à Voter, and Scantegrity II,” *The USENIX Journal of Election Technology and Systems*, vol. 2, no. 3, pp. 26–56, 2014.
- [10] C. Z. Acemyan, P. Kortum, M. D. Byrne, and D. S. Wallach, “From error to error: Why voters could not cast a ballot and verify their vote with Helios, Prêt à Voter, and Scantegrity II,” *USENIX Journal of Election Technology and Systems*, vol. 3, no. 2, pp. 1–19, 2015.
- [11] K. Gjøsteen and A. S. Lund, “An experiment on the security of the Norwegian electronic voting protocol,” *Annals of Telecommunications*, vol. 71, no. 7-8, pp. 299–307, 2016.
- [12] C. Z. Acemyan, P. Kortum, M. D. Byrne, and D. S. Wallach, “Summative Usability Assessments of STAR-Vote: A Cryptographically Secure e2e Voting System That Has Been Empirically Proven to Be Easy to Use,” *Human Factors*, pp. 1–24, 2018.
- [13] K. Marky, O. Kulyk, K. Renaud, and M. Volkamer, “What Did I Really Vote For?,” in *ACM CHI*, p. 176, 2018.
- [14] V. Distler, M.-L. Zollinger, C. Lallemand, P. Roenne, P. Ryan, and V. Koenig, “Security–visible, yet unseen? how displaying security mechanisms impacts user experience and perceived security,” in *ACM CHI*, pp. 605:1–605:13, 2019.
- [15] K. Marky, M.-L. Zollinger, P. Roenne, P. Y. Ryan, T. Grube, and K. Kunze, “Investigating usability and user experience of individually verifiable internet voting schemes,” *ACM Trans. Comput.-Hum. Interact.*, vol. 28, no. 5, 2021.
- [16] O. Kulyk, M. Volkamer, M. Müller, and K. Renaud, “Towards improving the efficacy of code-based verification in internet voting,” in *VOTING Workshop at Financial Crypto*, Springer, 2020.
- [17] K. Marky, V. Zimmermann, M. Funk, J. Daubert, K. Bleck, and M. Mühlhäuser, “Improving the Usability and UX of the Swiss Internet Voting Interface,” in *ACM CHI*, 2020.
- [18] D. Chaum, “Surevote: technical overview,” in *Proceedings of the workshop on trustworthy elections (WOTE’01)*, 2001.
- [19] J. Helbach and J. Schwenk, “Secure internet voting with code sheets,” in *E-Voting and Identity*, pp. 166–177, Springer, 2007.
- [20] R. Joaquim, C. Ribeiro, and P. Ferreira, “Veryvote: A voter verifiable code voting system,” in *E-Voting and Identity*, pp. 106–121, Springer, 2009.
- [21] P. Y. Ryan and V. Teague, “Pretty good democracy,” in *Security Protocols Workshop*, vol. 17, pp. 111–130, Springer, 2009.

- [22] J. Budurushi, S. Neumann, M. M. Olembo, and M. Volkamer, “Pretty Understandable Democracy - A Secure and Understandable Internet Voting Scheme,” in *ARES*, pp. 198–207, 2013.
- [23] K. Marky, M. Schmitz, F. Lange, and M. Mühlhäuser, “Usability of Code Voting Modalities,” in *ACM CHI*, 2019.
- [24] O. Kulyk, S. Neumann, J. Budurushi, and M. Volkamer, “Nothing comes for free: How much usability can you sacrifice for security?,” *IEEE Security & Privacy*, vol. 15, no. 3, pp. 24–29, 2017.
- [25] O. Kulyk, J. Ludwig, M. Volkamer, R. E. Koenig, and P. Locher, “Usable verifiable secrecy-preserving e-voting,” in *6th Joint International Conference on Electronic Voting*, pp. 337 – 353, University of Tartu Press, 2021.
- [26] A. Aviv, P. Černý, S. Clark, E. Cronin, G. Shah, M. Sherr, and M. Blaze, “Security Evaluation of ES&S Voting Machines and Election Management System,” in *Proceedings of the Conference on Electronic Voting Technology*, EVT’08, (USA), USENIX Association, 2008.
- [27] S. Wolchok, E. Wustrow, D. Isabel, and J. A. Halderman, “Attacking the Washington, D.C. Internet Voting System,” in *Financial Cryptography and Data Security*, pp. 114–128, 2012.
- [28] D. Springall, T. Finkenauer, Z. Durumeric, J. Kitcat, H. Hursti, M. MacAlpine, and J. A. Halderman, “Security Analysis of the Estonian Internet Voting System,” in *CCS*, p. 703–715, ACM, 2014.
- [29] M.-L. Zollinger, E. Estaji, P. Y. Ryan, and K. Marky, “‘Just for the Sake of Transparency’: Exploring Voter Mental Models of Verifiability,” in *International Joint Conference on Electronic Voting*, pp. 155–170, Springer, 2021.
- [30] *Verordnung der Bundeskanzlei über die elektronische Stimmabgabe (VEleS) (July 1st 2018)*. Die Schweizerische Bundeskanzlei , 2018. <https://www.fedlex.admin.ch/eli/cc/2013/859/de>, [Online, February 16th 2022].
- [31] “Specification of the Swiss voting system,” <https://gitlab.com/swisspost-evoting>, [Online, February 16th 2022].
- [32] A. Bangor, P. Kortum, and J. Miller, “Determining what individual sus scores mean: Adding an adjective rating scale,” *Journal of Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.
- [33] A. Bangor, P. T. Kortum, and J. T. Miller, “An Empirical Evaluation of the System Usability Scale,” *International Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.
- [34] L. F. Cranor, “In search of the perfect voting technology: No easy answers,” in *Secure Electronic Voting*, pp. 17–30, Springer, 2003.
- [35] C. Z. Acemyan, P. Kortum, and F. L. Oswald, “The Trust in Voting Systems (TVS) Measure,” *International Journal of Technology and Human Interaction (IJTHI)*, vol. 18, no. 1, pp. 1–23, 2022.
- [36] L. J. Osterweil, M. Bishop, H. M. Conboy, H. Phan, B. I. Simidchieva, G. S. Avrunin, L. A. Clarke, and S. Peisert, “A Comprehensive Framework for Using Iterative Analysis to Improve Human-Intensive Process Security: An Election Example,” 2017.
- [37] M. Bishop, M. Doroud, C. Gates, and J. Hunker, “Attribution in the future internet: The second summer of the sisterhood,” *The Institute Ecole Supérieure en Informatique Electronique et Automatique, Laval, France 5-6 July 2012 Edited by*, p. 63, 2012.
- [38] C. of Europe, “Recommendation cm/rec(2017)5[1] of the committee of ministers to member states on standards for e-voting.” https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680726f6f#globalcontainer, last visited 24.05.2022.

Appendix

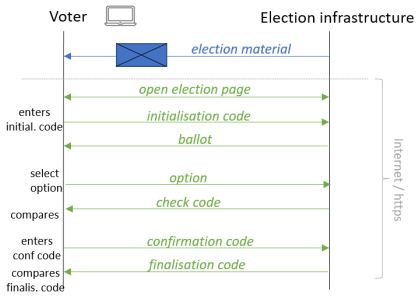


Figure 4: Vote casting with the Swiss voting scheme.

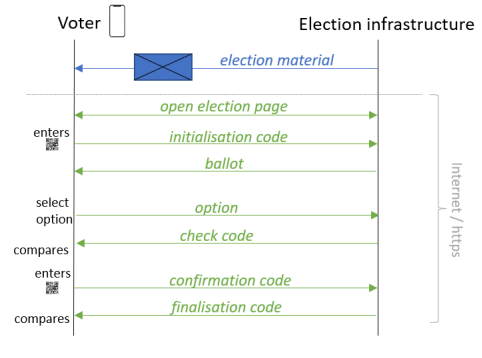


Figure 7: Vote casting with the proposal-standard-voting-with-QR-codes.

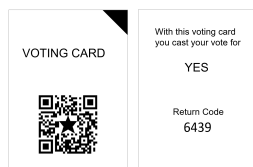


Figure 5: Voting Card (front and back side).

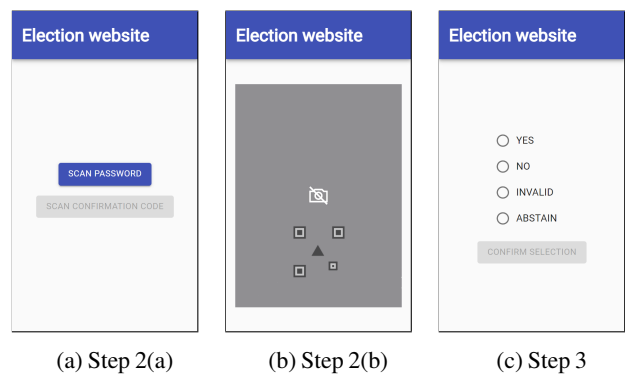


Figure 8: Voting webpage for the proposal-standard-voting-with-QR-codes, only displaying steps that are different from Figure 6.

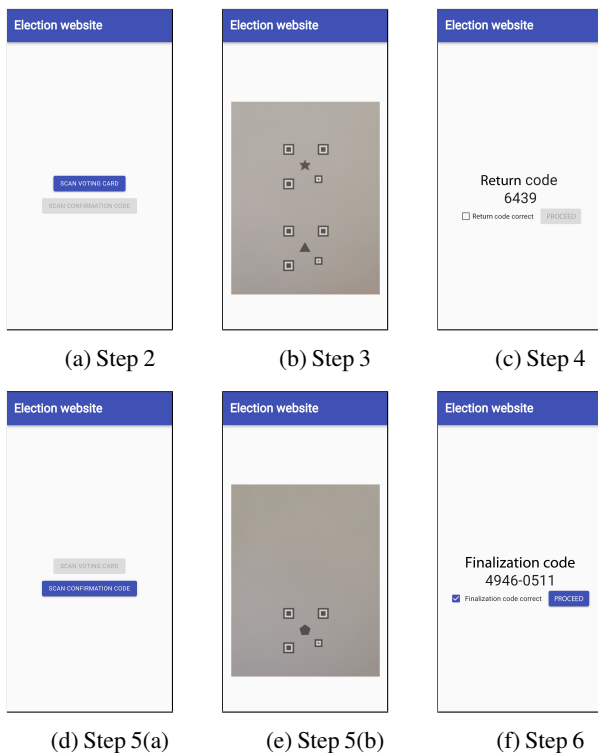


Figure 6: Voting webpage for proposal-code-voting-with-QR-codes.

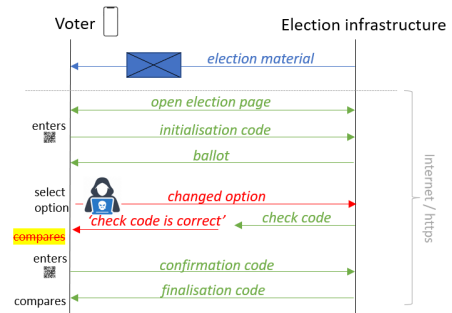


Figure 9: Manipulation for the proposal-standard-voting-with-QR-codes.

Polling Sheet

SUPPORT 0800 99 88 66

Before you start: This voting card allows you to participate in the referendum on the following topic:

Do you want to accept the initiative **"For responsible business – protecting human rights and the environment"**

To accept the popular initiative, vote **YES**, to reject it, vote **NO**. You are also able to **ABSTAIN** or **INVALIDATE** your vote.

For each of the four options, you have received a voting card with a QR code in your voting material.

In the event of problems or irregularities, only call the telephone number provided at the top of these voting instructions!

You are now able to start the voting procedure. Open the inner side of these voting instructions and start with **Step 1. Selection**.

(a) front

5. Confirmation: Now, click "Scan confirmation code" on the election website. Scan the code below.

CONFIRMATION CODE

6. Finalizing: The finalizing code is shown on the election website. **If this is not the case, immediately contact the support at 0800 99 88 66!**

To reveal the finalizing code below, scratch it with a coin or your finger.

FINALIZING CODE

4946-0511

Check if the code matches the code on the election website. **If the code does not match, contact the support immediately.**

Confirm the match on the election website. If this is the case, casting of the vote is complete.

Missing Finalizing code

Wrong Finalizing code

Vote casting complete

(d) back

1. Selection: Decide on one of the voting options and place the **corresponding voting card** onto the right side of this leaflet. Place the highlighted corner in the top right.

To avoid accidental scanning, return the remaining voting cards into the envelope.

1. Election website: Open the election website on your smartphone: **2021.wahl-webseite.de**

3. Vote: On the election website, click "Scan voting code". To do so, **grant** the election website **camera access**. Scan both QR-Codes on the right side at the same time as depicted.

4. Check code: The election website now shows a **check code**. **If no check code is shown, immediately contact the support at 0800 99 88 66!**

Please **check** if the check code on the election website matches the code in the list above next to the option you chose. **If this is not the case, contact the support immediately.**

Return the remaining voting card to the other cards in the envelope, to avoid accidental scanning. Now confirm the match on the election website.

No code

Wrong code

Continue on the next page

(b) inner - left

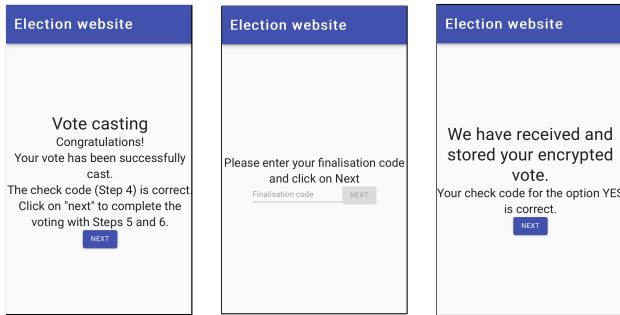
SUPPORT 0800 99 88 66

PLACE VOTING CARD HERE

QR code with a triangle marker

(c) inner - right

Figure 10: Polling sheet for the proposal-code-voting-with-QR-codes system.



(a) proposal-code-voting-with-QR-codes, Step 4 (b) proposal-code-voting-with-QR-codes, Step 6 (c) proposal-standard-voting-with-QR-codes, Step 4

Figure 11: Manipulation of the website for both proposal-standard-voting-with-QR-codes and proposal-code-voting-with-QR-codes.

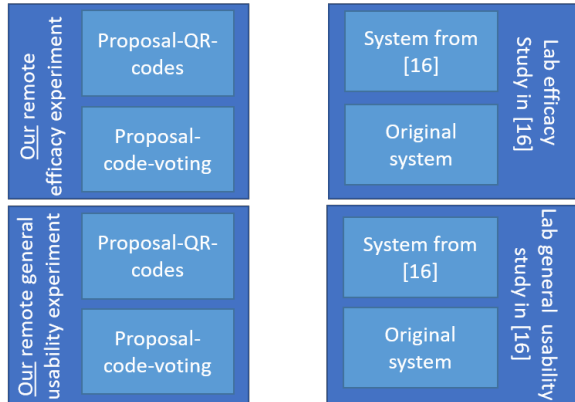


Figure 12: Overview of the considered groups.

Hypothesis	Estimate	Statistic	p	Effect size
$H_{1,1}$	-0.00	394	0.658	0.0508
$H_{1,2}$	-2.50	342	0.739	0.0830
$H_{2,1}$	-5.00	375	0.932	0.184
$H_{2,2}$	-5.00	306	0.975	0.25

Table 4: Comparison of general usability (evaluating RQ1) - p-values without adjustments for multiple comparisons.

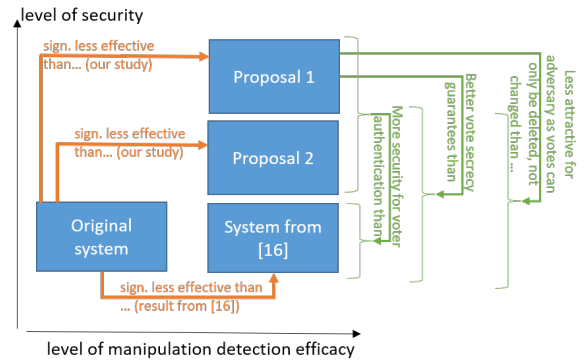




Figure 13: Overall result (the text on the arrows should be read in the following way: system A [-] <text on arrow > [->] system B means system A is <text on arrow > system B. e.g. the original system is sign. less effective than the proposal-code-voting-with-QR-codes). Proposal 1 is proposal-code-voting-with-QR-codes. Proposal 2 is proposal-standard-voting-with-QR-codes.

Polling Sheet  **SUPPORT 0800 99 88 66**

Before you start: This voting card allows you to participate in the referendum on the following topic:

Do you want to accept the initiative **“For responsible business – protecting human rights and the environment”**

To accept the popular initiative, vote **YES**, to reject it, vote **NO**. You are also able to **ABSTAIN** or **INVALIDATE** your vote.

 **In the event of problems or irregularities, only call the telephone number provided at the top of this polling sheet!**

1. Election website: Open the election website on your smartphone:
bern.wahl-webseite.de

2. Password: On the election website, click “Scan Password”. To do so, grant the election website **camera access**. Scan the QR-Code below to start the election procedure.

PASSWORD

3. Vote: Now decide on one of the voting options and confirm your choice.

Continue on the next page

(a) front



(d) back (blank)

4. Check code: The election website now shows a check code. **If no check code is shown, immediately contact the support at 0800 99 88 66!**

CHECK CODES

YES	6439
NO	8971
INVALID	4789
ABSTAIN	7526

Please **check** if the check code on the election website matches the code in the list above next to the option you chose. **If this is not the case, contact the support immediately.** Confirm the match on the election website.

5. Confirmation: Now, click “Scan confirmation code” on the election website. Scan the code below.

CONFIRMATION CODE

Continue on the next page

(b) inner - left

SUPPORT 0800 99 88 66

6. Finalizing: The finalizing code is shown on the election website. **If this is not the case, immediately contact the support at 0800 99 88 66!**

To reveal the finalizing code below, scratch it with a coin or your finger.

FINALIZING CODE

4946-0511

Check if the code matches the code on the election website. **If the code does not match, contact the support immediately.** Confirm the match on the election website. If this is the case, casting of the vote is complete.

Vote casting complete

(c) inner - right

Figure 14: Polling sheet for the proposal-standard-voting-with-QR-codes.

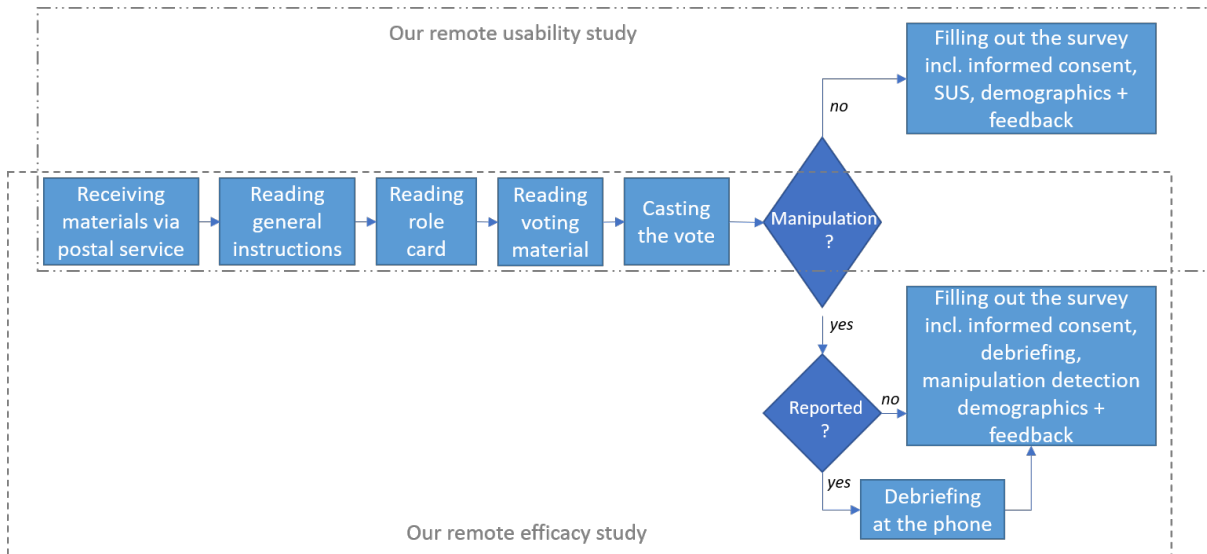


Figure 15: Study procedure for both the (general) usability and the efficacy study.

2. Debriefing

At this point, we would like to inform you that the vote casting contained a deviation. It was included intentionally by us. We simulated an attempt at vote manipulation.

An important aspect of the usability of online voting systems is to allow voters to reliably check if the voting is verifiably correct. In other words: the systems shall allow voters to reliably recognize manipulations. Not recognizing the deviation indicates poor usability of the evaluated online voting system.

Please consider that the study would not be valid if we had informed you about the discrepancy at the beginning of the study. This would have likely influenced you to further look for discrepancies than would be the case for a usual vote. Please also consider that the objective of this study was not to test you, but to test the system. Not recognizing the discrepancy is a symptom of insufficient usability of the online voting system, not inability the participants.

We hope you understand this approach. We would however understand if you decide to withdraw your participation after you received this information. If you continue with the study you will help to improve the usability of online voting systems. If you decide withdraw your participation, your information on this platform will not be evaluated.

If you would like to receive a personal debriefing, please contact the principal investigator ([redacted for review]).

I read the debriefing and was sufficiently debriefed.

Next

Figure 16: Debriefing of participants at the beginning of the questionnaire

Code	Recognised, but did not call, because:
not critical	I did recognise the manipulation as such. The website informed me that the check code was correct (without it being shown). I was satisfied with that.
not critical	I assumed the voting was correct even without another check code.
not critical	It was suggested that everything was in order. I would have wished that the mandatory adherence to the provided steps was indicated even more clearly. Such technical measures should be arranged.
not critical	I did not take it seriously enough.
not critical	I did not know that it was a manipulation. In the respective step, I received feedback from the platform that the last step was successful. Hence I did not call.
not critical	Too much effort.
not critical	Because it is just before 23 o'clock and I did not want to wake the study examiner.
not critical	I assumed there was a reason for it.
other	Unsafe.
mistake by user or examiner	I recognised that step 4 could not be performed as described and initially tried to correct a user error on my end by redoing the previous steps. By pressing the next-option everything proceeded until the end and, because I was attested a successful voting, I did not further question this – I'm just a sheep in such matters [smirking Emoji]. I wish you to gain lots of knowledge with this study.
mistake by user or examiner	I find the effort to call someone too much. Especially in the case of a fictional study. Moreover, I did find it very peculiar that both the role card and cover letter were printed double-sided with different salutations on each side. I initially did not see the second page and the cover letter and role card did not match, so I assumed that the study was flawed.
mistake by user or examiner	I was late with the test and assumed a flaw in the creation of the material.
plausible reason	I am currently abroad and a call would have been costly. I had planned to contact the support via E-Mail after completion.
plausible reason	Answering machine.
plausible reason	It was late at night, I did not want to call anyone at that time.
plausible reason	Recklessness. Time (nearly 23 o'clock). Assumption, that it was right anyhow.
plausible reason	I did call, but no one answered.

Table 5: Stated reasons participants recognised the manipulation but did not call the support

Presenting Suspicious Details in User-Facing E-mail Headers Does Not Improve Phishing Detection

Sarah Y. Zheng
UCL

Ingolf Becker
UCL

Abstract

Phishing requires humans to fall for impersonated sources. Sender authenticity can often be inferred from e-mail header information commonly displayed by e-mail clients, such as sender and recipient details. People may be biased by convincing e-mail content and overlook these details, and subsequently fall for phishing. This study tests whether people are better at detecting phishing e-mails when they are only presented with user-facing e-mail headers, instead of full e-mails. Results from a representative sample show that most phishing e-mails were detected by less than 30% of the participants, regardless of which e-mail part was displayed. In fact, phishing detection was worst when only e-mail headers were provided. Thus, people still fall for phishing, because they do not recognize online impersonation tactics. No personal traits, e-mail characteristics, nor URL interactions reliably predicted phishing detection abilities. These findings highlight the need for novel approaches to help users with evaluating e-mail authenticity.

1 Introduction

Phishing is a form of deceiving humans to obtain sensitive information in cyberspace. For example, people may receive e-mails that ostensibly come from genuine sources. Nearly half of all security breaches in 2021 involved some form of phishing [53]. Public campaigns and organizational policies have been warning people about it for years. Yet, individuals still receive and fall for them [27, 18, 3, 38]. With the steady growth of global digitization efforts, phishing appears to re-

main a powerful threat that is unlikely to decline [23, 19, 53]. It is therefore essential to understand why people fall for it.

Previous works suggest that people disproportionately infer e-mail legitimacy from e-mail message content and less so from details in typical *user-facing e-mail header information* (e.g., subject, sender e-mail address, sender display name, timestamp) [24, 57, 40, 60]. For instance, one study found that only the presence or absence of e-mail message features and none of the e-mail header-based features predicted whether participants processed e-mails as genuine or phishing [40]. Moreover, less self-reported attention to sender details was found to predict higher phishing susceptibility [57]. Qualitative studies with general users and IT experts also found that they primarily process e-mails and websites based on content relevance, rather than header details [24, 60]. Since e-mail messages can easily be manipulated and sender details in e-mail headers less so, e-mail headers often contain more reliable indicators of e-mail authenticity. It is thus conceivable that general phishing susceptibility could be driven by users' inattention to e-mail header information typically displayed in e-mail user interfaces. Remarkably, this hypothesis has not been tested empirically before.

This study aims to see if people are better at detecting phishing e-mails when they can only see e-mail header details. If so, simple changes in inbox user interfaces (UIs) that shift people's attention to e-mail headers could help to reduce phishing susceptibility, e.g., by highlighting an external sender's e-mail address. Participants are expected to be better at detecting phishing e-mails with suspicious source details when they can only see the e-mail headers, compared to when the full e-mails are displayed. Participants who are presented with full e-mails are expected to be at least as accurate as those who only see the e-mail message contents and subject lines, if people indeed are generally ignorant toward e-mail headers. This approach gives users the benefit of the doubt on their ability to recognize e-mail impersonation tactics.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

1.1 Contributions

- This is the first study to test whether people fall for phishing because they have overlooked suspicious details in common user-facing e-mail headers. Results show that most participants were not able to detect most of the phishing e-mails in the face of suspicious signals, regardless of whether they saw full e-mails or just e-mail header details. This strongly suggests that most people do not recognize deception tactics that are often used in phishing, even when they cannot be distracted by persuasive e-mail messages. This finding has important implications for tool developments to support users in phishing detection.
- This study used a realistic e-mail filtering task with rendered e-mails instead of screenshots. This allowed for more reliable measurements of phishing detection ability and tracking natural user interactions with e-mails, e.g., hovering over links.
- Participants performed a task where phishing detection was a secondary task.
- The sample was representative for age and gender of the UK population (N=252).
- It provides additional evidence that demographics, personality traits and privacy concerns do not reliably predict individual phishing susceptibility.

The next section discusses prior research on phishing and misinformation susceptibility, as well as anti-phishing interventions that motivated the design of the present study. Section 3 details the study’s setup and analysis approach. The results in Section 4 are structured around three key findings. Their implications are described in Section 5, before concluding the paper in Section 6.

2 Related literature

Prior studies explored online deceptions such as fake news and phishing through information processing theories, and factors that may explain individual differences in phishing susceptibility.

2.1 Inattention and online deception susceptibility

People use e-mail to communicate about relevant issues and not to detect phishing, so they may primarily read the e-mail message and use cues such as linguistic errors to infer an e-mail’s authenticity from [24, 15, 8, 40, 60] and overlook suspicious indicators in source details.

The recent surge of human fake news detection research provides an interesting parallel for understanding how people

process digital information and detect suspicious online content. Studies in lab settings as well as with real life Twitter data have provided evidence that user inattentiveness may drive belief in fake news [44, 43]. These findings suggest that susceptibility to fake news is mainly driven by “peripheral” cognitive processing.

Some theoretical models of phishing susceptibility align on the same notion of two distinct human information processing routes: 1. a systematic or “central” route based on careful assessment of phishing features that makes people less likely to fall for phishing, and 2. a less careful, “peripheral” route that increases people’s susceptibility to phishing [33, 56, 37, 17]. As people’s capacity for central information processing may be bound by their cognitive functioning, phishing susceptibility may be particularly related to markers of cognitive functioning (e.g., attention, memory) and not necessarily someone’s age [16].

It may neither be realistic to expect people to use central processing for all e-mails they receive, i.e., carefully checking all details of every incoming e-mail, when most of their e-mails are genuine. Indeed, lower phishing e-mail prevalence has been associated with worse phishing detection [49, 50]. Participants in Singh et al. performed a phishing training task in which either 25, 50 or 75 percent of the e-mails were phishing. Those exposed to higher phishing proportions detected more of the phishing e-mails, but were less precise in doing so. That is, they also marked more legitimate e-mails as phishing. Sawyer and Hancock found a similar effect with more realistic proportions of phishing attacks and termed it the “prevalence paradox”: participants who responded to 300 e-mails of which 1% were phishing, performed worse than participants who were presented with 5% or 20% phishing [49].

2.2 Measuring individual phishing susceptibility

There is arguably no single quintessential phishing e-mail with which people’s general phishing susceptibility can be measured [28]. Consequently, various phishing e-mail tactics have been described and used in phishing e-mail studies. For example, tactics based on Cialdini and Goldstein’s six principles of persuasion [62, 32, 39, 12] are: authoritativeness and urgency of e-mail messages [7], e-mails adapted to recipients’ contexts [22], and positive (e.g., monetary gain) versus negative (e.g., losing something valuable if not complying with the sender) e-mail content [14, 61].

The only common tactic used in online deceptions such as phishing seems to be impersonation, where adversaries manipulate information to create the impression that the digital content came from the claimed source. Still, many studies aimed at finding personal traits associated with higher phishing susceptibility relied on one type of phishing e-mail sent out to non-representative participant samples [63, 21, 4, 55, 36, 9, 1]. While these works all found associations between

various personal factors and engagement with the simulated phishing e-mails (e.g., clicking on the phishing link or entering personal details), these results need to be interpreted in light of their specific phishing types and sample contexts.

A general theory for online deception susceptibility, whether in the context of phishing or other scams, would predict higher susceptibility through *any* factor known to encourage inattentive information processing. In this view, personal traits such as age and gender are not the most reliable determinants of phishing susceptibility, since they do not necessarily indicate overall inattentiveness. Situational factors such as stress, distractions and user interfaces more likely affect people’s information processing capacity [11], and thence, phishing susceptibility.

The present study tests if presenting users only with commonly displayed e-mail header information increases their ability to recognize phishing e-mails. It uses a task design that addresses the aforementioned methodological limitations in five ways: 1. participants were presented with a diverse set of phishing e-mails, 2. it used a more naturalistic phishing proportion of approximately 17% compared to previous survey-based studies with 50% phishing [52, 13], 3. it rendered all e-mails from HTML instead of screenshots, which allowed tracking user interactions with the e-mails, 4. the participants sample was representative of the UK population, 5. it used a task context in which phishing detection was not the primary task, to avoid measuring biased phishing detection abilities [41].

3 Methods

This study examines if people’s phishing susceptibility is driven by inattention to suspicious e-mail source details. If so, merely presenting them with common e-mail header displays should improve their phishing detection ability. It also tests what factors (personal traits, e-mail characteristics, user interactions with URLs) could reliably predict phishing detection abilities. This section describes the experimental conditions, participants recruitment, task flow, e-mails selection, ethics approval and analysis approach. The task application, e-mail stimuli, scripts, processed data sets and supplementary materials can all be found on the OSF project page: https://osf.io/j9dm8/?view_only=212393f11473447d4bea74be547afbd17.

3.1 Study design

This study followed a between-subjects design with three e-mail display conditions: “Control”, “Headerless” and “Bodyless”. In the Control condition, full e-mails were displayed with typical *user-facing* e-mail header information and body content. That is, headers consisted of the subject line, sender name and e-mail address, recipient(s) e-mail address(es), time

it was sent and carbon copy (CC) e-mail address(es) where applicable. In Headerless, only the subject line and body content were displayed. In Bodyless, only said header information was displayed. See Figure 1 for an example full genuine e-mail display and labeling options in Control. Figure 2 shows the same e-mail as displayed in Headerless and Bodyless. E-mail headers of five of the curated phishing variants included sufficient information indicative of malice, see Table 1 and Section 3.4. Participants were randomly allocated to one of the three conditions.

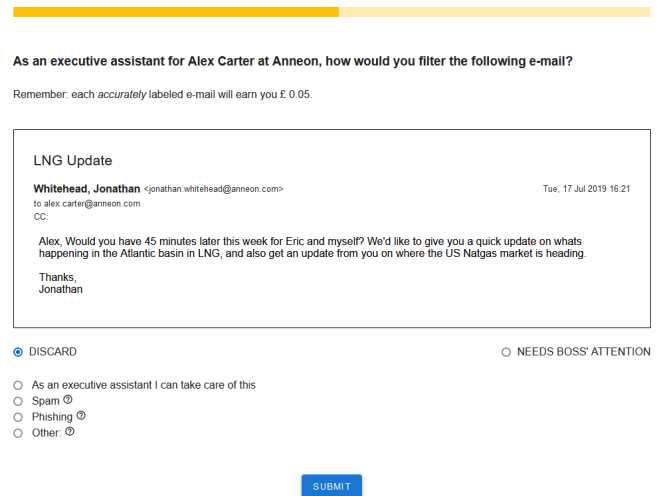


Figure 1: Example display of a genuine e-mail in the main task in the Control condition. The formatting of e-mail header information mirrored that of Gmail. E-mails were displayed one at a time in a solid black frame below the main task question (“As an executive assistant to Alex Carter at Anneon, how would you filter the following e-mail?”). To encourage participants to perform well in the task, a reminder of bonus payment for each correctly labeled e-mail was displayed as well. Initially, participants only see the two primary labeling options (“DISCARD” and “NEEDS BOSS’ ATTENTION”). When they select “DISCARD”, they need to select a specific reason for why they would discard the given e-mail. This step-wise approach is meant to avoid potentially priming users with the “spam” and “phishing” labels, and mimics people’s tendency to filter e-mails based on relevance first.

3.2 Participants

Eighty-four participants were recruited through Prolific for each display condition (total N=252). The sample was representative of the British population in terms of age ($\mu=46.46$, $SD=16.99$, range=18–68) and gender (50% male). The required sample size was informed by a desired statistical power of 0.80 with a 95% confidence interval for correctly interpreting a medium-sized difference in phishing detection accuracy between three groups. See supplementary materials on OSF

As an executive assistant for Alex Carter at Anneon, how would you filter the following e-mail?

Remember: each accurately labeled e-mail will earn you £ 0.05.

LNG Update

Alex, Would you have 45 minutes later this week for Eric and myself? We'd like to give you a quick update on whats happening in the Atlantic basin in LNG, and also get an update from you on where the US Natgas market is heading.

Thanks,
Jonathan

DISCARD NEEDS BOSS' ATTENTION

As an executive assistant for Alex Carter at Anneon, how would you filter the following e-mail?

Remember: each accurately labeled e-mail will earn you £ 0.05.

LNG Update

Whitehead, Jonathan <jonathan.whitehead@anneon.com>
to alex.carter@anneon.com
CC:

Tue, 17 Jul 2019 16:21

DISCARD NEEDS BOSS' ATTENTION

Figure 2: Example genuine e-mail display in Headerless (left) and Bodyless (right) conditions. No primary label selection was made yet, hence no further “DISCARD” reasons are displayed here.

for a complete overview of sample statistics. Two participants failed at least two out of three attention check questions in the questionnaires after the labeling task and were excluded from hierarchical regression analyses. This left 83 participants in Control and Headerless and 84 in Bodyless. All participants indicated that they responded honestly and to the best of their ability.

3.3 Task

After giving consent and solving a reCAPTCHA challenge to prevent bot responses, participants saw the task instructions and answered experiential questions on how much professional experience they have with executive assistance tasks and how many e-mails they receive on a daily basis. To cater to the secondary nature of security behaviors [41], the study was disguised as research for a new job application assessment.

Cover story Participants were told to imagine they are working as an executive assistant (EA) for their boss Alex Carter at a fictive petrochemical company called Anneon. Their main task was to filter e-mails for their boss by labeling each e-mail as “Needs boss’ attention” or “Discard” it. This distinction was made to further avoid giving participants the impression that the task was about phishing detection. Only when participants chose “Discard”, a specific discard reason had to be selected: “As an executive assistant, I can take care of this”, “Spam”, “Phishing” or “Other”. If they selected “Other”, they had to provide a brief free-text format explanation. The task instructions explained the scope of work of the executive assistant and the fictitious boss and showed example e-mails for what should be labeled “Needs boss’ attention” or “As an executive assistant, I can take care of this”. Tooltips were added to the “Spam”, “Phishing” and “Other” options with definitions of the respective labels, identical to the ones participants saw in the instructions. See Figure 1 and Appendix A for the full task instructions.

Task display E-mails were displayed one at a time, center-aligned, with a maximum width of 960 pixels and a 3 pixels thick solid black border with 10 pixels padding all around. The e-mails display order was randomized to avoid successively presenting phishing e-mails. All participants saw this pseudo-randomized order. The e-mail header format mimicked that of Gmail, since Gmail is the most used private e-mail provider. Labeling options were presented with radio buttons beneath each e-mail and a “submit” button to go to the next e-mail. Revisiting previously labeled e-mails was disabled by blocking backward navigation. A progress bar and the lines “As an executive assistant for Alex Carter at Anneon, how would you filter the following e-mail? Remember: each accurately labeled e-mail will earn you £0.05.” were continuously displayed at the top of the screen during the labeling task to encourage participants to perform well in the task. The task was developed as a single page application using Vue.js to minimize potential connectivity-related latency issues, and hosted on Google Firebase.

Post-task surveys After labeling all e-mails, participants filled in the short Big Five Inventory (BFI-S [31]) and Internet Users’ Information Privacy Concerns (IUIPC [34]) questionnaires. Three attention check questions of the form “Please select (option)” with 7-point Likert scale answer categories were added pseudo-randomly between the main questionnaire items. Finally, participants answered questions on their age, education level, gender, occupational status, income, estimated knowledge about cybersecurity (7-point scale, 1=“No knowledge at all”, 7=“Very knowledgeable”), likelihood to fall for phishing (“Very unlikely”, “Unlikely”, “Likely” or “Very likely”), frequency of receiving phishing e-mails (“Multiple times per day”, “Daily”, “Weekly”, “Monthly” or “Rarely”), whether they responded honestly and to the best of their ability in the study (“yes” or “no”) and whether the study could be improved in any way. Responses to the latter showed that participants found the task “straightforward” and “easy to navigate”. See Supplementary Materials for details.

Table 1: Displayed header details of the phishing e-mails set. For the full e-mail bodies, please refer to the Supplementary Materials.

e-mail	subject	sender name	sender e-mail	recipient e-mail
p1	URGENT	Sam Jones	sam.jones@anneon.com	alex.carter@anneon.com
p2	are you available?	Jeffrey Skilling	j.skilling@gmail.com	alex.carter@anneon.com
p3	From Mrs.Ameena Essa.	hillb439@gmail.com	hillb439@gmail.com	alex.carter@anneon.com
p4	Alex Carter	Barrister Paul Heywood	office.heywood@gmail.com	alex.carter@anneon.com
p5	Re: [Daily News Update Report] [Account Service] Microsoft account unusual sign-in activity: An order was issued grazie ordine on 06/11/2020.DLIBVCZA	Apple	ponco-gaming2443724@fajardoyandustone.com	mailtdsecure@m-lidscured.com
p6	Anneon File Cash Position Report - Oct19 (1).xlsx has been shared with you	SharePoint Online <no-reply@sharepointonline.com>	esrtn365@microsoft.com	alex.carter@anneon.com
p7	Action Required: Update your payment information now	Microsoft Online Services	no-reply@email.microsoftonline.com	alex.carter@anneon.com
p8	ZOOM Conference Call - April 06, 2020 @ 8:30 - 9:15am	anneon.com ZoomCall	zoom@anneon.formidable.it	alex.carter@anneon.com

3.4 E-mail stimuli

To present participants a realistic proportion of genuine versus phishing e-mails, 47 e-mails were selected of which eight (ca. 17%) were phishing and two were spam. Table 1 shows the header information of the phishing e-mails. See Supplementary Materials for the full e-mails set, including their bodies.

Legitimate e-mails The 37 legitimate e-mails were adapted from the Enron e-mails data set (as retrieved from <http://www.cs.cmu.edu/~enron/>), which contains actual business e-mails from former Enron employees. These were sanitized by substituting all original mentions of “Enron” with “Anneon” to prevent any potential response bias in the case of knowing the Enron scandal and bankruptcy from 2001. Personally addressed e-mails were made to target a gender-neutral executive named Alex Carter. Hyperlinks were replaced by links that opened a blank page in a new window. Despite the age of these business e-mails, they still resemble a natural source of electronic communication representative of corporate e-mails today, as they were mostly sent in plain text and concern realistic ongoing business topics.

Phishing e-mails Four of the phishing e-mails and the spam e-mails were adapted from actual e-mails received by the researchers. Four additional phishing e-mails were selected from various online sources with actual phishing examples. Participants in Bodyless were specifically expected to recognize the suspicious header details in phishing e-mails 1, 5, 6, 7 and 8. In Headerless, phishing e-mails 3–8 were expected to be detected by most participants. Phishing e-mail 2 was only reasonably detectable in Control.

Phishing e-mails 1 and 2 exemplified spear phishing e-mails, which could be detected through careful appraisal of the domain of the sender’s e-mail address. If these two phishing e-mails came from an `anneon.com` e-mail domain, they would have been virtually impossible for participants to detect. However, the domains were `anneon.com`, representing a homograph attack that should be detectable in Bodyless, and `gmail.com`, respectively. Phishing e-mails 3 and 4 were “Nigerian prince”-style phishing e-mails in which the sender tells a story about a diseased or deceased relative, after which they seek some form of financial help. The latter two e-mails could not reasonably be detected in Bodyless.

Phishing e-mails 5, 6 and 7 impersonated Apple or Microsoft and asked recipients to log in to secure their account, update payment details or to view a file that was shared with them through SharePoint. Phishing e-mail 5 contained mismatching sender details, a phishing sign that should be detectable in Bodyless. Phishing e-mails 6 and 7 came from suspicious sender e-mail domains that resembled Microsoft’s. Phishing e-mail 8 exemplified a Zoom phishing e-mail that surfaced during the COVID-19 pandemic in 2020 and also

used a non-sensible e-mail domain. Phishing e-mails 5–8 all contained malicious links, of which the original URLs were made visible on hover through the HTML link “title” attribute. Clicking on them would open a blank page in a new window. Together, this selection represented a diverse set of phishing sub-types.

3.5 Ethics

This study was reviewed and received approval prior to any data collection by the authors’ institution’s Research Ethics Committee. Participants were compensated at a recommended rate of £7.50 per hour and typically completed the study in 20–35 minutes. Each correctly labeled e-mail yielded a bonus payment of £0.05. Participants in the Headerless condition always received an additional bonus of £0.10 and participants in Bodyless always received an additional £0.15, given that phishing e-mails 1–2 and 2–4 were not reasonably detectable in the respective conditions. The average bonus payment across conditions was £1.25. All responses were collected anonymously.

3.6 E-mail characteristics

Twenty-five e-mail characteristics were computed for all 47 e-mails. This allowed for testing if participants consistently use common e-mail characteristics to infer e-mail authenticity from. For example, whether the sender e-mail address domain was the fictive company name (Anneon), whether the e-mail contained a personal greeting and how urgent the message sounds. E-mail body and subject valence, arousal and dominance (VAD) scores were based on the NRC dictionary [35] to give an indication of authoritativeness and emotional weight of each e-mail’s content. For each e-mail, the VAD-scores of all words in the subject line or body that were found in the dictionary were summed and divided over the total number of words in the subject or e-mail body, respectively. Language quality of e-mail body content was based on the number of linguistic errors as found with Beautiful Soup [48], divided over the total number of words in the e-mail body. An overview of all computed e-mail characteristics is included in the Supplementary Materials.

3.7 Analyses

Phishing detection by display condition. To test for the effect of e-mail display on each phishing e-mail’s detection proportion, χ^2 -tests were run with equal expected detection proportions for all display conditions under the null hypothesis. Detection proportion is the number of participants that labeled the given phishing e-mail as “phishing”, divided by the total number of participants in the respective condition. Phishing detection ability is computed as participants’ phishing detection precision (i.e., the proportion of e-mails the

Table 2: **Detection proportions for each phishing e-mail per display condition.** Boldfaced expectations are supported by the respective χ^2 -test for equal proportions.

e-mail	expectation	Control	Headerless	Bodyless	χ^2	p
p1	worst detection in Headerless	0.16	0.06	0.14	3.80	0.15
p2	highest detection in Control	0.11	0.04	0.13	4.52	0.10
p3	worst detection in Bodyless	0.87	0.79	0.16	42.5	<.001
p4	worst detection in Bodyless	0.80	0.75	0.12	43.4	<.001
p5	equal detection for all conditions	0.81	0.70	0.70	0.87	0.65
p6	equal detection for all conditions	0.29	0.06	0.19	12.1	<0.01
p7	equal detection for all conditions	0.46	0.55	0.51	0.58	0.75
p8	equal detection for all conditions	0.06	0.04	0.02	1.40	0.50

participant labeled as phishing, that indeed were phishing) and phishing detection recall (i.e., the proportion of all phishing e-mails that the participant detected). Using both metrics allows for more thorough estimates of phishing detection ability, given there were less phishing than legitimate e-mails. Participants with low phishing detection precision and/or low phishing detection recall are regarded as particularly susceptible to phishing. Additional analyses were run with both “Phishing” and “Spam” as true positive labels in the detection ability metrics.

E-mail characteristics regressions. To see if participants used “rule of thumb” tactics in deciding which e-mails had to be discarded as phishing, multiple regressions were computed in R to predict phishing detection proportions for all 47 e-mails, i.e., including false positives, per condition. E-mail characteristics based on sender information (e.g., if the company name was present in the sender e-mail domain) were not included in the Headerless model and body content features (e.g., body content dominance score) were not included in the Bodyless model. The Control model incorporated all e-mail characteristics.

Hierarchical regressions. Hierarchical linear regressions were run to predict phishing detection precision and recall in each condition. This step-wise approach allows for examining the added predictive value of every set of personal traits.

Step 1 included all demographic traits as predictors of phishing detection ability (age, gender, education level, occupational status, income). Step 2 added experiential question responses (prior professional experience with executive assistance, self-reported knowledge of cybersecurity, expectation to fall for phishing, self-reported phishing reception frequency and estimated daily amount of e-mails received). Step 3 added participants’ mean scores on the three IUIPC dimensions (awareness, collection, control). Step 4 added mean scores on the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism). Gender and occupational status were treated as categorical variables,

all others as continuous.

Normality of residuals checks were done visually with QQ plots. When additional steps did not significantly reduce the residual sum of squares (RSS), only the effects in the more parsimonious step were interpreted. All hierarchical regressions were analyzed in R [51]. χ^2 -, t- and correlation tests were performed using *scipy* version 1.5.4 [54] in Python 3. All results were interpreted against a two-tailed significance level of 0.05, unless noted otherwise.

4 Results

The results are divided into three key analyses. Section 4.1 shows the overall detection rates for each phishing e-mail per display condition. Section 4.2 describes the fitted e-mail feature regressions. Section 4.3 describes the hierarchical regressions on personal traits to predict phishing susceptibility.

4.1 Phishing detection varies widely by e-mail type and is the worst in Bodyless

According to the overall hypothesis, higher phishing detection proportions would be expected in the Bodyless condition for phishing e-mails 1, 5, 6, 7 and 8, which contained clearly suspicious header details. Table 2 shows the phishing detection proportions per condition per e-mail, and corresponding χ^2 -tests for proportional equality. Phishing e-mail 5 was detected by the majority of participants in all conditions, as well as the “Nigerian prince”-style phishing e-mails (3 and 4) in Control and Headerless ($\chi^2(2, N = 252) = 42.49, p < .001$). Phishing e-mail 7 about updating Microsoft payment details was detected at around chance level in all conditions. The remaining four phishing e-mails were at most detected by 29% of participants across all three conditions.

Most relative detection proportions were as expected, except for phishing e-mails 1, 2 and 6, although the lower detection proportions in Headerless for phishing e-mails 1 and 2 could also be considered as expected at trend level ($\chi^2(2, N = 252) = 3.80, p = .150$; $\chi^2(2, N = 252) = 4.52$,

Table 3: Number of participants who hovered over phishing URLs.

e-mail	condition	N hovered	N labeled as phishing or spam	URL in body
p5	Headerless	1	1	https://apple.ngrok.io/3p8sf9JeGzr60+haC9F9mxANtLM
p6	Headerless	2	0	http://25.245.256.02/excel/3p8sf9JeGzr60+haC9F9mxANtLM
p7	Control	3	1	http://office365.microsoft.netgriokgth.com
p8	Control	1	0	https://ngrok.io/b31d032cfdcf47a399990a71e43c5d2a

$p = .100$). Detection of phishing e-mail 6 was worst in Headerless ($\chi^2(2, N = 252) = 12.13$, $p = .010$), while a detection proportion comparable to Control would have been expected in the best case scenario. That is, if participants in Headerless hovered over the hyperlink and recognized the suspicious URL.

Of note is that in Bodyless, one participant labeled phishing e-mail 1 as “Other”, despite correctly identifying the homograph attack. They commented “No informative subject, accent on the e in sam.jones anneon email address”. This suggests that people may perfectly spot the discrepancy in sender details, but lack the knowledge that these are intentional deception tactics.

4.2 Phishing detection is not predicted by e-mail characteristics

To see if people use consistent rules to infer suspiciousness from e-mail characteristics, linear regressions were run with e-mail characteristics to predict the phishing detection proportions in each display condition. In Control, the full model yielded a significant regression ($F(24, 22) = 2.187$, $p = .035$, $R^2 = 0.7047$, $R^2_{adj} = 0.3825$) where more linguistic errors ($\beta = -3.334$, $p = .023$) and presence of the Anneon company name in the sender e-mail address ($\beta = -0.200$, $p = .038$) predicted lower phishing proportions. The former can be explained by the absence of linguistic errors in the phishing e-mails and presence of some grammatical errors and typos in some legitimate e-mails. In Headerless, a multiple regression predicting phishing detection proportions with only e-mail header-based features was not significant ($F(19, 27) = 1.425$, $p = .195$, $R^2 = 0.500$, $R^2_{adj} = 0.149$).

In Bodyless, a significant regression ($F(9, 37) = 3.186$, $p = .006$, $R^2 = 0.450$, $R^2_{adj} = 0.298$) was fit with all e-mail header-based features. Longer subject lines were found to predict higher phishing detection proportions ($\beta = 0.002$, $p < 0.001$). Phishing e-mail 5 had the longest subject line of all e-mails and had the highest detection rate in Bodyless, which explains this small, but highly significant effect. Since none of the other e-mail characteristics significantly predicted phishing detection across conditions, people do not seem to use consistent strategies in differentiating phishing from genuine e-mails.

Even when participants hovered over phishing URLs, most did not raise suspicion. One common piece of security advice is to check the true URLs of links in e-mails, by hovering over them [46]. To see if people do so, this study tracked and analyzed user interactions with e-mail links. Phishing e-mails 5, 6, 7 and 8 contained malicious URLs. Seven participants hovered over at least one of them. In two cases, the e-mail was labeled as “Phishing”. One of the three participants who hovered over the URL in phishing e-mail 7, labeled the e-mail as “Other”. For phishing e-mails 5 and 6, no URL hovers were observed in Control, nor for phishing e-mails 7 and 8 in Headerless—see Table 3. This suggests that most participants who labeled phishing e-mails as phishing did not base their judgments on the true URL of linked e-mail contents or did not know what to do with this information.

4.3 Phishing detection is not reliably predicted by personal traits

Overall, phishing detection accuracies varied greatly between participants. The mean phishing detection recall score was 0.46 (SD=0.2) in Control, 0.25 (SD=0.18) in Bodyless and 0.37 (SD=0.16) in Headerless, showing that most people detected less than half of all phishing e-mails. The mean phishing detection precision score was 0.93 (SD=0.18) in Control, 0.73 (SD=0.37) in Bodyless and 0.81 (SD=0.27) in Headerless, suggesting that when people think an e-mail is phishing, they are mostly correct.

To investigate individual differences in phishing susceptibility, hierarchical linear regressions were run to see if personal traits can reliably predict participants’ phishing detection recall and precision scores. Table 4 shows ANOVA results that compare the added value of each hierarchical regression step in predicting phishing detection recall from personal traits. None of the steps significantly reduced model RSS compared to step 1 in Control and Headerless, and step 1 regressions did not significantly predict phishing detection recall in Control ($F(7, 75) = 1.292$, $p = .266$, $R^2 = 0.108$, $R^2_{adj} = 0.024$), nor in Headerless ($F(7, 75) = 1.590$, $p = .152$, $R^2 = 0.130$, $R^2_{adj} = 0.048$). Only adding experiential question responses in step 2 in Bodyless significantly reduced RSS compared to step 1 and showed a significant regression ($F(12, 71) = 3.014$, $p = .002$, $R^2 = 0.338$, $R^2_{adj} = 0.226$).

Table 4: **Hierarchical regressions predicting phishing detection recall.** Only step 2 in Bodyless significantly improved phishing detection recall predictions compared to step 1. None of the hierarchical regression steps significantly improved model fits in Control and Headerless compared to step 1. Colors indicate whether the regression fit at the respective step was significant (green) or not (yellow), e.g., step 1 regressions in Control and Headerless were non-significant. RSS = Residual Sum of Squares

step	predictors	Control			Headerless			Bodyless		
		RSS	F (df)	<i>p</i>	RSS	F (df)	<i>p</i>	RSS	F (df)	<i>p</i>
1	demographics	3.13			1.90			2.31		
2	experiential questions	2.80	1.61 (5, 70)	.171	1.82	0.679 (5, 70)	.642	1.80	3.94 (5, 71)	.004
3	privacy concerns	2.70	0.826 (3, 67)	.484	1.65	2.22 (3, 67)	.095	1.78	0.251 (3, 68)	.860
4	Big Five	2.55	0.696 (5, 62)	.629	1.55	0.875 (5, 62)	.503	1.64	1.08 (5, 63)	.379

Table 5: **Hierarchical regressions predicting phishing detection precision.** None of the hierarchical regression steps in any condition significantly improved fit results compared to step 1. Colors indicate whether the regression fit at the respective step was significant (green) or not (yellow), e.g., only the step 1 regression in Control was significant. RSS = Residual Sum of Squares

step	predictors	Control			Headerless			Bodyless		
		RSS	F (df)	<i>p</i>	RSS	F (df)	<i>p</i>	RSS	F (df)	<i>p</i>
1	demographics	2.94			5.29			9.66		
2	experiential questions	2.82	0.659 (5, 70)	.659	5.18	0.332 (5, 70)	.892	8.65	1.71 (5, 71)	.144
3	privacy concerns	2.53	2.63 (3, 67)	.058	4.89	1.36 (3, 67)	.263	8.20	1.27 (3, 68)	.294
4	Big Five	2.30	1.24 (5, 62)	.304	4.40	1.36 (5, 62)	.251	7.42	1.31 (5, 63)	.271

Lower phishing detection recall was predicted by higher age ($\beta = -0.003, p = .040$), less frequent self-reported phishing reception ($\beta = -0.057, p = .003$) and more professional experience with executive assistance work ($\beta = -0.054, p < 0.001$).

None of the steps in the hierarchical regressions showed significant reductions in model residuals when predicting phishing detection precision from personal traits (see Table 5). Therefore, only step 1 regressions are reported further. In Control, the step 1 multiple regression with only demographic traits was significant ($F(7,75) = 2.436, p = .026, R^2 = 0.185, R^2_{adj} = 0.109$). Higher education level predicted higher phishing detection precision ($\beta = 0.037, p = .013$) and higher income predicted lower phishing detection precision ($\beta = -0.001, p = .018$). Step 1 regressions did not significantly predict phishing detection precision in Bodyless ($F(7,76) = 1.620, p = .143, R^2 = 0.130, R^2_{adj} = 0.050$), nor in Headerless ($F(7,75) = 1.309, p = .258, R^2 = 0.109, R^2_{adj} = 0.026$) conditions. Effect sizes of all significant predictors were small and arguably of limited meaningful value. Note that using a different order of regression steps did not change the results.

Higher age was associated with slower labeling responses in Control ($r = 0.355, p = .002$) and Bodyless ($r = 0.341, p = .002$). That is, older participants were slower at the task

overall. However, no significant associations were found between mean labeling RTs and phishing detection recall or precision in any of the display conditions. This further implies that demographics do not reliably predict phishing detection ability.

Adding “spam” as an accurate phishing detection label does not lead to more consistent results.

Some participants may have confused the meaning of “spam” and “phishing”. Hence, additional regressions with personal traits were run where both “spam” and “phishing” were regarded as accurate (true positive) labels for phishing e-mails and false positive labels for legitimate e-mails. This approach yielded prediction improvements for a step 2 regression in Control and step 4 regression in Headerless. Both regressions were significant. In the step 2 regression in Control (demographics and experiential questions; $F(12,70) = 1.973, p = .040, R^2 = 0.253, R^2_{adj} = 0.125$), older participants had a somewhat higher phishing detection recall score ($\beta = -0.004, p = .014$). The step 4 regression in Headerless (including all personal traits) predicted phishing detection recall at trend level ($F(20,62) = 1.637, p = .072, R^2 = 0.346, R^2_{adj} = 0.135$), where higher phishing detection recall was predicted by higher mean extraversion ($\beta = 0.040, p = .003$) and neuroticism ($\beta = 0.033, p = .048$). Higher mean agreeableness predicted lower phishing detection recall ($\beta = -0.040,$

$p = .012$). None of the steps in the hierarchical regressions for the Bodyless condition were significant, meaning none of the personal traits predicted phishing detection recall in Bodyless, even when “spam” was considered as an accurate phishing detection label.

In Control, phishing detection precision was significantly predicted in a step 3 regression with demographics, experiential questions and privacy concerns ($F(7,76) = 1.995$, $p = .029$, $R^2 = 0.309$, $R^2_{adj} = 0.154$). Less frequent self-reported phishing reception ($\beta = -0.041$, $p = .017$) and higher IUIPC “control” dimension scores ($\beta = -0.043$, $p = .014$) predicted lower phishing detection precision. None of the hierarchical regressions predicting phishing detection precision in Bodyless and Headerless were significant.

Altogether, whereas more personal traits were found to predict phishing detection recall by including “spam” as an accurate phishing detection label, the effects remained inconsistent over conditions and effect sizes were of limited meaning. These analyses strongly suggest that personal traits (e.g., demographics, personality traits, privacy concerns) do not consistently relate to how susceptible people are to phishing e-mails.

5 Discussion

Given the rising and increasingly sophisticated threat of phishing e-mails, it is essential to understand why people fall for them and to develop new solutions that reduce phishing victimization. This study highlights the possibility that phishing susceptibility is caused by inattention to suspicious source details found in e-mail headers. It tested the phishing detection ability of a representative sample in an e-mail processing task with different display conditions. Contrary to expectations, participants were not better at detecting phishing when only e-mail header details were displayed. Since the vast majority did detect phishing e-mail 5 in all conditions and the “Nigerian prince” scams in Control and Headerless, low participant motivation to perform well at the task is an unlikely explanation for the low overall detection rates. These findings show that people do not necessarily have a blind spot for e-mail source details, but instead do not recognize deception tactics commonly used in phishing.

The lack of e-mail characteristics predictive of phishing detection proportions confirm the idea that users do not rely on consistent tactics to gauge e-mail authenticity. One heuristic to do so, for instance, is checking if the sender e-mail address domain corresponds with the organization the sender claims to be from. If people used this rule, most participants should have detected phishing e-mails 1, 2 and 6 in Control. Another often given advice to avoid getting phished is to always inspect the actual URL of links in e-mail content, by hovering over them. If people adopted this advice, phishing e-mails 5, 6, 7 and 8 should have been detected by the majority as well. The lack of participants who did so suggests that common

anti-phishing advice is not used or that they do not know what to do with the gathered information [46], and may reflect people’s general misreading of URL domains when they hover over links [42, 2].

Many existing efforts to reduce phishing victimization rely on some form of training and are widely implemented in organizations and public campaigns already [5, 25, 46, 47, 6, 45, 52, 20, 58, 27, 30, 30, 29]. If the general public followed common cybersecurity advice, this study should have found higher average phishing detection proportions. The low detection rates imply the need for alternative solutions that help people recognize deception tactics used in phishing e-mails.

A strategy would be to target at-risk individuals with personalized anti-phishing interventions. If traits such as demographics were reliable predictors of phishing susceptibility, cybersecurity training could be targeted more specifically at certain demographic groups. However, the present study used an improved sample and task design, and still found no consistent relations between phishing detection and demographics, personality traits, privacy concerns, self-reported cybersecurity knowledge, nor self-reported phishing susceptibility. This accord with results from studies that also used role-playing tasks with a larger variety of phishing e-mails [13, 28, 26]. Whereas more research is needed to see if people with certain traits may be more susceptible to specific types of phishing (e.g., see [32]), interventions solely based on personal traits are not well-justified by the current body of research. Another under-researched direction is to profile within-individual differences in attention and situational changes to predict phishing susceptibility.

Taken together, this study emphasizes the need for research on user-centered techniques to reduce phishing susceptibility. In this approach, knowledge about online deception tactics needs to be accessible and usable for users in real-time. This moves away from conventional time-limited training programs and calls for more interdisciplinary collaboration between software developers and social scientists. An encouraging example from work on fake news detection showed that simply asking Twitter users to think about the veracity of social media articles reduced content sharing from untrustworthy sources [43]. More studies are needed to test similar tactics in e-mail inbox interfaces. Examples include explaining URLs to users when they hover over them [45, 58, 5] and Outlook’s external sender warnings. New experiments are being conducted by the authors on new e-mail functionalities in this realm, e.g., showing explainable suspicion scores and changing text colors for suspicious e-mails. Such interventions could provide cost-effective alternatives to anti-phishing training programs that suffer from questionable long-term effectiveness [27, 47] or phishing simulations that bear the risk of damaging employee relationships [59].

5.1 Limitations

The business context of this task may have been difficult to empathize with for participants without corporate experience, although business experience was not needed to recognize the suspicious details in the provided phishing e-mails. Moreover, knowing the business context may not necessarily lead to better phishing detection. Current phishing attacks are often deliberately adapted to organizational contexts, as in phishing e-mails 1, 2, 6 and 8, and real employees have fallen for them. Unfamiliarity with the business context may in fact have prompted people to read the e-mail contents more carefully before deciding what to do with them.

Next, the task interface did not fully mimic an actual inbox. Only the single e-mail display mimicked e-mail displays in Gmail. The task may have been more convincing if situated in an actual inbox, although similar setups were used in previous works [52, 41, 13]. It is also possible that people in Bodyless still ignored sender e-mail addresses and merely based their judgments on the subject line and sender name. Various online and offline eye tracking methods were considered to measure participants' visual attention, but none were able to differentiate users' gaze at such granular levels.

Lastly, this study only asked for participants' self-reported cybersecurity knowledge and not their actual amount of prior cybersecurity education or anti-phishing training. However, equal distributions of variance in cybersecurity training can be expected in each experimental condition, since participants were randomly allocated to either of them.

6 Conclusion

Phishing e-mails are a growing and increasingly sophisticated threat in our daily lives. Whereas e-mail messages can easily be manipulated, altering actual source details is more difficult to achieve. Consequently, phishing e-mails will often show suspicious signs in e-mail header details. When people fail to pay attention to them, they may especially be prone to falling for phishing e-mails. The present study compared people's phishing susceptibility when only e-mail headers were displayed, to when they saw full e-mail messages in a realistic task context. Presenting people merely with e-mail header details was expected to improve phishing detection. Surprisingly, this was not the case. Phishing susceptibility did not seem to be caused by blindness to source details. Instead, the results imply that people do not recognize deception tactics that are often used in phishing. The findings also affirmed that personal traits do not reliably predict phishing susceptibility. Altogether, this study encourages more interdisciplinary development of alternative user-centered tools that help us in the challenge against phishing.

7 Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was funded by the Dawes Centre for Future Crime at UCL.

References

- [1] Hossein Abroshan, Jan Devos, Geert Poels, and Eric Laermans. Phishing Happens beyond Technology: The Effects of Human Behaviors and Demographics on Each Step of a Phishing Process. *IEEE Access*, 9:44928–44949, 2021. ISSN: 21693536. DOI: [10.1109/ACCESS.2021.3066383](https://doi.org/10.1109/ACCESS.2021.3066383).
- [2] Sara Albakry, Kami Vaniea, and Maria K. Wolters. What is this URL's Destination? Empirical Evaluation of Users' URL Reading. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 2020. ISBN: 9781450367080. DOI: [10.1145/3313831.3376168](https://doi.org/10.1145/3313831.3376168).
- [3] Hussain Aldawood and Geoffrey Skinner. Reviewing cyber security social engineering training and awareness programs-pitfalls and ongoing issues. *Future Internet*, 11(3), 2019. ISSN: 19995903. DOI: [10.3390/fi11030073](https://doi.org/10.3390/fi11030073).
- [4] Ibrahim Alseadoon, M. F.I. Othman, and Taizan Chan. What is the influence of users' characteristics on their ability to detect phishing emails? *Lecture Notes in Electrical Engineering*, 315:949–962, 2015. ISSN: 18761119. DOI: [10.1007/978-3-319-07674-4_89](https://doi.org/10.1007/978-3-319-07674-4_89).
- [5] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I don't need an expert! making url phishing features human comprehensible. *Conference on Human Factors in Computing Systems - Proceedings*, 2021. DOI: [10.1145/3411764.3445574](https://doi.org/10.1145/3411764.3445574).
- [6] Aurélien Baillon, Jeroen De Bruin, Aysil Emirmahmutoglu, Evelien Van De Veer, and Bram Van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. *PLoS ONE*, 14(12), 2019. ISSN: 19326203. DOI: [10.1371/journal.pone.0224216](https://doi.org/10.1371/journal.pone.0224216).
- [7] Maxim Baryshevtsev and Joseph McGlynn. Persuasive Appeals Predict Credibility Judgments of Phishing Messages. *Cyberpsychology, Behavior, and Social Networking*, 23(5):297–302, 2020. ISSN: 21522723. DOI: [10.1089/cyber.2019.0592](https://doi.org/10.1089/cyber.2019.0592).
- [8] Mark Blythe, Helen Petrie, and John Clark. F for fake: four studies on how we fall for phish. In pages 3469–3478, 2011. DOI: [10.1145/1978942.1979459](https://doi.org/10.1145/1978942.1979459).

- [9] Frank Kun Yueh Chou, Abbott Po Shun Chen, and Vincent Cheng Lung Lo. Mindless response or mindful interpretation: Examining the effect of message influence on phishing susceptibility. *Sustainability (Switzerland)*, 13(4):1–25, 2021. ISSN: 20711050. DOI: [10.3390/su13041651](https://doi.org/10.3390/su13041651).
- [10] Robert B. Cialdini and Noah J. Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621, 2004. ISSN: 00664308. DOI: [10.1146/annurev.psych.55.090902.142015](https://doi.org/10.1146/annurev.psych.55.090902.142015).
- [11] Ronald V. Clarke. Situational crime prevention. In redacted by Richard Wortley and Michael Townsley, *Environmental Criminology and Crime Analysis*, Crime Science Series. Routledge, New York, 2nd edition edition, 2008.
- [12] Marco De Bona and Federica Paci. A real world study on employees’ susceptibility to phishing attacks. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, 2020. ISBN: 9781450388337. DOI: [10.1145/3407023.3409179](https://doi.org/10.1145/3407023.3409179).
- [13] Rachna Dhamija, J. Doug Tygar, and Marti Hearst. Why phishing works. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 581–590, 2006. DOI: [10.1145/1124772.1124861](https://doi.org/10.1145/1124772.1124861).
- [14] Alejandra Diaz, Alan T. Sherman, and Anupam Joshi. Phishing in an academic community: A study of user susceptibility and behavior. *Cryptologia*, 44(1):53–67, 2020. ISSN: 0161-1194. DOI: [10.1080/01611194.2019.1623343](https://doi.org/10.1080/01611194.2019.1623343).
- [15] Julie S. Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral response to phishing risk. Technical report, 2007, pages 37–44. DOI: [10.1145/1299015.1299019](https://doi.org/10.1145/1299015.1299019).
- [16] Natalie C. Ebner et al. Uncovering Susceptibility Risk to Online Deception in Aging. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 75(3):522–533, 2020. ISSN: 10795014. DOI: [10.1093/geronb/gby036](https://doi.org/10.1093/geronb/gby036).
- [17] Edwin Donald Frauenstein and Stephen Flowerday. Susceptibility to phishing on social network sites: A personality information processing model. *Computers and Security*, 94, 2020. ISSN: 01674048. DOI: [10.1016/j.cose.2020.101862](https://doi.org/10.1016/j.cose.2020.101862).
- [18] Steven Furnell, Kieran Millet, and M. Papadaki. Fifteen years of phishing: can technology save us? *Computer Fraud and Security*, 2019(7):11–16, 2019. ISSN: 13613723. DOI: [10.1016/S1361-3723\(19\)30074-0](https://doi.org/10.1016/S1361-3723(19)30074-0).
- [19] Adam Kavon Ghazi-Tehrani and Henry N. Pontell. Phishing Evolves: Analyzing the Enduring Cybercrime. *Victims and Offenders*, 16(3):316–342, 2021. ISSN: 15564991. DOI: [10.1080/15564886.2020.1829224](https://doi.org/10.1080/15564886.2020.1829224).
- [20] C. J. Gokul, Sankalp Pandit, Sukanya Vaddepalli, Harshal Tupsamudre, Vijayanand Banahatti, and Sachin Lodha. Phishy - A serious game to train enterprise users on phishing awareness. In *CHI PLAY 2018 - Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pages 169–181, 2018. ISBN: 9781450359689. DOI: [10.1145/3270316.3273042](https://doi.org/10.1145/3270316.3273042).
- [21] Tzipora Halevi, James Lewis, and Nasir Memon. A pilot study of cyber security and privacy related behavior and personality traits. *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web:737–744*, 2013. DOI: [10.1145/2487788.2488034](https://doi.org/10.1145/2487788.2488034).
- [22] Farkhondeh Hassandoust, Harminder Singh, and Jocelyn Williams. The role of contextualization in users’ vulnerability to phishing attempts. *Australasian Journal of Information Systems*, 24, 2020. ISSN: 14498618. DOI: [10.3127/AJIS.V24I0.2693](https://doi.org/10.3127/AJIS.V24I0.2693).
- [23] Joseph M. Hatfield. Social engineering in cybersecurity: The evolution of a concept. *Computers and Security*, 73:102–113, 2018. ISSN: 01674048. DOI: [10.1016/j.cose.2017.10.008](https://doi.org/10.1016/j.cose.2017.10.008).
- [24] Markus Jakobsson. The Human Factor in Phishing. *Privacy Security of Consumer Information*, 7:1–19, 2007. URL: <http://markus-jakobsson.com/papers/jakobsson-psci07.pdf>.
- [25] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. Don’t click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences*, 10(1):1–41, 2020. ISSN: 21921962. DOI: [10.1186/s13673-020-00237-7](https://doi.org/10.1186/s13673-020-00237-7).
- [26] Helen S. Jones, John N. Towse, Nicholas Race, and Timothy Harrison. Email fraud: The search for psychological predictors of susceptibility. *PLoS ONE*, 14(1):1–15, 2019. ISSN: 19326203. DOI: [10.1371/journal.pone.0209684](https://doi.org/10.1371/journal.pone.0209684).
- [27] Iacovos Kirlappos and M. Angela Sasse. Security education against Phishing: A modest proposal for a Major Rethink. *IEEE Security and Privacy*, 10(2):24–32, 2012. ISSN: 15407993. DOI: [10.1109/MSP.2011.179](https://doi.org/10.1109/MSP.2011.179).

- [28] Sabina Kleitman, Marvin K.H. Law, and Judy Kay. It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLoS ONE*, 13(10), 2018. ISSN: 19326203. DOI: [10.1371/journal.pone.0205089](https://doi.org/10.1371/journal.pone.0205089).
- [29] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Lessons from a real world evaluation of anti-phishing training. *eCrime Researchers Summit*, 2008. DOI: [10.1109/ECRIME.2008.4696970](https://doi.org/10.1109/ECRIME.2008.4696970).
- [30] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology*, 10(2), 2010. ISSN: 15335399. DOI: [10.1145/1754393.1754396](https://doi.org/10.1145/1754393.1754396).
- [31] Frieder R. Lang, Dennis John, Oliver Lüdtkke, Jürgen Schupp, and Gert G. Wagner. Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior Research Methods*, 43(2):548–567, 2011. ISSN: 1554351X. DOI: [10.3758/s13428-011-0066-z](https://doi.org/10.3758/s13428-011-0066-z).
- [32] Tian Lin, Daniel E. Capecci, Donovan M. Ellis, Harold A. Rocha, Sandeep Dommaraju, Daniela S. Oliveira, and Natalie C. Ebner. Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Transactions on Computer-Human Interaction*, 26(5):32, 2019. ISSN: 15577325. DOI: [10.1145/3336141](https://doi.org/10.1145/3336141).
- [33] Xin (Robert) Luo, Wei Zhang, Stephen Burd, and Alessandro Seazzu. Investigating phishing victimization with the heuristic–systematic model: a theoretical framework and an exploration. *Computers & Security*, 38:28–38, 2013. ISSN: 0167-4048. DOI: [10.1016/j.cose.2012.12.003](https://doi.org/10.1016/j.cose.2012.12.003).
- [34] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004. ISSN: 10477047. DOI: [10.1287/isre.1040.0032](https://doi.org/10.1287/isre.1040.0032).
- [35] Saif M. Mohammad. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text, 2016. DOI: [10.1016/B978-0-08-100508-8.00009-6](https://doi.org/10.1016/B978-0-08-100508-8.00009-6).
- [36] Gregory D. Moody, Dennis F. Galletta, and Brian Kimball Dunn. Which phish get caught An exploratory study of individuals susceptibility to phishing. *European Journal of Information Systems*, 26(6):564–584, 2017. ISSN: 14769344. DOI: [10.1057/s41303-017-0058-x](https://doi.org/10.1057/s41303-017-0058-x).
- [37] Paula M.W. Musuva, Katherine W. Getao, and Christopher K. Chepken. A new approach to modelling the effects of cognitive processing and threat detection on phishing susceptibility. *Computers in Human Behavior*, 94:154–175, 2019. ISSN: 07475632. DOI: [10.1016/j.chb.2018.12.036](https://doi.org/10.1016/j.chb.2018.12.036).
- [38] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail Joon Ahn. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. *Proceedings of the 29th USENIX Security Symposium*:361–377, 2020. DOI: [10.5555/3489212.3489233](https://doi.org/10.5555/3489212.3489233).
- [39] Kathryn Parsons, Marcus Butavicius, Paul Delfabbro, and Meredith Lillie. Predicting susceptibility to social influence in phishing emails. *International Journal of Human Computer Studies*, 128(February):17–26, 2019. ISSN: 10959300. DOI: [10.1016/j.ijhcs.2019.02.007](https://doi.org/10.1016/j.ijhcs.2019.02.007).
- [40] Kathryn Parsons, Marcus Butavicius, Malcolm Pattinson, Agata McCormac, Dragana Calic, and Cate Jerram. Do users focus on the correct cues to differentiate between phishing and genuine emails? In *ACIS 2015 Proceedings - 26th Australasian Conference on Information Systems*, 2015. ISBN: 9780646953373.
- [41] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. *IFIP Advances in Information and Communication Technology*, 405:366–378, 2013. ISSN: 1868422X. DOI: [10.1007/978-3-642-39218-4_27](https://doi.org/10.1007/978-3-642-39218-4_27).
- [42] Ed Pearson, Cindy L. Bethel, Andrew F. Jarosz, and Mitchell E. Berman. "To click or not to click is the question": Fraudulent URL identification accuracy in a community sample. In *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*, pages 659–664, 2017. ISBN: 9781538616451. DOI: [10.1109/SMC.2017.8122682](https://doi.org/10.1109/SMC.2017.8122682).
- [43] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, 2021. ISSN: 14764687. DOI: [10.1038/s41586-021-03344-2](https://doi.org/10.1038/s41586-021-03344-2).
- [44] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7):770–780, 2020. ISSN: 14679280. DOI: [10.1177/0956797620939054](https://doi.org/10.1177/0956797620939054).

- [45] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. *Conference on Human Factors in Computing Systems*, 2019. DOI: [10.1145/3290605.3300748](https://doi.org/10.1145/3290605.3300748).
- [46] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *Proceedings of the 29th USENIX Security Symposium*, pages 89–108, 2020. ISBN: 9781939133175. URL: <https://www.usenix.org/conference/usenixsecurity20/presentation/redmiles>.
- [47] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. *Proceedings of the 16th Symposium on Usable Privacy and Security, SOUPS 2020*:259–284, 2020. URL: <https://www.usenix.org/conference/soups2020/presentation/reinheimer>.
- [48] Leonard Richardson. Beautiful Soup Documentation. *Media.Readthedocs.Org*:1–72, 2016. URL: <https://media.readthedocs.org/pdf/beautiful-soup-4/latest/beautiful-soup-4.pdf>.
- [49] Ben D. Sawyer and Peter A. Hancock. Hacking the Human: The Prevalence Paradox in Cybersecurity. *Human Factors*, 60(5):597–609, 2018. ISSN: 15478181. DOI: [10.1177/0018720818780472](https://doi.org/10.1177/0018720818780472).
- [50] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):453–457, 2019. ISSN: 2169-5067. DOI: [10.1177/1071181319631355](https://doi.org/10.1177/1071181319631355).
- [51] Team R Development Core. A Language and Environment for Statistical Computing, Vienna, Austria, 2018. URL: <http://www.r-project.org>.
- [52] René van Bavel, Nuria Rodríguez-Priego, José Vila, and Pam Briggs. Using protection motivation theory in the design of nudges to improve online security behavior. *International Journal of Human Computer Studies*, 123:29–39, 2019. ISSN: 10959300. DOI: [10.1016/j.ijhcs.2018.11.003](https://doi.org/10.1016/j.ijhcs.2018.11.003).
- [53] Verizon. 2021 data breach investigations report, 2021.
- [54] Pauli Virtanen et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. ISSN: 15487105. DOI: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2).
- [55] Arun Vishwanath. Examining the Distinct Antecedents of E-Mail Habits and its Influence on the Outcomes of a Phishing Attack. *Journal of Computer-Mediated Communication*, 20(5):570–584, 2015. ISSN: 10836101. DOI: [10.1111/jcc4.12126](https://doi.org/10.1111/jcc4.12126).
- [56] Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research*, 45(8):1146–1166, 2018. ISSN: 15523810. DOI: [10.1177/0093650215627483](https://doi.org/10.1177/0093650215627483).
- [57] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H. Raghav Rao. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3):576–586, 2011. ISSN: 01679236. DOI: [10.1016/j.dss.2011.03.002](https://doi.org/10.1016/j.dss.2011.03.002).
- [58] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers and Security*, 71:100–113, 2017. ISSN: 01674048. DOI: [10.1016/j.cose.2017.02.004](https://doi.org/10.1016/j.cose.2017.02.004).
- [59] Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. Analysing Simulated Phishing Campaigns for Staff:312–328, 2020. DOI: [10.1007/978-3-030-66504-3_19](https://doi.org/10.1007/978-3-030-66504-3_19).
- [60] Rick Wash. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 2020. ISSN: 25730142. DOI: [10.1145/3415231](https://doi.org/10.1145/3415231).
- [61] Emma J. Williams and Danielle Polage. How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. *Behaviour and Information Technology*, 38(2):184–197, 2019. ISSN: 13623001. DOI: [10.1080/0144929X.2018.1519599](https://doi.org/10.1080/0144929X.2018.1519599).
- [62] Ryan T. Wright, Matthew L. Jensen, Jason Bennett Thatcher, Michael Dinger, and Kent Marett. Influence techniques in phishing attacks: An examination of vulnerability and resistance. *Information Systems Research*, 25(2):385–400, 2014. ISSN: 15265536. DOI: [10.1287/isre.2014.0522](https://doi.org/10.1287/isre.2014.0522).
- [63] Ryan T. Wright and Kent Marett. The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived. *Journal of Management Information Systems*, 27(1):273–303, 2010. ISSN: 07421222. DOI: [10.2753/MIS0742-1222270111](https://doi.org/10.2753/MIS0742-1222270111).

A Task instructions

Display of e-mail examples in the instructions was adapted according to the participant's randomly assigned display condition. The screenshots below are taken from the Bodyless condition.

Imagine that you are working as the executive assistant for Alex Carter, a top executive at a company called Anneon. Anneon is a multinational corporation in the oil and gas industry.

Typical tasks of an executive assistant include, but are not limited to:

- managing the executive's calendar: (re)scheduling meetings;
- signing approved documents on behalf of Alex;
- managing all business and personal travel;
- preparing research and meeting packs.

Now you are asked to filter Alex' e-mails, so that Alex *only* gets to see relevant e-mails.

How many e-mails do you receive on an average day?

Please type the amount here

Is this type of work something you have done professionally in the past?

- no I am not sure yes, occasionally yes, at least part-time yes, full-time

SUBMIT

You will see one e-mail at a time.

For each e-mail, you need to indicate whether you think it needs your boss' attention or whether to "discard" it. To do so, you need to **select the applicable label beneath the e-mail** and click "SUBMIT".

Your screen will look like this:

As an executive assistant for Alex Carter at Anneon, how would you filter the following e-mail?

Remember: each *accurately* labeled e-mail will earn you £ 0.05.

ConEd Lakewood Deal (PJM East)

Blair, Greg <greg.blair@anneon.com> Tue, 6 Nov 2020 08:00
to harry.arora@anneon.com, iris.mack@anneon.com, brandon.cavazos@anneon.com
CC: alex.carter@anneon.com

Harry, Iris and Brandon:

I asked ConEd if they were prepared to hear our OFFER to buy peaking capacity from the Lakewood expansion. They are "in the midst of preparing a draft term sheet for purchasing the output/tolling" so we'll have to wait to see the parameters of sale.

I also asked if they would like to see PJM WEST synthetic toll offer from us. Unfortunately, NO.

To the extent it matters, the on-line date is now set for late 2020, so they would like to see our offer for a Jan '21 start date. Given that these guys are doing a real comparison against bricks and mortar, we would be looking at a \$63 million option premium for 500MW @ 3.50/KW-mo over a three-year term; \$84 million for a four-year term, I would hope we could be able to find a way to hedge the risk of too many \$500 hours killing us. Is there any cap we could buy to protect our downside risk here and still make this deal attractive? The heat rate call is struck at 10,900!

DISCARD NEEDS BOSS' ATTENTION

The yellow bar at the top indicates your progress.

If you label an e-mail as "NEEDS BOSS' ATTENTION", it will go to your boss Alex Carter.
If you label an e-mail as "DISCARD", you will be asked to specify the reason:

DISCARD NEEDS BOSS' ATTENTION

As an executive assistant I can take care of this

Spam [?]

Phishing [?]

Other: please specify... [?]

As a top executive, Alex is responsible for making strategic and financial decisions to keep the business running as profitably as possible. These decisions can range from foreign investment deals, finding new business partners, to changing local plant operations.

Next to this, Alex acts as a representative of Anneon who occasionally needs to present about the company at internal or external events.

Any e-mail that fits the scope of these responsibilities needs to be labeled as "NEEDS BOSS' ATTENTION".

"As an executive assistant, I can take care of this" indicates that the e-mail falls within the scope of the typical tasks of an executive assistant as outlined before.

Spam e-mails are messages that the recipient did not ask for and are sent to many people at once, often with commercial aims.

Phishing e-mails are fraudulent communications that appear to come from a reputable source. Their objective is to obtain sensitive data from their victim.

Select "Other" if you believe none of the other categories apply, with a brief comment with your reasoning.

After each submit, you will be presented with the next e-mail.

You cannot revise e-mails that you have already labeled.

The yellow bar at the top indicates your progress.

This example e-mail would be labeled as "DISCARD" with the reason "As an executive assistant, I can take care of this":

Prints ready

Smith, Jane <jane.smith@anneon.com>
to <alex.carter@anneon.com>

Mon, 2 Apr 2018 10:31:09 -0700 (PDT)

Hi Alex,

The Q1 report is printed in threefold as requested, please check your locker.

Jane

Printed reports for Alex can be managed by the executive assistant.

This example e-mail would be labeled as "NEEDS BOSS' ATTENTION":

NOPR (RM96-1-019)

Hess, Theresa <theresa.hess@anneon.com>
to alex.carter@anneon.com, shelley.holmes@anneon.com

Fri, 19 Oct 2019 15:34:32 -0700 (PDT)

CMS Energy (Bill Grygar and Kim Van Pelt) are asking whether the pipelines want to meet to discuss the NOPR and its affect on pipelines. And, whether the pipelines want to file comments. They say they haven't heard anything from INGAA or whether INGAA has plans to discuss or comment. They've suggested meeting the week of Oct 29. This notice was sent only to the GISB EC reps.

If INGAA is addressing, we won't want to duplicate their meetings. Shelley -- have you heard from INGAA? I've told CMS that we'll get back with them next week.

Theresa

NOPR stands for Notice of Proposed Rulemaking. Alex needs to be informed about NOPRs, as they legally affect Anneon's business operations.

Evaluating the Usability of Privacy Choice Mechanisms

Hana Habib
Carnegie Mellon University

Lorrie Faith Cranor
Carnegie Mellon University

Abstract

Privacy choice interfaces commonly take the form of cookie consent banners, advertising choices, sharing settings, and prompts to enable location and other system services. However, a growing body of research has repeatedly demonstrated that existing consent and privacy choice mechanisms are difficult for people to use. Our work synthesizes the approaches used in prior usability evaluations of privacy choice interactions and contributes a framework for conducting future evaluations. We first identify a comprehensive definition of usability for the privacy-choice context consisting of seven aspects: user needs, ability & effort, awareness, comprehension, sentiment, decision reversal, and nudging patterns. We then classify research methods and study designs for performing privacy choice usability evaluations. Next, we draw on classic approaches to usability testing and prior work in this space to identify a framework that can be applied to evaluations of different types of privacy choice interactions. Usability evaluations applying this framework can yield design recommendations that would improve the usability of these choice mechanisms, ameliorating some of the considerable user burden involved in privacy management.

1 Introduction

Consumer privacy protection has long been rooted in the notice and choice paradigm. This model assumes that companies notify users about how they handle their data and consumers exercise privacy choices according to their preferences. Thus, companies implement web and app interfaces with privacy

choice mechanisms that allow users to make choices about some form of collection or use of their personal data, including device permission prompts, cookie consent notices, social media audience settings, targeted advertising opt-outs, and mailing list opt-outs. The possible design space of privacy choice mechanisms is broad, resulting in interfaces that vary in type of choice, functionality, timing, channel, and modality [15]. Despite the availability of privacy controls, the notice and choice model arguably has not resulted in effective consumer privacy protection, in part due to the poor usability of privacy choice mechanisms [53].

The design of privacy choice and consent interfaces can significantly impact users' privacy outcomes. Historically, companies have had economic motivation to encourage users to share their data through such interactions and may not have exerted more than minimal effort in testing the usability of their privacy choice and consent interfaces. Furthermore, privacy choice interfaces require usability considerations beyond those considered for typical user interfaces. Generally, users make privacy decisions when trying to accomplish a different goal (e.g., browse a website or make an online purchase), which means that a choice interface that interferes with the primary goal might score high with respect to the usability of the privacy decision but low with respect to the primary goal.

Prior usability evaluations of privacy choice mechanisms have highlighted several obstacles to their effective use. For example, some privacy choice mechanisms may be difficult to configure without substantial technical knowledge [36]. Some seem to require that users put aside their preconceived assumptions and read explanations that most users readily skip over [51]. Furthermore, the use of dark patterns may nudge users toward less privacy-protective options provided in the interface [56]. Prior studies often include actionable design recommendations for a particular privacy choice context (e.g., [21, 40, 60]).

The expanding literature on privacy choice interfaces has explored a variety of usability considerations for privacy choice interactions, utilizing a spectrum of usability testing methods from the field of human-computer interaction. In this

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

work, we distill the usability aspects explored and methods used in prior work into a framework that can inform the design of future usability evaluations of privacy choice interactions. To develop this framework we adopt our prior work [20] presenting a comprehensive definition of usability for the context of privacy choice mechanisms consisting of seven objectives: user needs, ability & effort, awareness, comprehension, sentiment, decision reversal, and nudging patterns. We then categorize different research methods and study designs that can be used to perform usability evaluations of privacy choice interfaces. Next, drawing on classic approaches to usability testing and prior evaluations of privacy controls, we construct the Privacy Choice Evaluation Framework that can be applied to future evaluations of privacy choice interfaces. The framework provides criteria for evaluating each aspect of usability through the relevant evaluation approaches, serving as a guide for organizations that want to ensure provided privacy controls are effective in enabling consumers to manage their privacy. Furthermore, regulators can make use of this framework as they work to hold companies accountable to rigorous usability testing of privacy choice and consent processes.

After presenting our framework, we present an overview of the literature on privacy choice evaluations, illustrating the applicability of our framework. We then discuss additional considerations, guidance for organizations, and limitations of privacy choice usability. Our appendix includes guidance on using the evaluation framework through a detailed example.

2 Defining Privacy Choice Usability

To consider the holistic usability of privacy choice interfaces, it is important to first identify aspects of usability that are relevant to the privacy choice experience. We adopt our previous work [20], which reviewed definitions of usability drawn from academics and practitioners in the privacy, HCI, and user experience (UX) fields and identified seven distinct aspects of privacy choice usability. We use these seven aspects of usability to provide an organizing structure for our framework.

User Needs: Whether a privacy choice interface addresses the intended users' privacy needs in a particular privacy choice context. Also includes accuracy and completeness of the interface in addressing these needs. *Components from previous definitions:* **Effectiveness** (Feng et al. [15], ISO [27], Quesenbery [52]), **Useful** (Schaub and Cranor [54], Morville UX Honeycomb [44])

User Ability & Effort: Whether a privacy choice interface allows the intended users to accomplish a particular privacy goal and with minimal effort. *Components from previous definitions:* **Efficiency** (Feng et al. [15], ISO [27], Quesenbery [52], Nielsen [45]), **Usable** (Schaub and Cranor [54],

Morville UX Honeycomb [44]), **Accessible** by “non-experts” (Morville UX Honeycomb [44])

User Awareness: Whether the intended users are aware that a particular privacy choice exists within a privacy choice interface, and if they are able to find it. *Components from previous definitions:* **User awareness** (Feng et al. [15]), **Findable** (Schaub and Cranor [54], Morville UX Honeycomb [44]), **Easy to learn** - initial orientation (Quesenbery [52], Nielsen [45])

User Comprehension: Whether the intended users understand what a particular privacy choice does and the implications of their decisions. *Components from previous definitions:* **Comprehensiveness** (Feng et al. [15]), **Understandability** (Schaub and Cranor [54]), **Easy to learn** - continued learning (Quesenbery [52])

User Sentiment: Whether the intended users are satisfied with a privacy choice interface and options it provides. This includes whether users have faith that the privacy choice will be honored. *Components from previous definitions:* **Satisfaction** (ISO [27], Nielsen [45]), **Engaging** (Quesenbery [52]), **Desirable** (Morville UX Honeycomb [44]), **Credible** (Morville UX Honeycomb [44])

Decision Reversal: Whether a privacy choice interface allows the intended users to correct an error or change their decision. This also includes the effort required to do so. *Components from previous definitions:* **Error tolerant** (Quesenbery [52], Nielsen [45])

Nudging Patterns: Whether the design of a privacy choice interface leads the intended users to select certain choices in the interface over others (including dark patterns that lead users to less privacy-protective options). *Components from previous definitions:* **Neutrality** (Feng et al. [15])

3 Privacy Choice Evaluation Approaches

This section describes research methods and study designs that can be applied to privacy choice evaluations. While it is not a comprehensive list of all possible evaluation techniques, it demonstrates a wide breadth and diversity of approaches.

3.1 Expert Evaluation Methods

Inspection-based approaches, in which usability obstacles are identified through a systematic review of the interface by a domain expert, can be adapted to evaluate the usability of privacy choice and consent interfaces. Such approaches may be particularly beneficial in evaluating privacy choice interfaces in contexts where users may lack requisite background

privacy knowledge or experience. Prior examples of privacy choice usability studies conducted through expert evaluation include Grey et al.'s interaction criticism approach and Soe et al.'s heuristic evaluation of cookie consent banners [19, 56]. Here we provide a brief description of five inspection-based methods that could be used in evaluating for different usability aspects. Additional information about these approaches can be found in the HCI literature (e.g., [66]).

Perspective-based UI Inspection: One or more people evaluate the privacy choice interface from the perspective of a particular type of user (super-user, less-tech savvy, person with disability) or through the lens of a specific normative value, in this case privacy.

Individual Expert Review: One or more experts in HCI, the privacy choice domain, or the product conducts a review to find usability problems in a privacy choice interface according to the usability aspect(s) being evaluated.

Cognitive Walkthrough: An expert or team interacts with a privacy choice interface to identify usability issues that primarily impact user awareness. This method is based on the theory that users learn through exploration.

Heuristic Evaluation: An individual or team evaluates a privacy choice interface design against a list of UX principles (e.g. Nielsen Heuristics [46]) or other pre-defined criteria (e.g., regulatory requirements).

Formal Usability Evaluation: Trained inspectors conduct coordinated, individual usability assessments of a privacy choice interface (similar to formal code inspections). This may include collecting information about the shortest path, minimum number of actions, and time taken to complete a privacy choice task.

3.2 User Study Designs

User studies provide perspectives from individuals who are more likely to represent the opinions and behaviors of end-users of the privacy choice interface. Such evaluations of privacy choice interfaces can be implemented through different research methods and study designs as outlined below. Studies may combine elements to explore how well a privacy choice interface addresses particular usability aspects.

3.2.1 No Task Assigned

Self-reported: Self-report methods can help with understanding users' experiences with a privacy choice interface in the context of their actual use of the system. This can provide valuable insight even for privacy interfaces that users may

encounter infrequently. Furthermore, self-report methods can help understand users' privacy needs for a particular context. These studies can be conducted through surveys, interviews, and focus groups utilizing qualitative prompts, measurement scales, and other question types. Examples of prior self-report studies related to privacy choice interfaces include Malkin et al.'s survey of smart speaker users [40] and Colnago et al.'s interview study informing the design of a privacy assistant [11].

Observed: Observation studies primarily involve measurement of users' behavior when interacting with a deployed privacy choice interface, sometimes as part of an A/B test. Examples of such metrics include the average amount of time spent before making a privacy choice or percentage of users who click a particular option. Such studies provide an advantage over other study designs by providing insight into when and how users are actually interacting with an interface, which is particularly useful for the privacy choice context as privacy management is typically not users' primary reason for engaging with a system. However, observation studies do not typically provide an explanation as to why users interact with it in the way that they do, unless paired with an interview or survey. Previous observation studies of cookie consent interface designs include Utz et al.'s field study evaluation [61] and the logistics company DHL's A/B tests [49].

3.2.2 Participants Assigned Privacy Task

In their natural use of a system, users may encounter a particular privacy choice interface so infrequently that it may be difficult for researchers to assess its usability. Thus evaluating for some usability aspects may require explicitly assigning privacy-related tasks to ensure that users interact with the interface being evaluated. Additionally, participants are typically asked questions before or after task completion (or both). These user studies can be implemented through surveys, experiments, or lab usability studies.

Hypothetical Privacy Scenario: Participants are given a realistic scenario motivating a privacy choice and are asked how they would use a privacy choice interface (or other mechanism) to make that choice. An example of a hypothetical scenario that was used by Habib et al. [21] and Kumar et al. [5] to introduce tasks involving email opt-outs is "You just got the 10th update email from this website today. Now you want to stop receiving them."

Participant Inspection: Participants are shown a privacy choice interface and are encouraged to fully engage with it prior to answering questions (e.g., to measure their awareness or comprehension). Typically, participants are allowed to reference the interface while they are answering questions. Tsai

et al.'s online experiment used participant inspection to compare the design of a new Android permission manager tool with Android's native permission management interface [60].

Participant Quick Review: Participants are shown a privacy choice interface but are only allowed a short period to engage with it (e.g., 3 seconds). Typically, participants are not allowed to reference the interface while they are answering questions. Quick review may also be done as part of a task in which participants are exposed to the interface, but answer questions about it after they complete the task and the interface is no longer in view. Cranor et al. used quick review to evaluate whether participants noticed a “Do Not Sell My Personal Information” opt-out link and icon in the footer of an e-commerce website after their attention was directed to a nearby link [12].

Make Personal Privacy Choices: Participants are shown a privacy choice interface and are asked how they would interact with it according to their own personal privacy preferences. For example, Krsek et al.'s experiment asked participants to select their preferences for Facebook privacy settings under different nudging conditions [33].

3.2.3 Participants Assigned Distraction Task

Considering that privacy/security are often secondary priorities when users interact with a system, simulating this in an online experiment or lab usability study might require assigning participants a “distraction task.” Examples of distraction tasks include shopping for a particular item, or finding information on a website. Participants should encounter the choice interface or an indicator leading to it during their task.

Privacy Choice Prompt Appears: Participants are asked to complete a task that is unrelated to the privacy choice interface being evaluated, but are exposed to the privacy choice interface at some point in the study. For example, in Bermejo Fernández et al.'s online experiment evaluating the usability of cookie consent interfaces, a cookie consent banner appeared as participants arrived at the website to complete a survey about smart home devices [6].

Participant Seeks Out Privacy Settings: Participants are asked to complete a task that is unrelated to privacy but as part of the interface they can see the current privacy settings. During the course of task completion they may choose to change their privacy settings according to their preferences. Vaniea et al. conducted a series of lab studies in which participants were assigned photo management tasks during which they had the opportunity to observe and change the access control settings for each photo [62]. However, the authors report a number

of challenges they encountered while conducting these studies using this approach, including making sure participants understood the somewhat-complex desired access control policy, and balancing the need to make participants aware of the access control settings with a desire not to prime participants to think about access control more than they normally would.

4 The Privacy Choice Evaluation Framework

We introduce the Privacy Choice Evaluation Framework, summarized in Table 1, which provides a set of criteria that can be used in usability evaluations of privacy choice interfaces. We structure the framework according to the seven usability aspects defined in Section 2. For each criterion included in the framework, we highlight the study approaches described in Section 3 and describe measures or example prompts that can be incorporated into a usability study. We refer to established usability metrics and heuristics when appropriate, or specific components of existing usability scales that are applicable to the privacy choice context. It is important to note that the usability requirements and acceptable thresholds for meeting them are not universal, but rather depend on the context of the privacy choice interface. Many factors, including intended user groups, complexity of options, and devices used to display the privacy choice interface, influence whether a given privacy choice interface is sufficiently usable. The framework also considers the types of privacy choice interfaces relevant to each criterion, in terms of the Timing component of the privacy choices design space: *on-demand* (privacy settings pages that the user seeks out) or *interruptive* (privacy choice interfaces that appear at setup, just-in-time, are context-aware, are periodic, or are personalized) [15]. Furthermore, we provide citations to prior privacy choice evaluations when applicable to demonstrate possible implementations of the listed criteria.

4.1 User Needs

Prior to designing an interface, design teams often complete a needs assessment using qualitative approaches to better understand how and why users might use the interface. It is important to assess whether a resulting interface design is aligned with the identified needs and how completely it addresses them. Assessing user needs is relevant to both interruptive and on-demand privacy choice interfaces. Some evaluations in other parts of this framework rely on an understanding of user needs associated with a privacy interface.

4.1.1 Users' Privacy Objectives

This criterion pertains to understanding users' privacy objectives when using a particular system. Assessments of users' privacy objectives can be conducted as self-reported evaluations of past experiences or user studies involving assigned tasks. Prior work evaluating users' privacy objective when

Framework Criterion	Usability Aspect						Evaluation Approach				Interface Timing			
	Needs	Ability & Effort	Awareness	Comprehension	Sentiment	Decision Reversal	Nudging Patterns	Expert Evaluation	Self-Report	Observation	Privacy Task	Distraction Task	Interruptive	On-demand
Users' privacy objectives	●						●	●		●	●	●	●	●
Users' intentions	●							●		●	●	●	●	●
Interface completeness	●							●					●	●
Interface accuracy	●							●					●	●
Ability - make privacy choice		●				●	●			●	●	●	●	●
Time taken - make privacy choice		●				●	●			●	●	●	●	●
User actions - make privacy choice		●				●	●			●	●	●	●	●
Perceived effort - make privacy choice		●				●		●			●	●	●	●
Estimated effort - make privacy choice		●				●		●					●	●
Awareness of choice existence			●			●	●		●			●		●
Ability - find privacy choice			●			●	●			●	●			●
Time taken - find privacy choice			●			●	●			●	●			●
User actions - find privacy choice			●			●	●			●	●			●
Perceived effort - find privacy choice			●			●		●			●			●
Estimated effort - find privacy choice			●			●		●						●
Objective knowledge - focused attention				●		●	●		●		●		●	●
Objective knowledge - unfocused attention				●		●	●		●		●		●	●
Perceived effort - comprehension				●		●		●		●	●		●	●
Estimated effort - comprehension				●		●		●					●	●
Perceived transparency & control					●		●		●		●		●	●
Subjective knowledge					●		●		●		●		●	●
Levels of comfort & trust					●		●		●		●		●	●
Investment in decision-making					●		●		●		●		●	●
Impact on individual welfare						●		●		●	●		●	●
Unintended societal consequences						●		●					●	●
Alignment with regulatory objectives						●		●					●	●
Individual autonomy						●		●	●	●	●		●	●

Table 1: A summary of the Privacy Choice Evaluation Framework which provides an overview of the evaluation criteria (grouped by the usability aspect in Section 4 under which they are described). Marked are the applicable usability aspects (defined in Section 2), evaluation approaches (described in Section 3), and timing of privacy choice interface (interruptive and/or on-demand) for each criterion.

using a privacy choice interface includes Fiesler et al.'s survey of Facebook users, which asked "Why did you choose this privacy setting?" for each post shared by their participants. Additional example prompts include:

- What settings or controls related to [domain of privacy choice] would you like to have available to you, if any? [for initial exploration into user needs prior to designing the privacy choice interface]
- What *other* settings or controls related to [domain of privacy choice] would you like to have available to you, if any? [for further exploration into user needs related to an existing privacy choice interface design]

4.1.2 Users' Intentions

Similar to exploring users' objectives, it is important to assess why users interact with privacy choice interfaces in the way that they do, including evaluating users' decision strategies. Assessing users' intentions requires participants to reflect on what they were trying to achieve in a past interaction with a privacy choice interface, which could be privacy related (e.g., trying to prevent a certain type of data collection) or more practical (e.g., to continue to the main website). This can be conducted as self-reported evaluation of past experiences or user studies involving assigned tasks. An example prompt to assess users' intentions is: What were you trying to achieve when you [interacted with the choice interface]?

4.1.3 Interface Completeness

The criterion assesses how completely an implemented privacy choice interface achieves users' needs through an expert evaluation. This requires having some knowledge of users' objectives through a user study and thus ideally should be done in conjunction with the criterion described in 4.1.1. Such evaluations could include heuristics such as:

- Does the interface meet the needs of different types of users (e.g. those who want fine-grained controls and those who want simplicity.)?
- Does it allow users to achieve all of their stated objectives, or only some of them?

Some interfaces may be incomplete because they do not allow users to make desired privacy choices at all, for example not offering the option to post anonymously on a social media platform. Others may offer desired choices, but not at the level of granularity desired by some users, for example allowing social media users to restrict the audience of their posts to friends, but not allowing them to restrict the audience to only a particular subset of their friends.

4.1.4 Interface Accuracy

In addition to how completely a privacy choice interface meets users' needs, an expert evaluation can also assess how accurately it achieves users' needs. This requires having some knowledge of users' intentions when using a privacy choice interface, and could be done in conjunction with a user study exploring the criterion described in 4.1.2. These evaluations could include evaluating whether there is a mismatch between what the user said they were trying to achieve and what the interface actually does, and identifying how the interface helps users accomplish their goals. For example, some cookie consent interfaces give users a choice of "accept all cookies," "reject all cookies," or "manage cookies." Reject all cookies is not an accurate label on most websites where it actually rejects all non-essential cookies but not the "strictly necessary" cookies needed for the site to function and which are permitted under GDPR.

4.2 User Ability & Effort

Usability testing often involves quantitative measures that estimate the effort involved in using an interface. These metrics can be used to compare interfaces (e.g., a previous version of the interface, alternate designs, or the interface of a similar product). Measuring perceived effort is relevant to both interruptive and on-demand choice interfaces. For on-demand privacy choice interfaces, much of the effort involved in using the interface will likely be in finding where it is (which we discuss as separate criteria in 4.3), but users could possibly make other errors such as forgetting to save their choices or toggling a choice in the wrong direction.

4.2.1 Ability to Make a Privacy Choice

Ability evaluations may assess whether users are able to complete the end-to-end interaction required to make a privacy choice, as well as the type and extent of assistance they require. Ability to make a privacy choice can be measured through observational field studies or user studies involving task assignment. Prior work evaluating for ability to make a privacy choice includes Chalhoub et al.'s ethnographic study which surfaced participants' inability to configure privacy settings on their smart home devices [7].

4.2.2 Time Taken to Make Privacy Choice

Time is one measure of the effort required to use an interface, and can be measured through both observation and user study tasks. However, the raw time to complete a privacy decision may be an imperfect measure if users are multi-tasking or thinking aloud during a moderated study. Alternative time-based metrics include time-based efficiency and overall relative efficiency [43]. An example of prior work

that included timing metrics in their usability testing is Garlach and Suthers’s study evaluating the effectiveness of the AdChoices icon in the mobile environment [18].

4.2.3 User Actions Required to Make Privacy Choice

Another measure of effort is the number and type of user actions (e.g., clicks, hovers, form fields) required to complete a privacy choice. This may sometimes be a more reliable measure than time and may also reveal common user errors that result in extra user actions. User actions can be measured through observation as well as user study tasks. Habib et al. tracked clicks, scrolls, form field, check boxes, and hovers in a lab usability study of opt-out and data deletion interfaces [21].

4.2.4 Perceived Effort in Making a Privacy Choice

After completing a task that requires using a privacy choice interface, participants can be asked questions related to the perceived ease or difficulty of their experience. Alternatively, these questions can be asked about participants’ prior experiences with a privacy choice interface outside of the study environment. Work by Tsai et al. and Habib et al. reported perceived effort by asking participants a version of the the Single Ease Question (SEQ) (“Overall, how easy or difficult was it to perform this task?”) to evaluate different privacy choice interfaces [21, 22, 60]. Other commonly used prompts that measure perceived effort on a Likert scale include items 2, 3, 4, and 8 on the System Usability Scale (SUS) [35].

4.2.5 Estimated Effort Required to Make a Choice

Expert evaluation approaches can be used to estimate users’ ability and effort in using a privacy choice interface to accomplish a particular goal. Such evaluations may include a set of design heuristics specific to the privacy choice interface or established usability heuristics (e.g., items 1-3, 7, 8 of the Nielsen heuristics [46]). Estimating ability and effort could also be done in conjunction with 4.2.2 and 4.2.3, as it may be helpful to compare the ability and effort of an “expert” with prior knowledge of the privacy choice interaction to those of user study participants. Habib et al. estimated the effort involved in using privacy opt-outs and data deletion mechanisms by counting the user actions in the shortest interaction path required to opt-out or delete data [24].

4.3 User Awareness

For privacy choice interfaces to be usable, it is necessary to ensure that users recognize that the privacy choice(s) exist and that they are able to find them. Awareness may be measured together or separately from user ability & effort (Section 4.2) as it is part of the interaction required to use a privacy choice interface. Testing for awareness may be less important for interfaces that interrupt the user’s primary goal, compared to

on-demand privacy settings pages that users must seek out. Furthermore, for step-wise privacy choice interfaces, in which choices are incrementally revealed, it may be sufficient to evaluate whether users are aware of the general types of options available, rather than every option offered in the interface.

4.3.1 Awareness of Choice Existence

Assessing awareness of privacy choice interfaces and available options, sometimes referred to as *discoverability* [3], requires study participants to have prior experience with the system but not necessarily the particular interface being evaluated. Thus, self-report evaluations or user studies with distraction tasks are appropriate for evaluating awareness. For interruptive interfaces, evaluating for this criterion might include whether participants can recall the specific choice interface or available privacy options, whether participants realized they were asked to make a privacy choice during a distraction task, and if can they identify which choice they made. For on-demand privacy choices, users might be asked about their own privacy objectives or told about objectives that some users have, and then asked whether they think there is an interface that might help them achieve this objective (as a follow-up researchers may then assess the users’ ability to find it). An example of prior work measuring awareness is Cranor et al.’s study that evaluated whether participants noticed an opt-out link and icon present on the page [12].

4.3.2 Ability to Find Privacy Choice

This criterion can be incorporated into user studies that implement the criterion described in 4.2.1, as finding the privacy choice interface is typically the bulk of a privacy choice interaction for on-demand privacy choices. It may include assessing whether participants were able to find the choice interface without assistance, and for moderated studies, what hints aided participants in finding the privacy choice.

4.3.3 Time Taken to Find the Privacy Choice

Similarly, this criterion can be studied with the criterion described in 4.2.2. For example, Garlach and Suthers report the time taken by their study participants to find the AdChoices icon on a mobile device [18].

4.3.4 User Actions Taken to Find the Privacy Choice

Participants’ interaction path while trying to find the privacy choice can also be studied alongside the criterion in 4.2.3.

4.3.5 Perceived Effort in Finding the Privacy Choice

This criterion is similar to that described in 4.2.4. After completing a study task that requires participants to seek out the privacy choice interface, participants can be asked questions

related to the perceived ease or difficulty in finding the privacy choice. For example, participants in Chen et al.'s study were asked to rate the difficulty of finding different app privacy settings [8]. Alternatively, participants can be asked about prior experiences with a privacy choice interface outside of the study environment in self-report studies.

4.3.6 Estimated Effort in Finding the Privacy Choice

Expert evaluation approaches can be used to estimate the difficulty of finding a privacy choice. Cognitive walkthroughs of the system may be especially relevant when evaluating the learnability of the privacy choice interaction [66]. Established usability heuristics (e.g., items 4 and 6 of the Nielsen heuristics [46]) also address findability. Estimating effort in finding the privacy choice could be done in conjunction with 4.3.3 and 4.3.4. Similar to the criterion described in 4.2.5, comparing the ability of an “expert” with prior knowledge of the system with those of study participants to find the privacy choice interface may suggest usability issues in the interaction if there is a large gap.

4.4 User Comprehension

For a privacy choice interface to be effective, it is important to ensure that users understand what it does and identify any misconceptions. When evaluating for comprehension, it is important to evaluate whether users understand the options that are available to them and the implications of their decision, given their (often) incomplete understanding of the technologies relevant to the privacy choice.

4.4.1 Objective Knowledge with Focused Attention

To better understand whether users can comprehend information provided in a privacy choice interface (either interruptive or on-demand), user study participants can be asked objective knowledge questions when it is presumed that their attention was focused on the privacy choice interface. This criterion can be assessed through user studies that involve privacy tasks, as well as self-report studies that ask participants to recall their experience with the privacy choice interface being evaluated. Koelle et al.'s study evaluating opt-in and opt-out gestures assessed objective knowledge by asking “What does the gesture shown in the video above mean to you” [31]. Evaluating for objective knowledge could also include asking if participants understand the privacy benefits and risks associated with different options, and if applicable, whether participants recognize whether a privacy choice is optional or mandatory.

4.4.2 Objective Knowledge with Unfocused Attention

It is also important to assess whether users understand the options available to them and implications of a decision made

through interruptive privacy choice interfaces that they encounter when their attention is focused elsewhere in their interactions with a system. Similar to measuring awareness of a privacy choice described in 4.3.1, measuring objective knowledge with unfocused attention might require assigning participants to a distraction task, or having them recall their past experiences in a self-report study. Comparing objective knowledge when attention was focused on the privacy choice interface to when it was focused elsewhere may also help to reveal comprehension issues. For example, Pearman et al. asked participants about practices described in a HIPAA authorization they had encountered while trying to use a chatbot as part of a distraction task and later asked them to review the authorization again and revisit their answers [51].

4.4.3 Perceived Effort in Comprehending Choices

Similar to assessing the perceived effort to make a privacy choice (4.2.1), user study participants can be asked questions related to the perceived ease or difficulty in learning or comprehending the privacy choices. Similarly, this criterion can be assessed for both interruptive and on-demand interfaces after completing a study task that exposed them to the privacy choices. Alternatively, these questions can be asked about participants' prior experiences with a privacy choice interface outside of the study environment. Example prompts and measures to evaluate perceived learnability include: “what (if anything) was difficult to understand about the privacy choice interface” and items 5, 6, 7 and 10 on the SUS [35].

4.4.4 Estimated Effort in Comprehending Choices

Similar to the criteria described in 4.2.5 and 4.3.6, expert evaluation approaches can assess the difficulty in learning or comprehending a privacy choice interface. Such evaluations may assess whether particular types of users might have greater difficulty in learning or comprehending what the choice interface does, as well as what aid might be required to learn available choices. Furthermore, item 10 of the Nielsen heuristics also pertains to learnability [46].

4.5 User Sentiment

Different facets of user sentiment assess users' satisfaction with a privacy choice interface after they have had some exposure to it. This exposure may occur through a study task, or during their past interactions with a system. Evaluating for sentiment is applicable to both interruptive and on-demand privacy choice interfaces, and may be assessed through Likert measures accompanied with qualitative prompts.

4.5.1 Perceived Transparency & Control

This criterion assesses whether the privacy choice interface provides an appropriate level of transparency and control re-

lated to how user data is handled. Participants may be asked how transparent they feel the evaluated privacy choice interface is related to the use of their data, and to what extent they feel that it provides sufficient control over their data.

4.5.2 Subjective Knowledge

Assessing for subjective knowledge involves capturing users' interpretations of their ability to effectively use the privacy choice interface, as well as if they experience feelings of regret. Korff and Böhme used the TMC scale to measure participants' satisfaction, regret, and feelings of being overwhelmed after interacting with a privacy choice interface related to disclosure on a business networking website [32]. Example prompts and metrics related to subjective knowledge include to what extent participants feel informed about their choices, how capable they feel in making a decision, and how confident they are in their privacy choice (e.g., item 9 of SUS [35]).

4.5.3 Levels of Comfort and Trust

Ideally, privacy choice interfaces should empower users by providing control over their data. Thus it is important to evaluate whether after interacting with a privacy choice interface users are comfortable with how their data will be used, as well as to what extent they feel that their privacy decision will be honored. Mathur et al. argue that privacy choice interfaces should be evaluated on whether they are detrimental to the collective welfare [41]. In the context of privacy choice interfaces, dark patterns may result in a loss of trust or skepticism (e.g., in the company, in companies using similar privacy choice interfaces), and could contribute to feelings of resignation. Korff and Böhme also used the PCRT scale to measure participants' perceived comfort, risk, and trust in the privacy choice interface evaluated [32].

4.5.4 Investment in Decision-Making

This criterion pertains to whether the design of the privacy choice interface sufficiently motivates users to make an informed privacy decision. An example of prior work that assessed investment in decision-making is Cranor et al.'s user study that asked participants how likely they would be to click on the do-not-sell icon and link texts being evaluated [12]. Other means of measuring investment include asking participants how carefully they considered their privacy choice and describing how they made their privacy decision.

4.6 Decision Reversal

For privacy choices to be usable, users need to be able to change their privacy choice decision, both immediately after an interaction with a privacy choice interface and, if applicable, at a later time through user settings offered through the website or app. This allows for users to correct an error they

may have made in their initial privacy choice as well as circumstances in which users change their mind about how their data may be used or collected. The criteria for evaluating user ability & effort described in Section 4.2 related to making an initial privacy choice can be adapted to measure users' ability and effort in reversing their privacy decision (both immediately after making an initial decision and at a later point in time in which the choice interface or a settings page must be revisited). Similarly, those related to user awareness (Section 4.3 and user comprehension (Section 4.4) can be utilized to ensure that users can find and understand the information and processes that are part of reversing their privacy decision. Assessing for reversal through user studies involves assigning participants a privacy choice task in which they must undo or modify their initial privacy choice. This aspect of usability is applicable to both interruptive and on-demand interfaces.

4.7 Nudging Patterns

In contrast to the other usability aspects that are applicable to almost any type of user interface, evaluating for nudging patterns is especially relevant to contexts in which users are asked to give up something, such as their personal data. Privacy choice interfaces often exhibit dark patterns that nudge users to less privacy-protective outcomes to the benefit of the company. This usually occurs when privacy-protective options are made less salient or more cumbersome to use than the alternatives. Furthermore, legislation such as the General Data Protection Regulation (GDPR) and the California Privacy Rights Act (CPRA) make the use of dark patterns in privacy choice interfaces, particularly those related to consent, illegal [14, 48]. As such it is important for designers to be aware of the way they are nudging consumers and evaluate whether this nudging could be a dark pattern. In some contexts, it may even be appropriate for interfaces to nudge users to privacy-protective choices [1]. To evaluate interruptive and on-demand privacy choice interfaces for dark patterns, we propose criteria aligned to the normative perspectives described by Mathur et al. with regards to privacy [41].

4.7.1 Impact of Individual Welfare

Mathur et al. suggest measuring a "welfarist conception of privacy" [41]. In the privacy choice context, one such calculation is the financial value of the data disclosed because of a particular design pattern. User studies involving study tasks or self-reporting of data could also examine the proportion of users whose needs were not satisfied by a particular design. These measures could also highlight whether individual welfare could be improved with nudges toward privacy-protective choices. An example of prior work that has explored impact to individual welfare is Nouwen's et al.'s experiment that quantified the impact of different design elements in cookie consent interfaces on participants' consent decisions [47].

4.7.2 Unintended Societal Consequences

Another aspect of collective welfare is analyzing through expert evaluation approaches whether the privacy choice interface could lead to unintentional disclosure of personal information, and whether this could have negative societal-level impact. A prominent example is Facebook users unknowingly consenting to their data being shared with Cambridge Analytica, which used the data to influence global elections [42]. Gray et al.’s interaction criticism incorporated potential societal impact in a usability evaluation of cookie consent interfaces by including considerations such as “relevant business models and economic rationale, current and future role of technology, social acceptance or rejection of technology norms, agency of users and technology providers” [19].

4.7.3 Alignment with Regulatory Objectives

Expert evaluation approaches can also be used to ensure that designed privacy choice interfaces meet regulatory requirements. Both the GDPR and CPRA have provisions related to the usability of privacy choice interfaces, particularly to the consent of data collection [14, 48]. The CPRA explicitly bans dark patterns, defining them as “a user interface designed or manipulated with the substantial effect of subverting or impairing user autonomy, decision-making, or choice, as further defined by regulation.” [48]. Prior empirical evaluations of consent notices have identified dark patterns that likely violate the spirit of GDPR and could potentially lead to regulatory penalties. Particularly Nowens et al. and Soe et al. provide a list of design criteria for cookie consent notices to evaluate for the presence of dark patterns and potential violations of the GDPR [47, 56]. This includes that consent be explicit (e.g., require a click from the user), consent must be as easy to withdraw or refuse as it is to give, and the privacy choice interface contain no pre-selected boxes for non-necessary purposes [47]. Other potentially violating design patterns are the absence of actual choices in the interface (e.g., instructions to change privacy choices are simply described in a notice text), choice toggles that are unlabelled, and not using antonyms of the consent option to label the option denying consent [56].

4.7.4 Individual Autonomy

Mathur et al. suggest evaluating to what degree an interface interferes with a user’s ability to make “independent decisions” [41]. User study approaches can evaluate whether privacy choice interface designs lead users to choose certain privacy options over others by comparing privacy options selected through interfaces with suspected nudging patterns with those selected through other designs; cookie consent interface evaluations by Machuletz and Böhme [38] and Nouwens et al. [47] took this approach. Similarly, in some contexts it may be beneficial to evaluate whether interfaces utilizing

reflective design better enable individual autonomy, as suggested by Terpstra et al [59]. Individual autonomy could also be evaluated through criteria that align with other evaluation objectives including: whether there is an option aligned with users’ preferences available (4.1.1), whether users are able to choose their preferred option and the effort required (4.2.1, 4.2.2, 4.2.3), whether users are aware of the options available (4.3.1), whether users comprehend available options (4.4.1 and 4.4.2), and perceptions of autonomy (4.5.2 and 4.5.4).

5 Previous Privacy Choice Evaluations

This section presents an overview of a range of prior studies evaluating different types of privacy choice mechanisms. Though other work in this space may also be beneficial in informing the design of privacy choice interfaces, the studies described illustrate facets of the Privacy Choice Evaluation Framework through a variety of approaches. We focus our review on studies published over the past 10 years, with most published in the past five years.

A common privacy choice interface is related to allowing access to a specific hardware resource obtained from a device, like camera or location data. Previous studies have focused on user needs related to permission management in different contexts — including smartphone apps [26, 50], smart speakers [58], and smart glasses [13] — offering insights into the types of privacy controls that users desire. Other studies have uncovered limitations related to users’ ability to use and comprehend existing permission management schemes [7, 55], or compared their usability to alternative approaches [60, 65]. Additionally, Bahirat et al. evaluated the impact of nudging on smart home privacy choices using data collected through hypothetical contextual scenarios, finding that defaults and framing of choices impact users’ decision-making [4].

Interfaces that allow individuals to consent to different types of data processing are often used to meet legal requirements, such as those set by GDPR, Health Insurance Portability and Accountability Act (HIPAA), and Family Educational Rights and Privacy Act (FERPA). A growing body of work has explored the usability of cookie consent interfaces, finding that dark or nudging patterns that impact users’ choices are prevalent in current interface designs [6, 19, 20, 47, 56, 61]. Others have explored the usability of consent interfaces used in other contexts. For example using an inspection-based approach, Khalil et al. found that students’ ability to withdraw consent from Massive Open Online Course (MOOC) providers are limited due to lack of available options [30]. Additionally, Pearman et al. explored the usability of different health data disclosure authorization designs for a healthcare chatbot and argued for alternative approaches to capturing informed consent [51].

In contrast, other types of privacy choice interfaces allow users to opt out of the processing of their data, or to request deletion of their data. However, opt-out and deletion mecha-

nisms commonly used on websites and apps have been found to have usability issues related to awareness and ability & effort [18, 21, 24]. Other studies have explored visual icons as a potential means of increasing awareness of available opt-out choices through different user study designs, including participant inspection and assignment of a distraction task [12, 25]. Data deletion mechanisms have also been studied in the context of smart speakers; while users were found to be unaware of existing deletion options [40] the presence of available deletion mechanisms impacted users' trust in the system [9].

Privacy choice interfaces can also take the form of settings offered by a platform that users must typically seek out. Many studies have explored the usability of audience-related settings on content sharing and social media platforms, including user needs for audience control settings [16, 29, 57], tools that improve awareness of such settings [8], and users' ability to effectively use settings [28, 37, 39, 64]. Beyond audience settings, other studies have explored user needs for Facebook advertising controls through participant inspection [22], as well as the impact of social nudges on users' choice of Facebook privacy settings [33]. Outside of the social media context, Frik et al. collected self-reported data to explore the usability of smartphone privacy settings, highlighting issues of awareness and comprehension, among other usability issues [17].

Users can also make privacy choices through mechanisms decoupled from the original point of data collection. Past work utilizing participant inspection approaches has found that browser extensions may be effective in helping users become aware of available privacy opt-outs [5] and set their ideal privacy settings [34], but has highlighted that extensions themselves may be difficult for some users to configure [36]. Others have evaluated user needs for smartphone apps designed to aid privacy decision-making in different contexts [2, 11]. Furthermore, as traditional privacy choice mechanisms may be ill suited for some data collection scenarios by Internet of Things devices, others have explored the usability of alternative choice mechanisms [67] such as opt-out hand gestures [31, 68].

Altogether, this past work demonstrates the challenges of designing usable privacy choice mechanisms. Privacy decisions, such as content sharing, can be highly contextual [23]. Privacy choice interfaces must effectively communicate the scope of the privacy choice to allow users to make informed decisions [15]. The Privacy Choice Evaluation Framework presented can guide organizations in evaluating for aspects of usability pertinent to a particular privacy choice context.

6 Discussion

The Privacy Choice Evaluation Framework draws on evaluation approaches used in prior work to provide criteria to comprehensively evaluate the usability of privacy choice interfaces. The framework takes into account several considerations that make privacy choice interactions distinctive from in-

teractions with other types of interfaces. The criteria provided in the framework can help guide organizations in evaluating new and existing privacy choice interface designs, which are necessary to support effective consumer privacy protection.

6.1 Additional Considerations

Privacy choice interactions differ from other interactions in that users are typically not trying to achieve a privacy goal when they interact with a system. Thus, the way they interact with privacy choice interfaces will be heavily impacted by their primary goal, such as to use a website or app. This is particularly relevant for interruptive privacy interfaces, such as cookie consent banners, which users may be inclined to quickly dismiss. This creates a tension between usability and privacy; while such interfaces may impede users in their primary goal and worsen the overall usability of a system, they can force users to make a privacy decision and offer an opportunity to select privacy-protective options that they would not have set otherwise. Furthermore, when evaluating privacy choice interfaces it is important to consider that users' behaviors and attitudes toward such interfaces are heavily influenced by their past experiences with similar privacy choices. Users may form expectations about where to find certain privacy choices and how they function [21]. Additionally, achieving meaningful privacy choice for some choice contexts in which users are overexposed to choice interfaces might require overcoming habituation and privacy fatigue [10].

The research methods described in the framework describe how general approaches to usability testing can be adapted to evaluate privacy choice interfaces. To ensure that meaningful privacy choice mechanisms are available to a broad population of internet users with differing abilities, evaluations utilizing these approaches should be performed in conjunction with accessibility assessments for which there are established frameworks [63]. In addition to users with disabilities, it is important to evaluate certain privacy choice interfaces with other vulnerable populations, such as marginalized racial groups or gender identities. Not only might these groups have specific privacy needs on a platform, the way they use existing privacy choice interfaces may differ from other users. An expert evaluation could provide an initial understanding of the usability of privacy choice interfaces for a special population. User studies with participants recruited from these special populations should be conducted to further this understanding.

6.2 Guidance for Organizations

A detailed example of how organizations can apply the Privacy Choice Evaluation Framework for their own usability evaluations is provided in the appendix. The same criteria could also be applied in studies that compare multiple privacy choice interface designs to identify which design elements are beneficial or detrimental to different usability aspects. In

selecting evaluation approaches, several factors related to the organization conducting the evaluation and the interface being evaluated should be considered. Here we describe a few such practical considerations.

Design Stage of the Privacy Choice Interface: An important factor that impacts which types of usability evaluations of a privacy choice interface are suitable is where in the design process the evaluation is being conducted. Ideally, evaluating the usability of a particular design would be integrated into an iterative design process with multiple research methods so that usability issues can be addressed prior to the interface being deployed. These usability assessments should build on each other. For example, a usability assessment in the ideation design phase may involve using qualitative methods, such as interviews or focus groups, to better understand users' needs in the context of the privacy choice interface. Expert evaluations, online surveys, experiments, and lab usability studies may be conducted with prototypes of the privacy choice interface to assess how well users' needs are met, as well as to what extent other usability aspects, including ability & effort, awareness, and comprehension, are achieved. Once a privacy choice interface is deployed, expert evaluations and field studies may be used to confirm that the usability of the final design is similar to results from previous usability testing.

Data Needed for Organizational Decisions: When considering the scope of possible research methods for assessments of privacy choice interfaces, it is necessary to prioritize which and what type of data are most important to capture from an organizational perspective. For example, some organizations may have additional requirements related to privacy choice that must be examined through a usability evaluation and thus focus more on a subset of the described usability aspects. Furthermore, organizations may differ in how they weigh and use different types of data in design decision-making. User studies that involve empirical data, such as field studies, online experiments, or lab usability studies, typically provide the best representation of how users may perceive or react to a particular design once it is deployed. However, user studies involving self-reported data may still provide enough of this insight to help organizations move forward with certain decisions. Expert evaluations can also aid in organizational decision-making, particularly in contexts where user feedback may not be helpful (e.g., new technologies where the average user may not be aware of all possible interaction paths).

Availability of Resources: Another important consideration in planning usability evaluations is the time, budget, and skill set of the evaluation team. While expert evaluations are typically less costly than user studies in terms of time and budget, they require evaluators with specific legal, design, or privacy expertise. User studies involving primarily

quantitative data, such as surveys, can be deployed to a large number of participants (e.g., through online crowd-sourcing platforms) and analyzed in a short amount of time. Qualitative user studies may require more time for both data collection and analysis. Costs associated with user studies depend on factors such as the number of participants, length of the study, ease of recruiting qualified participants, amount of qualitative data to be analyzed, and depth of the analysis.

6.3 Limitations of Privacy Choice Usability

Better design of privacy choice interfaces, particularly those that allow users to decline data sharing just as easily as to agree to it, may be at odds with revenue-generating goals of a company. Though mounting consumer pressure should encourage companies to better privacy practices, it is still unclear whether this will translate to better consumer privacy protection. Privacy choice requirements in regulation, which include general requirements for usability, provide further incentive for companies to evaluate their privacy choice interfaces. While this framework could help organizations meet such usability requirements, and regulators to hold organizations accountable to better design practices, it is possible that interface designs that perform best in terms of usability (such as those that bundle certain privacy choices) would not be in full compliance with applicable legal requirements. Conversely, not all lawful designs of a privacy choice interface would perform well in meeting the framework's criteria.

Furthermore, even the most usable privacy choice interfaces place the burden of privacy management on users. In addition to privacy regulation, other mechanisms — such as technology supported decision-making and standardized privacy choice interfaces — are necessary to form a more effective consumer privacy protection framework. The Privacy Choice Evaluation Framework could serve as an initial step towards a more comprehensive implementation framework that could standardize interfaces for certain contexts. However until adoption of these privacy protection mechanisms becomes widespread, this framework provides immediately actionable guidance in improving privacy choice interfaces for users.

Acknowledgements

This research was supported in part by gifts from Facebook, the Carnegie Corporation of New York, and Innovators Network Foundation. We also would like to thank Alessandro Acquisti, Rebecca Balebako, Jessica Colnago, Yuanyuan Feng, Justin Hepler, Liz Keneski, Norman Sadeh, Hanna Schraffenberger, and Yixin Zou for their feedback on this work.

References

- [1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys*, 50(3), 2017.
- [2] Mamtaj Akter, Amy J. Godfrey, Jess Kropczynski, Heather R. Lipford, and Pamela J. Wisniewski. From parental control to joint family oversight: Can parents and teens manage mobile online safety and privacy as equals? *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), April 2022.
- [3] Nick Babich. Tips to improve discoverability in UX, April 2020. <https://xd.adobe.com/ideas/process/information-architecture/tips-to-improve-discoverability-in-ux/>.
- [4] Paritosh Bahirat, Martijn Willemsen, Yangyang He, Qizhang Sun, and Bart Knijnenburg. Overlooking context: How do defaults and framing reduce deliberation in smart home privacy decision-making? In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2021.
- [5] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference*, 2020.
- [6] Carlos Bermejo Fernández, Dimitris Chatzopoulos, Dimitrios Papadopoulos, and Pan Hui. This website uses nudging: Mturk workers' behaviour on cookie consent notices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021.
- [7] George Chalhoub, Martin J Kraemer, Norbert Nthala, and Ivan Flechais. "It did not give me an option to decline": A longitudinal analysis of the user experience of security and privacy in smart home products. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2021.
- [8] Yi Chen, Mingming Zha, Nan Zhang, Dandan Xu, Qianqian Zhao, Xuan Feng, Kan Yuan, Fnu Suya, Yuan Tian, Kai Chen, et al. Demystifying hidden privacy settings in mobile apps. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 570–586. IEEE, 2019.
- [9] Eugene Cho, S Shyam Sundar, Saeed Abdullah, and Nasim Motalebi. Will deleting history make Alexa more trustworthy? Effects of privacy and content customization on user experience of smart speakers. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2020.
- [10] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior*, 81:42–51, 2018.
- [11] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Informing the design of a personalized privacy assistant for the Internet of Things. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2020.
- [12] Lorrie Faith Cranor, Hana Habib, Yaxing Yao, Yixin Zou, Alessandro Acquisti, Joel Reidenberg, Norman Sadeh, and Florian Schaub. CCPA opt-out icon testing—phase 2. Technical report, Office of the California Attorney General, 2020. <https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/dns-icon-study-report-052822020.pdf>.
- [13] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 2377–2386. ACM, 2014.
- [14] European Parliament. Regulation (EU) 2016/679 of the European parliament and of the council, 2016. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.
- [15] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. A design space for privacy choices: Towards meaningful privacy control in the internet of things. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2021.
- [16] Casey Fiesler, Michaelanne Dye, Jessica L Feuston, Chaya Hiruncharoenvate, Clayton J Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S Bruckman, Munmun De Choudhury, et al. What (or who) is public? Privacy settings and social media content sharing. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pages 567–580. ACM, 2017.
- [17] Alisa Frik, Juliann Kim, Joshua Rafael Sanchez, and Joanne Ma. Users' expectations about and use of smartphone privacy and security settings. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2022.

- [18] Stacia Garlach and Daniel Suthers. ‘I’m supposed to see that?’ AdChoices usability in the mobile environment. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2018.
- [19] Colin M Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2021.
- [20] Hana Habib, Megan Li, Ellie Young, and Lorrie Faith Cranor. “‘Okay, whatever’”: An evaluation of cookie consent interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2022.
- [21] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. “It’s a scavenger hunt’’: Usability of websites’ opt-out and data deletion choices. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2020.
- [22] Hana Habib, Sarah Pearman, Ellie Young, Ishika Saxena, Robert Zhang, and Lorrie Faith Cranor. Identifying user needs for advertising controls on facebook. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), April 2022.
- [23] Hana Habib, Neil Shah, and Rajan Vaish. Impact of contextual factors on Snapchat public sharing. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2019.
- [24] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An empirical analysis of data deletion and opt-out choices on 150 websites. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. USENIX, 2019.
- [25] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. Toggles, dollar signs, and triangles: How to (in)effectively convey privacy choices with icons and link texts. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2021.
- [26] Qatrunnada Ismail, Tousif Ahmed, Apu Kapadia, and Michael K Reiter. Crowdsourced exploration of security configurations. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 467–476. ACM, 2015.
- [27] ISO Technical Committee 159. Ergonomics of human-system interaction, March 2018. <https://www.iso.org/standard/63500.html>.
- [28] Yousra Javed and Mohamed Shehab. Access control policy misconfiguration detection in online social networks. In *Proceedings of the International Conference on Social Computing (SocialCom)*, pages 544–549. IEEE, 2013.
- [29] Dilara Kekulluoglu, Kami Vaniea, and Walid Magdy. Understanding privacy switching behaviour on Twitter. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2022.
- [30] Mohammad Khalil, Paul Prinsloo, and Sharon Slade. The unbearable lightness of consent: Mapping MOOC providers’ response to consent. In *Proceedings of the Conference on Learning at Scale (L@S)*. ACM, 2018.
- [31] Marion Koelle, Swamy Ananthanarayan, Simon Czupalla, Wilko Heuten, and Susanne Boll. Your smart glasses’ camera bothers me! Exploring opt-in and opt-out gestures for privacy mediation. In *Proceedings of the Nordic Conference on Human-Computer Interaction (NordiCHI)*, pages 473–481, 2018.
- [32] Stefan Korff and Rainer Böhme. Too much choice: End-User privacy decisions in the context of choice proliferation. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS)*, pages 69–87, 2014.
- [33] Isadora Krsek, Kimi Wenzel, Sauvik Das, Jason I Hong, and Laura Dabbish. To self-persuade or be persuaded: Examining interventions for users’ privacy setting selection. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2022.
- [34] Oksana Kulyk, Peter Mayer, Melanie Volkamer, and Oliver Käfer. A concept and evaluation of usable and fine-grained privacy-friendly cookie settings interface. In *Proceedings of the International Conference on Trust, Security And Privacy in Computing and Communications/International Conference on Big Data Science and Engineering (TrustCom/BigDataSE)*, pages 1058–1063. IEEE, 2018.
- [35] Page Laubheimer. Beyond the NPS: Measuring perceived usability with the SUS, NASA-TLX, and the single ease question after tasks and usability tests, February 2018. <https://www.nngroup.com/articles/measuring-perceived-usability/>.
- [36] Pedro Giovanni Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Faith Cranor. Why Johnny can’t opt out: A usability evaluation of tools to limit online behavioral advertising. In *Proceedings of*

the Conference on Human Factors in Computing Systems (CHI). ACM, 2012.

- [37] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the Internet Measurement Conference (IMC)*, pages 61–70. ACM, 2011.
- [38] Dominique Machuletz and Rainer Böhme. Multiple purposes, multiple problems: A user study of consent dialogs after GDPR. *Proceedings on Privacy Enhancing Technologies*, 2020(2):481–498, 2020.
- [39] Michelle Madejski, Maritza Johnson, and Steven M Bellovin. A study of privacy settings errors in an online social network. In *Proceedings of the International Workshop on Security and Social Networking (SESOC)*, pages 340–345. IEEE, 2012.
- [40] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271, 2019.
- [41] Arunesh Mathur, Jonathan Mayer, and Mihir Kshirsagar. What makes a dark pattern...dark? Design attributes, normative considerations, and measurement methods. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2021.
- [42] Sam Meredith. Here’s everything you need to know about the Cambridge Analytica scandal. *CNBC*, March 2018. <https://www.cnbc.com/2018/03/21/facebook-cambridge-analytica-scandal-everything-you-need-to-know.html>.
- [43] Justin Mifsud. Usability metrics – a guide to quantify the usability of any system. <https://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/>.
- [44] Peter Morville. User experience design, June 2004. http://semanticstudios.com/user_experience_design/.
- [45] Jakob Nielsen. Usability 101: Introduction to usability, January 2012. <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>.
- [46] Jakob Nielsen. 10 usability heuristics for user interface design, November 2020. <https://www.nngroup.com/articles/ten-usability-heuristics/>.
- [47] Midas Nouwens, Iaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2020.
- [48] Office of the California Attorney General. The California Privacy Rights and Enforcement Act of 2020, 2019. <https://oag.ca.gov/system/files/initiatives/pdfs/19-0017%20%28Consumer%20Privacy%20%29.pdf>.
- [49] OneTrust. DHL increases CMP opt-in rates with A/B testing and OneTrust PreferenceChoice, June 2021. <https://www.cookiepro.com/wp-content/uploads/2021/06/20210421-OneTrust-DHL-CS-US-Digital.pdf>.
- [50] Sameer Patil, Roman Schlegel, Apu Kapadia, and Adam J Lee. Reflection or action? How feedback and control affect location sharing decisions. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 101–110. ACM, 2014.
- [51] Sarah Pearman, Eleanor Young, and Lorrie Faith Cranor. User-friendly yet rarely read: A case study on the redesign of an online HIPAA authorization. *Proceedings on Privacy Enhancing Technologies*, 2022(3), 2022.
- [52] Whitney Quesenbery. Balancing the 5Es: Usability. *Cutter IT Journal*, February 2004. <http://whitneyquesenbery.com/articles/5es-citj0204.pdf>.
- [53] John A Rothchild. Against notice and choice: The manifest failure of the proceduralist paradigm to protect privacy online (or anywhere else). *Cleveland State Law Review*, 66, 2017.
- [54] Florian Schaub and Lorrie Faith Cranor. Usable and useful privacy interfaces. In Travis Breaux, editor, *An Introduction to Privacy for Technology Professionals*, pages 176–299. IAPP, 2020.
- [55] Muhammad Umair Shah, Umair Rehman, Farkhund Iqbal, Fazli Wahid, Mohammed Hussain, and Ali Arsalan. Access permissions for Apple Watch applications: A study on users’ perceptions. In *Proceedings of the International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 2020.
- [56] Than Htut Soe, Oda Elise Nordberg, Frode Guribye, and Marija Slavkovik. Circumvention by design - dark patterns in cookie consent for online news outlets. In *Proceedings of the Nordic Conference on Human-Computer Interaction (NordiCHI)*, 2020.
- [57] Katherine Strater and Heather Richter. Examining privacy and disclosure in a social networking community. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2007.

- [58] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. Investigating users' preferences and expectations for always-listening voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 3(4), 2019.
- [59] Arnout Terpstra, Paul Graßl, and Hanna Schraffenberger. Think before you click: How reflective patterns contribute to privacy. In *Proceedings of the What Can CHI Do About Dark Patterns Workshop*. ACM, 2021.
- [60] Lynn Tsai, Primal Wijesekera, Joel Reardon, Irwin Reyes, Serge Egelman, David Wagner, Nathan Good, and Jung-Wei Chen. Turtle Guard: Helping android users apply contextual privacy preferences. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2017.
- [61] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed consent: Studying GDPR consent notices in the field. In *Proceedings of the Conference on Computer and Communications Security (CCS)*, pages 973–990. ACM, 2019.
- [62] Kami Vaniea, Lujó Bauer, Lorrie Faith Cranor, and Michael K Reiter. Studying access-control usability in the lab: Lessons learned from four studies. In *Proceedings of the Workshop on Learning from Authoritative Security Experiment Results*, pages 31–40, 2012.
- [63] W3C Web Accessibility Initiative. Web content accessibility guidelines (WCAG) 2.1, 2018. <https://www.w3.org/TR/WCAG21/>.
- [64] Yang Wang, Liang Gou, Anbang Xu, Michelle X Zhou, Huahai Yang, and Hernan Badenes. Veilme: An interactive visualization tool for privacy configuration of using personality traits. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 817–826, 2015.
- [65] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2018.
- [66] Chauncey Wilson. *User Interface Inspection Methods: A User-Centered Design Method*. Newnes, 2013.
- [67] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2019.
- [68] Yaxing Yao, Huichuan Xia, Yun Huang, and Yang Wang. Privacy mechanisms for drones: Perceptions of drone controllers and bystanders. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 6777–6788. ACM, 2017.

A Using the Evaluation Framework

Table 1 provides a summary of the criteria included in the Privacy Choice Evaluation Framework. The mapping of the criteria to usability aspects, evaluation approaches, and interface timings was informed by prior work and standard HCI practices. Organizations and other researchers can use this table as a reference when designing evaluation studies. First, researchers should identify their study goals, or the usability aspect(s) that they want to explore in their usability study. Then researchers should identify an evaluation approach that is suited for addressing their study goals, taking into account the considerations described in Section 6.2. For more comprehensive usability evaluations, researchers may choose to incorporate multiple evaluation approaches into their study. Last, the researchers should select the criteria that are appropriate for the particular privacy choice context, taking into account the timing of the choice interface being evaluated.

In prior work, we used the Privacy Choice Evaluation Framework to assess the usability of 12 cookie consent banner variants, touching on a large fraction of the criteria [20]. This may serve as a useful example for understanding how the framework might be used.

Here we describe a scenario where you might want to do a fairly comprehensive evaluation. Imagine you are working for a company developing a new skincare app that allows users to take photographs of their skin, get recommendations for skincare products, get referrals to dermatologists and skincare professionals, and discuss skincare issues with other users.

As you begin developing the app, you conduct focus groups to understand the interests of potential users. During this phase, it would be a good idea to also focus on users' privacy **needs** by conducting interviews, focus groups, or surveys to uncover *users' privacy objectives*, including the types of privacy choices they would like to have and whether there are any special requirements for this population of users — who might include acne-prone teenagers under age 18 and people who suffer from chronic skin conditions or are experiencing skin problems as a side effect of treatments for other conditions. Here it would be useful to find out whether course-grained controls over data sharing would meet users' needs or if (some) users would appreciate finer-grained controls over the type of data to be shared, with whom it is shared, or other privacy objectives. In this phase you may discover that some users have little sensitivity about discussing certain types of skincare concerns and are interested in getting advertisements and discounts on relevant products, while other users are interested in getting advice from experts and other users with

their condition, but are concerned about being identified as a person with a particular condition and do not want to receive related advertisements.

As low- to mid-fidelity prototypes of the app are developed (such as static wire frames or interactive prototypes), user studies can probe other aspects of the framework with participants representative of those expected to use the app, including the special populations identified. For example, a lab or online study might present prototypes to participants, ask them to step through some typical non-privacy tasks, and then ask them about what information they believe is being shared and what privacy choices are available to probe **awareness**, particularly *awareness of choice existence*. Then participants might be directed to *find the privacy choice* interface and *make privacy choices* they would like to have to evaluate their **ability & effort** using an **on demand** privacy choice interface. Researchers may want to ask participants about their *privacy objectives* and *privacy intentions* to confirm that choices meet the participant's needs and to see whether the choices the participant made align with their stated intentions. Participants might also be asked questions related to **comprehension** to assess their *objective knowledge* of the privacy choices available and what they do. Finally, participants may be asked questions pertaining to **sentiment**, for example to assess their *investment in decision-making*, *perceived levels of transparency and control over how their data will be used*, *self-efficacy in using the choice interface*, and *comfort and trust in the company's handling of their data*.

If the app includes a feature with an **interruptive** privacy choice interface, such as a prompt for the user to immediately make a decision about whether an uploaded photograph will be shared, users should be asked to perform a task that triggers the interruption and then similar evaluations should be conducted as with the on-demand interface. Here participants might be asked **comprehension** questions to assess their *objective knowledge* of the privacy choices available and what they do, both after completing the choice task with the choices no longer visible on screen (*unfocused attention*), and when revisiting the choice interface (*focused attention*). The evaluation of **user sentiment**, such as *investment in decision-making*, here is even more relevant than in the on-demand

task, as it allows an assessment of whether participants were trying to make a meaningful decision at the time the choice appeared or just swatting the prompt away. To evaluate **decision reversal**, participants may be asked what they would do if they wanted to change their privacy decision. This user study data might also be helpful for evaluating for potential **nudging patterns**, particularly whether the interface designs hinders *individual welfare* or *individual autonomy*.

A usable privacy **expert** may evaluate the privacy choice interface for **user needs, ability & effort, awareness, and comprehension**. An expert may examine *interface completeness* and *interface accuracy* related to the needs uncovered in prior evaluations, and also *estimate the effort needed to make a privacy choice*, *users' abilities to find the privacy choice*, and *comprehension of the choices*. A privacy legal expert might evaluate for potential **nudging patterns** by examining *alignment with regulatory objectives*, including any relevant laws concerning sensitive health information or children's privacy, as well as any *unintended societal consequences* of the interface.

As app development proceeds, some of these studies would be repeated with higher fidelity prototypes and eventually the finished app. Where potential problems are uncovered, alternate interfaces might be tested and compared. In some cases a very narrow study might be done to focus on a specific problem, for example, if users are having trouble understanding a particular privacy choice, an online survey might just probe comprehension of alternative ways of describing that choice. Once improved language is identified it should then be tested in the full app context.

The number of study participants and number of rounds of iteration will vary depending on the complexity of the app, number of problems surfaced in the initial studies, resources available, and objectives of the app developers. Different levels of rigor are needed for published academic papers than for internal testing. However, a company that is under regulatory scrutiny, trying to hold itself up as a privacy role model, or planning to publish the results of its internal testing may engage in more rigorous testing than a company that just wants to do enough testing to avoid major privacy pitfalls.

Detecting iPhone Security Compromise in Simulated Stalking Scenarios: Strategies and Obstacles

Andrea Gallardo
Carnegie Mellon University

Hanseul Kim
Carnegie Mellon University

Tianying Li
Carnegie Mellon University

Lujo Bauer
Carnegie Mellon University

Lorrie Cranor
Carnegie Mellon University

Abstract

Mobile phones can be abused for stalking, through methods such as location tracking, account compromise, and remote surveillance. We conducted eighteen remote semi-structured interviews in which we presented four hypothetical iPhone compromise scenarios that simulated technology-enabled abuse. We asked participants to provide advice for detecting and resolving each type of compromise. Using qualitative coding, we analyzed the interview data and identified the strategies of non-expert participants and the difficulties they faced in each scenario. We found that participants could readily delete an app and search in iOS settings or the home screen, but they were generally unable to identify or turn off location sharing in Google Maps or determine whether the iCloud account was improperly accessed. When following online advice for jailbreak detection, participants had difficulty finding a root checker app and resetting the phone. We identify underlying factors contributing to these difficulties and recommend improvements to iOS, Google Maps, and online advice to reduce the difficulties we identified.

1 Introduction

Mobile phones can be abused to enable stalking through methods such as location tracking, account compromise, and remote surveillance. For example, victims of intimate partner violence (IPV) may experience such technology-enabled abuse [23, 35, 53, 61, 62]. While experts can help victims detect and recover from technology-enabled abuse, little is known about the ability of victims to do this on their own, with the assis-

tance of non-experts in their social support network, or with the assistance of online educational materials.

We developed four hypothetical iPhone compromise scenarios that simulated technology-enabled abuse based on real-world scenarios faced by IPV victims and other victims of stalking. To gain insights into how non-experts in victims' social support networks might help them, we conducted 18 remote interviews in which we presented these scenarios and asked iPhone users recruited from Craigslist how they would help a friend detect and resolve each security compromise.

We found that while these non-expert participants were familiar with the iOS user interface (UI), most had difficulty detecting and resolving the problems simulated in our scenarios. For example, participants had difficulty associating Google Maps with location sharing controls. The challenges participants encountered were caused by discoverability issues in iOS and Google Maps UIs, such as a lack of indicators showing that another device has iCloud account access or that location is being shared with another user, as well as an absence of features that would help users know whether apps could be used to monitor them. We also found that online advice on detecting jailbreaking and resetting an iPhone often had impractical, inaccurate, or jargon-filled instructions.

Our paper makes the following novel contributions:

- Identifying strategies used by non-experts and specific difficulties they face, such as pinpointing the app transmitting the device's location to another user, as they attempt to detect and resolve four types of security compromise characteristic of technology-enabled abuse;
- Identifying underlying factors, e.g., lack of persistent notifications, contributing to difficulties we identified, most of which may be applicable across apps and platforms;
- Recommending specific changes to iOS, Google Maps, and online advice that would likely reduce the difficulties we identified and make it easier for non-experts to detect device or account compromise; and

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

- Highlighting the need to consider the stalking threat and victims’ ability to use devices when developing apps.

2 Related Work

In this section, we review prior work on technology-enabled abuse and interventions and online advice to prevent it.

2.1 IPV and Stalking Threat Model

Our study focuses on scenarios in which a malicious user leverages features of iOS, iCloud, and apps downloaded from the Apple App Store (“App Store”), for the purpose of stalking. People may experience such stalking in the context of IPV, also known as domestic violence or domestic abuse, but anyone with access to a victim’s mobile phone (e.g. coworkers, employers, roommates, relatives) may carry out these kinds of non-sophisticated (but often difficult to detect) attacks.

Exploitation of Technology by Abusers. The threat model of technology-enabled abuse faced by IPV victims does not require technical sophistication and is characterized by adversaries limited by the functionality of the system’s UI [23, 33, 35, 62]. Technology-enabled abusers can take advantage of access to devices and accounts by initially setting them up, enabling features or downloading apps for surveillance. They can also compromise security by guessing passwords or answers to security questions, or threatening or coercing the victim into giving them access to devices, accounts and their live location data [23, 28, 34, 35, 47].

Legitimate Apps Used for Abuse. Some apps have legitimate use purposes (e.g., navigation, anti-theft tracking, or child or employee monitoring) and can be downloaded from mobile app stores, but they can also be used for illegal or harmful purposes, such as stalking or spying [13, 32, 41, 45]. Though sometimes marketed as safety products that should not be used for abuse, they appear in search results for phrases like “track my girlfriend” and may be profitable as stalking tools [13, 16, 20, 46, 59]. Due to their valid purposes and legal ambiguity around use-cases, such “dual-use” apps will likely continue to be allowed on app stores [13, 55]. Researchers have begun developing tools to detect these apps, using machine learning classifiers and graph mining algorithms, and to warn people about potential surveillance [22, 29, 45]. In our study, we challenged participants to detect such an app.

Risk of Escalation. While security assessment tools and interventions can empower victims countering coercive control in tech abuse contexts, they can create new problems or burdens in abusive or coercive situations [48, 61]. Certain behaviors, such as cutting off surveillance methods, may risk endangering the survivor by escalating violence [21]. No prior research has tested general population awareness regarding such risks, so we included a question in our study to do so.

Seeking Advice from Friends. Research shows that most victims disclose abuse to at least one informal social support

network member (e.g., friend or family member) [25, 44, 49]. Research on technology-enabled abuse typically does not investigate the ability of these social supports, who are unlikely to be experts in identifying security compromise, to detect and remediate technology-enabled abuse [23, 34, 35].

Need for Usable Tools. IPV advocates have reported insufficient expertise to support victims of tech abuse, due to little to no training in preventing technology-based abuse [29, 35, 47]. Technical and clinical interventions have been developed, and online resources published, to help IPV and stalking victims [4, 6, 11, 24, 29, 57]. While researchers have suggested making usability improvements to UIs [22, 35, 40], they have not detailed specific usability problems. Our study surfaced consistent usability problems and areas for improvement not specifically identified by prior work.

Jailbreaking: Less Common but Dangerous. While increased usability and interventions may help counter unsophisticated attacks, a jailbroken iPhone presents a less common but more sophisticated threat that, if undetected, could endanger victims. Jailbreaking allows downloading apps banned by the App Store, such as spyware, and enables the ability to hide apps, potentially turning smartphones into surveillance devices. [18, 26, 31]. Prior work shows that victims consider the potentially dangerous risk of hidden surveillance when deciding whether to keep, replace, or destroy their devices [29, 35].

2.2 Online Security Advice for Survivors

IPV survivors have expressed a desire to learn more about privacy and to have more control over their digital assets, as well as dissatisfaction about using internet searches to do so [24]. Prior work has shown how unclear advice can make it difficult to assess technology-enabled security threats [54]. Research on information-seeking behaviors and responses to security advice suggests that non-tech savvy users may not effectively prioritize or follow security advice [37, 42, 43].

Many online articles provide advice on how to detect and prevent technology-enabled abuse, including jailbreaking and stalkerware [38, 51, 52, 58]. Advice varies in format and depth, from general advice [2] to concise lists of action items [14, 15] to step-by-step instructions [7, 38]. Some advice is technical and may be too complicated for non-tech-savvy users [36, 50], and some articles are outdated, recommending a jailbreak detection app that is no longer available on the App Store [5, 19]. Given the plethora and variety of such online advice, insight is needed into obstacles faced in implementing this advice. In one of our scenarios, we presented participants with two online articles and evaluated how easy it was for them to follow advice on detecting jailbreaking on an iPhone.

3 Methodology

In our interview study, participants encountered four hypothetical scenarios that reflect risks faced by victims of IPV and

stalking: location tracking, apps with remote access, account compromise, and jailbreaking. In this section, we describe our participant recruitment process, scenario and interview design, analysis process, and the limitations of our methodology.

3.1 Recruitment

We recruited participants who were “interested in mobile phone security” to participate in a 45 to 60 minute interview through Craigslist’s “Computer Gigs” section for three cities, Los Angeles, New York, and Pittsburgh, and offered \$20 as compensation. We screened for the following criteria: at least 18 years old, located in the U.S., fluent in English, has access to a device that can connect to internet, can run Zoom, and uses an iPhone. We also collected basic demographic information (see Appendix C) to diversify the sample demographics to be reflective of the U.S. population. Our screening survey received 176 responses, and we invited 77 respondents to participate in the study.

We intentionally did not recruit IPV victims or other stalking victims, to avoid re-traumatization by making them revisit memories of abuse [17, 30, 56]. Prior work has suggested taking a participant-centered approach when working with trauma victims and including mental health professionals who can provide services, if needed [29, 60]. However, our study did not involve services or interventions that would address the specific needs of victims.

We thus focused on the ability of members of a victim’s social support network to help detect and remediate security compromises, and recruited from the general population rather than self-identifying victims or survivors. We advertised seeking participants “interested in mobile phone security,” without screening for experience or expertise, to recruit participants who might have enough interest in mobile phone security to be willing to help a friend with iPhone security issues.

We limited our participants to iPhone users to simplify the study design, as iPhone user experience is relatively uniform compared to Android phones, which vary by manufacturer and Android version. The problems we investigated are consistent across recent iOS versions (Section 5).

3.2 Interview

We conducted remote, semi-structured interviews with 22 participants over Zoom from April through August 2021. We eliminated four interviews from our analysis, as we discovered during the interview that these participants did not meet our screening criteria. We recorded audio and video (of our shared iPhone screen) and transcribed interviews using Zoom and Otter.ai. We conducted two slightly different versions of the interview, each with nine participants (see Section 3.5 for more details). Our interview script is included in Appendix A. All of our study protocols were approved by our IRB.

We presented participants with four distinct exemplar scenarios that we selected based on a range of threats seen in prior work [13, 16, 18, 22, 23, 26, 28, 29, 33, 35, 47, 62], news articles [46, 55], and through one author’s experience as a technologist in a techclinic for IPV survivors. To simulate the scenarios, we used Zoom’s screen sharing feature to share the live screen of an iPhone (iOS versions 14.5-14.6) that had been reset and set up for this study.

To understand participants’ existing knowledge, we started three scenarios by asking them to define a mobile phone security concept related to the scenario (spyware, account compromise, and jailbreaking). In each scenario, we presented the iPhone screen remotely via Zoom, read the scenarios aloud to participants, and asked them to give us directions to interact with the iPhone and guide us through their strategies to help a hypothetical friend or coworker investigate their suspicions about stalking, account compromise, and surveillance. At the end of each scenario we asked participants how easy or difficult they had found the tasks, whether they would advise their friend to do anything else, and what they would do if they encountered that scenario in their own life.

We presented scenarios in an open-ended way, without specific instructions for how to approach the problem. However, when participants said they did not know what to do or lingered on irrelevant options, we gave them a hint for how to proceed, having developed a set of hints per scenario as part of our interview script, to help participants complete the task. This prevented us from being limited to observing only early obstacles that might otherwise prevent a participant from completing the task. Participants were not prohibited from using external resources: some directed us to do an internet search, and a few did internet searches on their personal devices.

Scenario 1. The first scenario was designed to explore participants’ strategies for determining whether someone had access to the iPhone’s location information, how easily they discovered the Location Sharing feature within Google Maps, and how easy they found it to disable this feature. We asked participants to imagine that their coworker asked for help confirming whether or not someone was tracking their (the coworker’s) location through their iPhone. Our iPhone’s location was being shared with “Mallory” via Google Maps’ Location Sharing feature, which can grant access to location information until revoked or for a certain length of time.

If participants attempted to resolve the problem by turning off iOS Location Services, we clarified that the coworker needed it on for navigation purposes and that we wanted to determine whether location was being shared with someone else. We gave participants hints to help them see that the Google Maps app was using their location, as shown in Figure 1.

Scenario 2. In the second scenario, we explored how participants investigated their friend’s suspicion that their intimate partner was remotely spying on their phone, whether they could detect that an app could be used to remotely access the device, and whether they could remove the app.

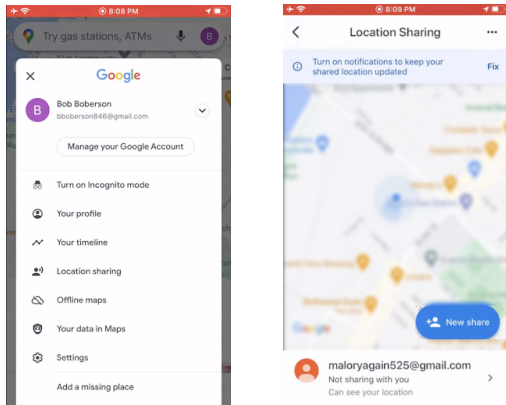


Figure 1: In Google Maps, selecting the top right circle revealed a drop-down menu with a “Location sharing” option (left). Selecting “Location sharing” revealed another account with access to the device’s location (right).

We first asked participants to describe what they thought spyware was. We then told them that their friend suspected that their significant other had “hacked” their phone and asked them to help their friend find out whether their significant other was remotely accessing their device using spyware.

On the test device, we installed an app that enables remote access and is marketed as a technical support tool, TeamViewer. Though iOS provides ongoing alerts while TeamViewer remotely accesses an iPhone, we envisioned a threat model in which the friend’s significant other has physical access and the passcode to the friend’s phone, and finds convenient times to remote into the friend’s phone to spy on them or to change settings that might enable spying, without needing to physically access the phone for long periods of time. Our intention was to observe how participants identified which app could remotely access the device, not whether remote access would be detectable while it was taking place.

After we noted their initial search attempts, we gave participants a hint that the suspected “spyware” was an app from the App Store and, if needed, another hint to search among the apps on the phone to find which app could be used as spyware. In the last nine interviews we added a hint to clarify that we were looking for an app that enables remote access to the device, not apps that simply appear suspicious. If participants were not familiar with the app or aware of its capabilities (as none were), we revealed that TeamViewer is an app that can be misused to enable spying, and that the friend’s significant other used it to remotely access the device without their friend’s permission. We then asked participants what precautions or advice they would suggest that their friend keep in mind, to see if they would consider escalation of abuse to be a possibility (see Section 2.1). We then told them that their friend decided it was safe to remove the spyware and asked them to guide us in removing the app.

We used the term “spyware” to describe what the friend sus-

pected was happening, i.e., spying. While we explained to participants that the TeamViewer app could be legitimately used for remote access or assistance, we continued to use the term “spyware” to capture the app use-case in the scenario’s context.

Scenario 3. In the third scenario, we explored whether participants could recognize indicators of iCloud account compromise and remove an unknown device’s access to an iCloud account. We asked them to describe what they thought account compromise was, then told them that their friend’s photos (and messages, in the second half of our interviews) were appearing and disappearing and asked them to help investigate. In the first nine interviews, we only mentioned photos, not messages (see Section 3.5).

We had logged into an iCloud account with two different devices, the test device and another device, and synced iCloud apps (Photos and iMessage) between them. We wanted to understand whether participants would intuit that changes in photos and messages could be a sign of iCloud compromise and discover an unknown device logged into the iCloud account, as well as how easy it was for them to remove an app from the list of devices logged into iCloud (see Figure 2).

Scenario 4. In this scenario, we investigated how easy it was for participants to follow online advice to detect whether a device is “jailbroken.” First, we asked participants to define “rooted” and “jailbroken.” We then asked them to imagine that their friend suspected their iPhone was jailbroken and wanted help following online advice for detecting jailbreaking.

We included a jailbreaking scenario because it enables downloading and hiding spyware banned by the App Store (see Section 2.1). As we wanted to study options available to general users, we searched online for articles with advice on how to detect jailbreaking and stalkerware. We chose articles by the FTC [8, 15] and Avast [12] because they might be recognizable, trusted and shared, and had simple instructions. In the first nine interviews, we presented participants with an article that suggests using a root checker app [15], and in the next nine, with an article that recommends checking whether the Cydia app (“Cydia”), a common app store for jailbroken iPhones, is installed on the phone [12]. Both articles suggest resetting the phone, which we asked participants to do in the last nine interviews. See Appendix B for the advice text.

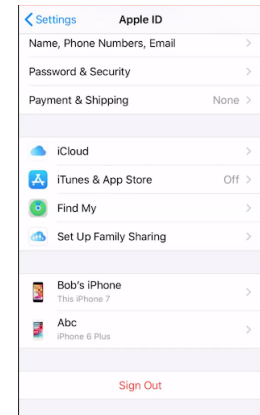


Figure 2: In iOS Settings, selecting the Apple ID and scrolling down revealed a list of devices logged into the iCloud account.

3.3 Data Analysis

We conducted a qualitative thematic analysis of the interviews by coding the interview transcripts as a group. Two of us coded all interviews, initially joined by a third researcher. Any initial disagreements were resolved through discussion. Since we conducted the coding collaboratively, it was not appropriate to calculate inter-rater reliability. We gathered keywords and ideas from the interviews and found common themes. Since the interview was divided into four scenarios, we coded in sessions dedicated to each scenario and developed a code book for each scenario. We reused several codes across scenarios, but there were also codes unique to each one. Our code book can be found in Appendix D.

To get a better idea of what was challenging or intuitive, we also analyzed the number of hints participants required for each scenario and their explanations about what they found easy or difficult about detecting and resolving the problems.

3.4 Limitations

As our study consisted of simulated scenarios, its design did not always reflect a completely realistic or typical situation. Additionally, it has some limitations due to the qualitative and remote nature of the study. While our study is limited to iOS and the apps studied, characteristics underpinning our findings are shared by OSes and apps more generally.

OS and App Selection. Many aspects of our study are applicable beyond iOS and Google Maps. Various navigation apps across OSes, including Google Maps (Scenario 1), Find My, and Waze, do not provide persistent notifications when the device transmits location to another user (see Section 5.1.1). Scenario 2 might be similarly difficult for Android users, as there are no obvious indicators on Android to inform users about apps' spying capabilities. While Scenario 3 is specific to iCloud, other OSes (and apps) offer cloud data syncing across multiple devices (e.g. Google). Popular instructions for detecting jailbreaking and (Android) rooting (Scenario 4) contain similar jargon (see Section 2), though future work could consider the usability of Android root detection apps.

Simulating Detection of Spying Capabilities. We encountered some challenges in designing a scenario to detect a legitimate app that could be misused for spying. In the initial nine interviews, our test phone had only three non-default apps installed: Zoom, Google Maps, and TeamViewer (see the left side of Figure 3). Though a few participants noted their unfamiliarity with TeamViewer, that does not mean they determined it could remotely access the device. See Section 3.5 for our modifications to this scenario. Additionally, we did not include hints to look at app permissions. TeamViewer and most other apps had not been set up or tested, which meant that additional permissions that might suggest remote access capability, such as Screen Recording or Accessibility permissions, had not been granted. While only two participants



Figure 3: Home screen in version 1 (left) and version 2 (right) of the study. The difference appears to have influenced whether participants noticed the remote access app.

looked at permissions for TeamViewer (which may have been easier to spot for the first nine participants, given that only three non-default apps were installed on the phone), we could have designed the scenario such that participants could look at app permissions to find that Screen Recording was enabled.

A Remote and Unfamiliar Test Device. In Scenario 3, three participants could not find the list of devices logged into the iCloud account because they did not scroll down. In Scenario 1, one participant said they did not see the “Stop” button (to stop sharing location) at the bottom of the screen, though our screen recording captured it. If the participants had been holding the test device, they may have intuitively scrolled down or been better able to look at the entire screen.

Additionally, if participants had really been helping their friend, they might be more familiar with the names and accounts logged into Google Maps or iCloud (Scenarios 1 and 3) and more suspicious of a stalker's email account or device.

Ordering Effects. Scenario order remained the same across interviews. While there may be ordering effects, each scenario required different skills. We observed participants routinely facing challenges in each subsequent scenario regardless of anything they may have learned from a previous one.

3.5 Study Modifications

We changed Scenarios 2 and 3 for the last nine interviews (P10-P18) to probe the impact of increasing the number of non-Apple apps installed on the phone and the number of apps involved in suspicious iCloud behavior, respectively. We wanted to see whether participants would continue to mention TeamViewer among 32 additional and potentially unfamiliar apps (see Figure 3) in Scenario 2 (they did not—a valuable contrast). We also wanted to see if participants would more easily detect iCloud compromise in Scenario 3, i.e., more intuitively link the suspicious behavior to the iCloud account, if we noted that two apps instead of one, photos *and* iMessage, were appearing and disappearing. In the first version, we only

mentioned photos, and seven of nine participants focused primarily on the Photos app and settings. Yet, responses were roughly the same, with five more participants focusing on photos. We discuss findings in Section 4.2 and Section 4.3.

To gain insight into subjective perspectives, we added questions about what participants would do if they experienced the same technical issue and, instead of asking them to rate difficulty on a scale, we made our questions more open-ended, asking how easy or difficult they found the scenario and why.

4 Results

We present findings surfaced by qualitative analysis of interview transcripts for each of the four scenarios. Our reports on frequency of behaviors are useful for understanding our participants but are not generalizable to a larger population.

4.1 Location Tracking

In this scenario, participants were asked to help their coworker find out whether someone (Mallory) was tracking their coworker's phone. Participants had difficulty associating a navigation app, Google Maps, with location sharing controls. While all participants appeared to understand that iOS Location Services were enabled and that Google Maps was always using Location Services, none of them suggested opening the app to investigate or disable location sharing. To stop location sharing, most participants attempted to fully turn off Location Services, which would prevent the co-worker from using navigation apps. With the hint to open the Google Maps app, most participants were able to confirm that location was being shared with Mallory and quickly stopped location sharing, though a few participants had difficulty navigating the UI. Most participants found this scenario to be difficult but suggested that it would have been easy, had they known to explore the app.

4.1.1 Detecting and Stopping Location Sharing

No participants discovered on their own that location was being shared with Mallory through Google Maps' Location Sharing feature. After we provided hints, 15 out of 16 participants¹ (all but P4) eventually discovered this.

Initial Strategies. Eight participants explored other iOS features, such as Tracking, Accessibility, and Control Center, which did not provide location-related information, and two participants used the iOS search bar to search for "location."

Participants' strategies in navigating the iOS Location Services UI varied. Eight participants looked at the Share My Location settings. Three participants, upon seeing that location was not being shared through this setting, concluded that location was therefore not being shared at all.

¹We are using data for 16 participants for Scenario 1 strategies, because the interviewer did not properly follow the interview script for P1 and P5.

Basically, you just see if Share My Location is on or off. Clearly it's off, so I guess that I would assume that someone is not tracking your location. (P17)

Ten participants suggested turning off iOS Location Services. This could be impractical if the user depends on using Location Services, e.g., for navigation. Three participants suggested changing Location Services settings for Google Maps, for example from "Always" to "Never," but these solutions do not permanently resolve the issue, as location could be transmitted upon re-enabling Location Services. In addition, three participants suggested turning off Find My iPhone, which would not work, since it was not the app transmitting data.

Hints. Our hints were intended to guide participants to look at the iOS Location Services settings and notice that Google Maps was the only app always using location, which we hoped would inspire them to open the app and investigate its Location Sharing feature. Three of 16 participants required Hint 1, to go into iOS Settings and search for something related to location. Six (including the prior three) required Hint 2, to search in iOS Privacy Settings and Location Services, to see that Google Maps was using location services. These participants had initially been looking at other iOS settings.

All 16 participants required Hint 3, to open Google Maps and check its settings. Seven participants required only this hint. Although three participants noted that Google Maps was set to "Always" use location, no one suggested exploring within-app settings. Four participants remarked on the difficulty, considering the process to involve too many steps:

I honestly never do this. It was just too much jumping around and knowing the difference between when to look into settings on the phone, versus when to look on settings in the specific app. (P12)

After opening Google Maps, six participants required Hint 4 to select the Location Sharing option in the drop-down menu.

Checking Google Maps Settings. Once participants received the hint to open Google Maps (which no participant suggested), nine were able to select the top-right circle and select Location Sharing settings from the resulting drop-down menu (Figure 1 left). Eight of these nine understood that the subsequent screen (Figure 1 right) showed that location was currently being shared with another user, Mallory. However, P7 and P15 did not understand the screen showing Mallory's email. P7 asked whether we were trying to communicate with Mallory, and P15 assumed Mallory's email was the coworker's email. P10 asked us to confirm whether Mallory's email could be trusted. In Section 3.4 we discuss the limitation that participants might be more likely to recognize their own or their coworker's details in a realistic scenario.

Seven participants had some difficulty navigating the Google Maps UI. P12 and P17 directed us to select the blue arrow icon, which only changes the angle of the map's view. P16 told us to look for a "blue thing" next to the iOS time,

perhaps referencing an iOS icon that appears when location is being used. P6 selected “Updates” and P9 selected “Contribute” at the bottom of the initial Google Maps screen. P13 said they had not known location could be shared via Google Maps and suggested looking at location search history.

Once participants saw that the location was being shared with Mallory, nine of them suggested that the process would have been easy if they had known to search in Google Maps.

You wouldn't really expect to use Google Maps. When it comes to someone sharing your location, you'd assume that it's one of the apps that's pre-installed on the iPhone. That was confusing, because I didn't think that I'd need to go there in the first place. But then, once I did know it's Google Maps and went to settings ... I can just find the thing that says something to do with sharing a location. (P17)

To go in the app itself, not just the iPhone, but the app settings, that's tricky in itself, so I had a little bit of issue to find that, but I mean, it was all there. (P14)

Strategies to Stop Location Sharing. After participants observed that location was being transmitted to Mallory, we asked them to guide us to stop location sharing. Eleven participants clicked the arrow next to Mallory’s email on the Location Sharing screen, which led to a screen with a “Stop” button, and then selected “Stop.” Three participants effectively stopped sharing through an alternate route. P10 and P15 selected the three dots next to Mallory’s email and went to the Google Account’s Location Sharing settings via Safari to stop sharing, and P7 blocked Mallory. Two participants chose sufficient but temporary solutions. P4 turned on incognito mode in Google Maps. P18 went back to iOS Location Services settings and changed Google Maps location settings from “Always” to “Never.” P8 mistook a nudge for a viable option, suggesting we click “Fix”(see the top of Figure 1).

P4 and P7 tried to add a “New Share” in Google Maps, rather than remove an existing share. P7 said that they could not see the “Stop” button at the bottom of the screen when they first saw the screen showing Mallory. P4 thought setting time to zero for a new share might stop the sharing.

Most participants (12 of 16) found stopping location sharing to be easy, with six noting that it was self-explanatory.

4.1.2 Security Precautions

When we asked participants what they would do if they thought someone had access to their location, some discussed additional security precautions they would take. P14 mentioned blocking and, later, consulting the police:

I would definitely block the person from my phone, make sure ... on social media, to get rid of that

person, because depending on if it's Facebook or anything [where] you can see the other person's location, you may not know their location is on. (P14)

P16 raised the possibility of escalation, which was the only mention of escalation in our entire study:

Based on my level of paranoia, if it seems like anything serious or fatal, some type of ongoing thing, like, let's say I removed that email and a new email popped up later on, I would try to probably download a VPN on my phone like something to just another layer of security, I guess. (P16)

Unfortunately, a VPN does not necessarily mask a location that is being shared through a navigation app, so this would not be an effective strategy.

4.2 Spyware and Apps That Can Spy

In this scenario, we asked participants to identify which app could be misused as spyware to remotely access their friend’s device. No participants successfully identified the remote access app, and most said they found the task of identifying it to be difficult. To find the app, most participants went into iOS settings and searched for an app with a suspicious name or with keywords such as “spy” in the name. We also asked what precautions or advice they would give to their friend after we revealed that TeamViewer was being used to spy on the friend. Most participants suggested deleting the app, and almost half suggested options that stalking victims may find difficult, such as ending the relationship or not allowing the significant other to access the phone. No participant mentioned the risk of escalating an abusive situation, which could threaten the friend’s safety. Most participants said that deleting the app was easy, as the process is the same for all iOS apps. Six participants suggested doing more than deleting the app, such as deleting the account used or deleting the app from the other device. We discuss the results in more detail below.

4.2.1 Definitions of Spyware

Before prompting participants with Scenario 2, we asked them what they thought spyware was. Eleven of eighteen participants described spyware as a virus or malicious file that could discover information about them or spy on them. Four participants did not mention viruses or malware but described spyware as tracking or collecting information about them. Five participants thought spyware had to do with tracking web browsing. Three participants had the misconception that spyware was a tool that could protect them, claiming that it can act as an antivirus (P7), “help prevent your computer from coming into contact with threatening sites” (P9) and “protect you from people trying to hack into your phone” (P16). We did not correct them, but they appeared to realize their misconception after we prompted them with the scenario.

4.2.2 (Not) Identifying the Remote Access App

No participants identified the app, TeamViewer, as an app capable of remotely accessing the friend's device. For the first half of our participants, who saw only three non-default apps installed on the phone (Zoom, Google Maps and Team Viewer), four of nine participants considered TeamViewer to be suspicious because they were unfamiliar with it and suggested deleting it. When 35 non-default apps were installed on the phone, no participants pointed out the TeamViewer app but most participants suggested deleting a hacking-simulation game app called "HackIt."

Initial Strategies. When we asked participants what instructions they would give their friend to see whether there was spyware on their phone, 13 participants suggested going into iOS Settings. Within iOS Settings, four participants visited Control Center, six participants went into Privacy, and two participants visited Accessibility. P14 looked through the iOS home screen. Four participants did not immediately engage with the iOS UI, suggesting other strategies, such as using a search engine to query "How do I find out if there is spyware on my phone?" (P3), installing antivirus software (P8), contacting Apple (P13), or asking their friend if they had "opened any weird emails or texts" (P18).

Hints. All participants except P14 required Hint 1, that the suspected "spyware" is an app downloaded from the App Store. Three participants pointed out that spyware can be hidden. P11 expressed doubt about being able to detect spyware. We describe the the initial attempts to find spyware below.

Three participants required Hint 2 (swipe through the home screen to reach the iOS app library) to search for apps on the phone. All nine participants in version 2 of the study required Hint 3 (look for an app that enables remote access to the device), though it did not appear to help them, since no one was able to figure out which app could have remote access.

Strategies to Find Spyware. Once participants knew that the spyware was an app from the App Store, they took different approaches to detect it. Six participants chose to look at apps by scrolling down on the iOS Settings screen, seven looked at apps on the home screen, two went to the App Store, and one used the iOS search bar. Two participants were not sure where to look and were given Hint 2.

Out of six participants who reviewed apps at the bottom of the iOS Settings, only P5 and P9 looked at app permissions:

There's three apps on this iPhone, I would probably go through every single app, and see if there's a certain setting that causes a red flag... (P5)

As we did not grant TeamViewer extra iOS permissions (see Section 3.4), P5 concluded that "there [was] nothing suspicious." P9 considered background app refresh suspicious.

We asked participants what they were looking for. Six participants said they were looking for keywords like "spyware" or app names they found weird or suspicious:

I would say apps with weird foreign names like Russian letters, Chinese letters or something. (P8)

Other participants suggested that connections across multiple apps (P6), power consumption (P6), and location sharing (P5) could be indicators of spyware. Some participants did not know what they were looking for:

I didn't really know what I was looking for. TeamViewer doesn't seem very malicious, but if it was my phone I would recognize that there is an app that I didn't download, you know, so yeah, it'd be different if it were my phone. (P11)

Recognizing and Understanding Apps. Fourteen participants found identifying the app capable of remote access or spying to be difficult. No participants were familiar with the app, so they were likely not aware of its ability to allow others to remotely control devices. P18 said that they had not known that remote access was possible. Three participants noted that solving the problem might involve being able to recognize unfamiliar apps or having their friend indicate which apps they might not recognize or remember downloading. After we showed P16 the App Library, they asked:

How would you know? Does the friend know what apps they already had and what they didn't? (P16)

This highlights the difficulty of finding such an app on behalf of someone else, a challenge faced by advocates, who may not know what apps the survivor installed or not:

It's weird looking at someone else's phone, it's like another world. (P10)

Participants also found it difficult to understand or learn the full capabilities of downloaded apps, including whether an app could be used as spyware.

It's kind of hard to tell. Sometimes you don't really know if someone has access to an app and is able to access your phone. (P8)

4.2.3 Removing the App and Deleting Data

All 18 participants found deleting the app to be easy, with seven saying that this was because they have done it before. Some participants thought that deleting the app on the phone would delete the app's data.

I believe, if we just like, completely delete the app, it should delete everything associated, all the data associated with that app. (P13)

Six participants, who were concerned about account data, suggested taking more steps after deleting the app, including erasing app data.

The only thing you would have to worry about is after you delete it, you know, make sure to erase your information, so you still don't have an active account with them. (P14)

Some participants' suggestions may not be feasible in IPV contexts, such as "delete the other half" of the spyware app on the significant other's phone (P2).

4.2.4 Advice to Friend Experiencing Tech Abuse

This scenario was the only scenario we situated within the context of IPV by asking the participant to imagine that their friend was being monitored by their significant other. Before asking participants to help their friend delete the app, we asked participants what precautions or advice they would suggest that their friend keep in mind.

Twelve participants said they would advise their friend to delete the app and ten suggested not letting other people handle or download things onto their device. Five participants suggested confronting the significant other, to figure out their intentions. Two of these suggested leaving the relationship:

Well, she didn't download it. Her loser boyfriend did. Get rid of the boyfriend. (P1)

I would definitely encourage them to leave the relationship. (P3)

4.3 iCloud Account Compromise

In this scenario, we asked participants to help their friend find out why some photos and messages were disappearing and new ones were appearing. With hints, we led them to discover an unknown device logged into the friend's iCloud account.

While most participants were familiar with the concept of iCloud and the iCloud account UI, they had difficulty discovering the list of devices on the Apple ID UI, even when they knew (or were given the hint) to search iCloud settings. Many participants looked at the iOS Photos settings, searching for things like Shared Albums. However, iOS Photos and Messages settings do not offer indications of other devices or device activity, so we had to nudge several participants towards the iCloud settings in iOS Settings. After seeing that an unknown device was logged into the iCloud account, most participants found it easy to remove the device. More than half of them suggested enhancing authentication mechanisms by changing the password or enabling multi-factor authentication.

4.3.1 Definitions of Account Compromise

Before starting the scenario, we asked participants what they thought account compromise was. Fifteen participants associated account compromise with another person (not the

account holder) gaining access to the account or data in it. Six participants mentioned password compromise, three participants mentioned data leaks, and one mentioned a different device being used to access the account.

4.3.2 Finding Devices Logged into the iCloud Account

Only P11 and P13 were able to quickly find the unknown device in the device list on the Apple ID UI. With hints, 13 of 18 participants were able to eventually find the device. While it appeared to be intuitive for most participants to check settings related to the apps showing suspicious behavior, i.e., the Photos or Messages settings or iCloud Photos settings, there is no indication in these settings that another device is logged into the iCloud account and syncing with the apps.

Initial Strategies. Four participants appeared to initially connect changes in the Photos or iMessage apps to iCloud syncing with another device and immediately checked iCloud settings. However, two still required a hint to find the list of devices logged into the iCloud account.

Twelve participants focused on the Photos app and checked the Photos app, Photos app settings in iOS Settings, or iCloud Photos settings. Of the six of these who checked Photos app settings from the iOS Settings, three checked whether the Shared Albums feature was enabled and tried to turn it off.

My thought process for Shared Albums would be, if it's sharing it with other people, there would be an option to see who, like if there's a drop down to see who else can see the photos. (P12)

Five participants opened the Photos app, and two of those five checked the Albums UI for any "Hidden" items. P6 checked whether the iCloud Photos feature was enabled.

Participants also checked other settings before reaching the iCloud settings. Three out of nine participants looked at Messages settings in iOS settings and asked whether the email address appearing next to "Send & Receive" was supposed to be logged in. We confirmed that this was the iCloud account of the device owner in the scenario. P17 thought the problem was caused by another third-party app.

Hints. We provided hints to help participants understand that syncing with other devices might be occurring (Hint 1) and that this syncing was occurring through iCloud (Hint 2), and to guide them to the iCloud settings in iOS Settings (Hint 3), which might lead them to the Apple ID UI's list of devices.

Only P11 and P13 did not require a hint, though we had to redirect P13 after they guided us to change the iCloud password, which can stop syncing the iCloud with other devices but does not help us discover another device. After this redirection, P13 immediately located the device list.

Eleven participants required Hint 1, that their friend used to sync their photos with other devices. Seven of those required Hint 2, which noted that the changes were happening due to

iCloud account syncing. Three of those 11 required Hint 3, which led them to the Apple ID UI. P5 only needed Hint 1.

Five participants who did not require Hint 1, about device syncing, still required at least one of the other hints about iCloud. One participant required all three hints.

Navigating the iCloud Settings. To see the list of devices logged into the device owner’s iCloud account, participants had to select the Apple ID (iCloud account) from the iOS Settings and scroll down. Half of them had difficulty finding this device list. Nine participants went into other options (including “Password & Security,” “Name, Phone Numbers, Email,” “Family Sharing,” and “iCloud,”) in the iCloud settings to find the device list. Three participants mentioned that they were familiar with iOS and iCloud but that finding the device list in the iCloud settings was not immediately apparent:

I think I knew generally to look under iCloud, but I just didn’t know the full screen. (P3)

P7 and P9 reached the screen listing logged-in devices but did not register its significance:

I didn’t even know that that was another phone that was interfering or connecting into. (P7)

4.3.3 Removing the Device from the iCloud Account

After finding the list of devices, 13 participants directed us to select the unknown device and select “Remove from Account” in the resulting screen showing the device information. Six participants noted how intuitive it was to identify the removal option, with two mentioning its red color. Five participants did not know how to proceed after finding the list of devices.

4.3.4 Security Precautions and Advice

Participants were asked what further advice they might give their friend or what they would do in the same scenario. Nine participants suggested changing the iCloud password and two of them also suggested using multi-factor authentication.

4.4 Online Advice to Detect Jailbreaking

In this scenario, we asked participants to help their friend follow online advice to find out whether an iPhone is jailbroken. After reading the online advice we showed them, most participants understood the instructions, but implementing the advice was not always easy. In our first nine interviews, we asked participants to follow online advice to find a “root checker app” (see Appendix B.1), and we found that no participants were able to successfully find such an app after searching in the App Store and on the web. In the next nine interviews, we asked participants to follow different online advice (see Appendix B.2), to search for an app called Cydia and “restore factory settings.” Most participants found it easy to search for Cydia, but to “restore factory settings,” most participants chose the wrong option.

4.4.1 Definitions of Rooted or Jailbroken

We started the fourth scenario by asking participants what they thought “rooted” or “jailbroken” meant. Since “jailbroken” is more commonly used in the context of iPhones than the term “rooted,” it was unsurprising that 13 of 18 participants (all iPhone users), were familiar with the term “jailbroken,” while only one participant was familiar with the term “rooted.” Five participants described jailbreaking as beneficial, for customizing devices or downloading paid apps for free, and two participants suggested it meant the device was stolen or hacked. However, some participants, including ones who indicated they were familiar with the term, expressed difficulty understanding the concept of jailbreaking.

4.4.2 Searching for a Root Checker App

In our first nine interviews, we asked participants to follow the FTC’s online advice to find a “root checker app” [8, 15]. We found that no participants were able to successfully find such an app. Six participants critiqued the article for not including any example apps, lacking details, and not being helpful.

To find a root checker app, seven participants searched in the App Store using the following terms: “root checker,” “root checking,” “jailbreaker,” “jailbreak checker,” “root checker app,” “rooted,” “root,” “security check,” “stalk,” and “stalker.” The most prominent app results for searches containing “root checker” and “jailbreak” were game apps. Four participants used web searches to look for a root checker app, and three of these said they would download apps mentioned in search results. However, these apps were either no longer available on the App Store or not able to detect jailbroken status. Two participants suggested that tutorial videos they discovered in their web searches would lead to a root checker app.

4.4.3 Finding Cydia and Resetting the Phone

In the last nine interviews of our study, participants followed Avast’s two-step online advice to: 1) check for the Cydia app, and 2) restore factory settings.

Finding Cydia. Eight participants used the iOS search bar to search for Cydia, an alternative app store for jailbroken iPhones. When Cydia did not appear in search results, five participants concluded that the app was not installed on the phone, but three participants were unsure whether it was installed or not. P14 and P17 appeared to think they were searching for Cydia in order to use it and suggested we download Cydia.

Restoring Factory Settings. To follow the advice’s second instruction, to “restore factory settings,” six of nine participants selected the “Reset All Settings” option, which does not delete apps or data, rather than “Erase All Content and Settings,” which does, from the iOS Reset menu. The article used the phrase “restore to factory settings,” but there is no menu or option using the word “factory setting” or “factory reset.” Five participants suggested that the wording of the

advice as well as of the iOS reset menus made implementing the instruction more difficult.

I would expect it to say factory reset and not have three different options that look very similar. (P17)

5 Discussion

In this section, we discuss usability challenges in Google Maps and iOS, recommendations to mitigate these challenges, the importance of effective online advice, and how our findings underscore the existing need for usable security options to detect and counter stalking and technology-enabled abuse.

Most of our recommendations focus on preventing easily exploitable threats (e.g., location sharing), since abusers often resort to unsophisticated attacks [22, 33, 35, 62]; one recommendation, to improve Reset options, targets sophisticated attacks (jailbreaking). Implementing better status indicators or persistent notifications of transmission of data to other users would likely have the most impact on users' ability to identify and remediate threats, since adversaries would be less able to leverage common or native apps. Such recommendations are in line with Jakob Nielsen's heuristic, "visibility of system status" [27], which emphasizes communicating current status "to keep users informed about what is going on, through appropriate feedback within reasonable time" and building trust by ensuring that "no action with consequences to users should be taken without informing them."

5.1 Usability Challenges

Our findings highlight usable security problems in Google Maps and iOS settings. While participants were relatively familiar with security risks, resolving security problems proved to be difficult or unintuitive. Despite iOS and app updates since we conducted our interviews, the features we explored (Google Maps' Location Sharing, iOS Apple ID/iCloud device list, and iOS Reset and Location Services menus) have remained essentially the same from April 2021 through June 2022 (iOS 14.5 through iOS 15.5).

5.1.1 Google Maps Usability Issues

In a threat model that assumes physical or remote access to a phone, an abuser can enable surveillance by misusing legitimate apps, such as navigation apps like Google Maps. Indeed, while Google bans stalkerware, which it defines as "[c]ode that collects and/or transmits personal or sensitive user data from a device without adequate notice or consent and doesn't display a persistent notification that this is happening," such navigation apps (Google Maps, Waze, Find My) do not provide persistent notifications that location is being transmitted to another user [9]. All participants had difficulty detecting another person's real-time access to the device's location, as

they did not check settings in Google Maps. Seven of them needed help locating the Location Sharing feature within Google Maps. Below, we make recommendations for how security, notifications, and transparency could be improved to alert users that their real-time location is being shared.

Authentication. Google Maps does not require users to re-authenticate to share their real-time location with another user by adding a New Share. This enables anyone with access to the phone to begin sharing with another user. The security of the device owner could be improved by requiring authentication to enable location sharing with a New Share.

Notifications. When a Google Maps user begins sharing their location with someone, two email notifications are immediately sent, one to the user and one to the contact with whom they're sharing, and a periodic email notification is sent to the user. While these are helpful, users who do not check their email often or at all, or whose notifications may have been deleted, could benefit from persistent notifications or periodic ones in different forms, e.g., SMS or in-app notifications.

Indicators. Upon opening the Google Maps app, there are no indicators that location is currently being shared. Users have to take the initiative to check the "Location Sharing" settings. To improve transparency, an indicator could alert users to the fact that they are currently sharing their location with someone. Some participants recommended such an "immediate indicator" (P6), e.g., a "glowing button" (P8).

While system status notifiers, persistent or periodic notifications, and security suggestions may be inconvenient for some users, if acted upon, they would reduce some range of opportunities for malicious parties to exploit apps for spying.

5.1.2 iOS UI Usability Issues

We identified some opportunities to improve the iOS UI's usability for people who are concerned about stalking.

Apps Using Location Services. iOS takes steps to protect users who use Location Services, by providing indicators during usage as well as periodic notifications about background location use [3]. However, none of the participants found it intuitive to investigate within-app settings in Google Maps (Scenario 1). iOS could further help users by informing them that apps using Location Services may be able to share the location with other people (even when only in limited modes, such as "While Using the App") and by recommending that users periodically check location settings within apps. While it would be impractical to catalog how to investigate settings in all apps, at least informing users of the possibility of location sharing could improve user awareness.

No Indicators of Devices in iOS App Settings. In Scenario 3, participants missed critical information about an unknown device because the Photos and iMessage app settings' UIs did not indicate that there was another device accessing the app data. Without our hint(s), many participants could not figure out that an unknown device was making changes in

the Photos and iMessage apps, and 16 participants required some hints to find the list of devices logged into the iCloud account. We recommend adding indicators regarding device access and activity within app settings.

Additionally, the list of logged-in devices and the “Remove Device” feature were placed at the bottom of their respective UI screens, which requires users to scroll down. Such critical information would ideally be more immediately visible.

Multiple Reset Options. In Scenario 4, participants struggled to choose between various reset options. There are six reset options in the reset menus of iOS 14 and iOS 15. Only upon selecting an option is a user given more information. We suggest providing clearer information to users about the differences between the reset options, especially regarding the difference between “Reset All Settings” and “Erase All Content and Settings.” Given that the latter option could potentially undo the changes made to the phone’s operating system by jailbreaking, while the former would not, highlighting the differences in an accessible way would make a considerable difference in a safety-critical situation.

Changes in iOS 15 and 16. Though iOS 15, released in September 2021, still has the issues we identified (e.g., multiple reset options), the “Record App Activity” feature in iOS Privacy Settings, which allows users to save a 7-day summary of when apps access their data, may help users identify apps using location or camera data and become more aware of app capabilities and activity. Apple announced in June 2022 that iOS 16 will include a tool called Safety Check, designed to help IPV victims revoke an abuser’s access to location and data [10]. This seems likely to address unknown or unwanted iCloud logins and privacy permissions, but it is unclear if it could revoke non-Apple apps’ within-app permissions. We encourage researchers and advocates to investigate whether these are usable security tools for victims of IPV and stalking.

5.1.3 Effective Communication and Advice

Participants had difficulty implementing advice to find a “root checker app” or “restore factory settings,” due to a lack of clear explanations and implementable instructions to end users about security.

Online Advice. All participants struggled to implement the FTC’s advice to find a root checker app. As there do not appear to be apps in the Apple App Store that are marketed as “root checker apps,” and Apple does not support a jailbreak detection feature for app developers [1], it does not seem practical to recommend that iPhone users seek out such apps. Additionally, it may be helpful to clarify which operating systems “rooting” and “jailbreaking” are associated with.

Though participants found the Avast article’s instructions relatively easy to implement, several encountered difficulty following the instruction to “restore factory settings.” We recommend that in online advice, instructions should match the language on the UI and should be updated to reflect changes

in the UI. In this case, it would help to note that the relevant option is “Erase All Content and Settings.”

Understanding Spyware and Its Many Forms. Additionally, given that three participants defined spyware as something beneficial that could protect them against online privacy or security threats, we discovered a potential issue with the term “spyware.” More research is needed on how to communicate effectively using computer security terms.

5.2 Including the Stalking Threat Model

Our study highlights the importance of including the stalking threat model in usable security design and research, i.e., focusing not only on use but also abuse of technology. Even though some of our participants suggested turning off certain settings as a solution, survivors should not have to give up using technology that may be essential to them, such as navigation apps [39]. As expected, most of our participants did not consider the risk of escalation, and some even gave advice to confront the abuser. The stalking threat model could be used to develop usable and intuitive UIs that help users safely detect and combat technology-enabled abuse and stalking.

Psychological Factors. While we did not interact with self-identifying victims of trauma, the confusion that our non-tech savvy participants expressed suggests that solving the problems we presented may be stressful.

After doing all these tasks, I just feel honestly a bit overwhelmed, but you know, good learning experience. . . . It’s just that I don’t know anything, and . . . it was a little like, a lot of booby traps. And yeah, it was just confusing, very, very confusing. (P12)

IPV and stalking victims experiencing trauma might also feel overwhelmed, likely more than our participants, as they try to detect surveillance. Usable tools and interfaces could make the process of detecting surveillance less difficult and confusing, and thereby perhaps cause less undue stress. While technology is a vector for abuse, it can also be a tool for survivors to enhance and maintain their safety.

6 Conclusion

This study focused on the qualitative analysis of 18 semi-structured interviews in which participants responded to four simulated-risk mobile phone security scenarios.

In four realistic scenarios simulating stalking and surveillance, the majority of non-tech savvy participants encountered significant usable security challenges, failing to use iOS and Google Maps UIs to detect and resolve security compromises. We recommend that companies make improvements to their interfaces and that writers of online security articles ensure their advice is clear and implementable. More research is needed on developing usable security tools and options to better detect and counter technology-enabled abuse.

Acknowledgments

This research was funded in part by the first author's GEM fellowship. We would also like to thank Sarah Pearman, Kevin Kim, and Chanaradee Leelamanthep for their work on an earlier version of this study.

References

- [1] Jailbroken detection check without App Store rejection. <https://developer.apple.com/forums/thread/66363?answerId=191199022#191199022>, Oct 2016.
- [2] Abuse of trust: How to identify and remove stalkerware. <https://nordvpn.com/blog/how-to-identify-stalkerware/>, Jul 2019.
- [3] About privacy and location services in iOS and iPadOS. <https://support.apple.com/en-us/HT203033>, Feb 2021.
- [4] Citizen Clinic - CLTC UC Berkeley Center for Long-Term Cybersecurity. <https://cltc.berkeley.edu/about-us/citizen-clinic/>, 2021.
- [5] Detecting and removing stalkerware. <https://goaskrose.com/stalkerware/>, 2021.
- [6] Safety net apps. <https://www.techsafety.org/safetynetapps>, 2021.
- [7] Stalkerware: Understanding and stopping technology-facilitated domestic violence. <https://staysafeonline.org/wp-content/uploads/2021/04/Stalkerware-Tip-Sheet-2021.pdf>, 2021.
- [8] Stalking apps: What to know. <https://www.consumer.ftc.gov/articles/stalking-apps-what-know>, May 2021.
- [9] Developer Program Policy - Play Console Help. https://support.google.com/googleplay/android-developer/answer/11987217?hl=en&ref_topic=9877065, May 2022.
- [10] Keynote - WWDC22 - Videos. <https://developer.apple.com/videos/play/wwdc2022/101/>, Jun 2022.
- [11] Laura Brignone and Jeffrey L. Edleson. The dating and domestic violence app rubric: synthesizing clinical best practices and digital health app standards for relationship violence prevention smartphone apps. *International Journal of Human-Computer Interaction*, 35(19), 2019.
- [12] Carly Burdova. What is jailbreaking and is it safe? <https://www.avast.com/c-jailbreaking#topic-6>, May 2021.
- [13] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The spyware used in intimate partner violence. In *2018 IEEE Symposium on Security and Privacy (SP)*, 2018.
- [14] Jacqueline Connor. Who's stalking: what to know about mobile spyware. <https://staysafeonline.org/blog/whos-stalking-know-mobile-spyware/>, Oct 2016.
- [15] Jacqueline Connor. Who's stalking: what to know about mobile spyware. <https://www.consumer.ftc.gov/blog/2016/09/whos-stalking-what-know-about-mobile-spyware>, Sep 2016.
- [16] Nicki Dell, Karen Levy, Damon McCoy, and Thomas Ristenpart. How domestic abusers use smartphones to spy on their partners. *Vox*, May 2018.
- [17] Melanie P. Duckworth and Victoria M. Follette. *Re-traumatization: Assessment, treatment, and prevention*. Routledge, 2012.
- [18] Brett Eterovic-Soric, Kim-Kwang Raymond Choo, Helen Ashman, and Sameera Mubarak. Stalking the stalkers – detecting and deterring stalking behaviours using technology: A review. *Computers & Security*, 70:278–289, 2017.
- [19] Blake Flournoy. How to check if an iPhone has been jailbroken. <https://www.techwalla.com/articles/how-to-check-if-an-iphone-has-been-jailbroken>, Dec 2018.
- [20] Lorenzo Franceschi-Bicchierai. Inside the “stalkerware” surveillance market, where ordinary people tap each other's phones. *WIRED*, Apr 2017.
- [21] Cynthia Fraser, Erica Olsen, Kaofeng Lee, Cindy Southworth, and Sarah Tucker. The new age of stalking: Technological implications for stalking. *Juvenile and Family Court Journal*, 61(4), 2010.
- [22] Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ristenpart, and Nicola Dell. “Is my phone hacked?” Analyzing clinical computer security interventions with survivors of intimate partner violence. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Nov 2019.

- [23] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. “A stalker’s paradise”: How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Apr 2018.
- [24] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), Dec 2017.
- [25] Jessica R. Goodkind, Tameka L. Gillum, Deborah I. Bybee, and Cris M. Sullivan. The impact of family and friends’ reactions on the well-being of women with abusive partners. *Violence Against Women*, 9(3):347–373, 2003.
- [26] Diarmaid Harkin and Adam Molnar. Operating-system design and its implications for victims of family violence: The comparative threat of smart phone spyware for android versus iPhone users. *Violence Against Women*, 27(6-7):851–875, 2021.
- [27] Aurora Harley. Visibility of system status. *Nielsen Norman Group*, Jun 2018.
- [28] Tirion Elizabeth Havard and Michelle Lefevre. Beyond the power and control wheel: How abusive men manipulate mobile phone technologies to facilitate coercive control. *Journal of Gender-Based Violence*, 4(2):223–239, Jun 2020.
- [29] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical computer security for victims of intimate partner violence. In *28th USENIX Security Symposium (USENIX Security 19)*, Aug 2019.
- [30] Tad Hirsch. Practicing without a license: Design research as psychotherapy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.
- [31] Alison Grace Johansen. Is jailbreaking legal and safe? <https://us.norton.com/internetsecurity-mobile-is-jailbreaking-legal-and-safe.html>, Mar 2019.
- [32] Cynthia Khoo, Kate Robertson, and Ronald Deibert. *Installing Fear: A Canadian Legal and Policy Analysis of Using, Developing, and Selling Smartphone Spyware and Stalkerware Applications*. Number 20 in Citizen Lab Research. Jun 2019.
- [33] Roxanne Leitão. Digital technologies and their role in intimate partner violence. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.
- [34] Roxanne Leitão. Technology-facilitated intimate partner abuse: a qualitative analysis of data from online domestic abuse forums. *Human-Computer Interaction*, 36(3), 2021.
- [35] Tara Matthews, Kathleen O’Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.
- [36] Nick Mooney. Jailbreak detector detector: An analysis of jailbreak detection methods and the tools used to evade them. <https://duo.com/blog/jailbreak-detector-detector>, Jan 2019.
- [37] James Nicholson, Lynne Coventry, and Pamela Briggs. “If it’s important it will be a headline”: Cybersecurity information seeking in older adults. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.
- [38] David Nield. How to check your devices for stalkerware. *WIRED*, Jul 2020.
- [39] Erica Olsen. Device and account security in safety planning for relocation with NortonLifeLock. <https://vimeo.com/631313869/ab089b7b24>, Oct 2021.
- [40] Cheul Young Park, Cori Faklaris, Siyan Zhao, Alex Scuito, Laura Dabbish, and Jason Hong. Share and share alike? An exploration of secure behaviors in romantic relationships. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, Aug 2018.
- [41] Christopher Parsons, Adam Molnar, Jakub Dalek, Jeffrey Knockel, Miles Kenyon, Bennett Haselton, Cynthia Khoo, and Ronald Deibert. The predator in your pocket: A multidisciplinary assessment of the stalkerware application industry. Citizen Lab Research Report No. 119. Jun 2019.
- [42] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

- [43] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *29th USENIX Security Symposium (USENIX Security 20)*, Aug 2020.
- [44] L. E. Rose, J. Campbell, and J. Kub. The role of social support and family relationships in women’s responses to battering. *Health Care for Women International*, 21(1):27–39, Feb 2000.
- [45] Kevin A. Roundy, Paula Barmaimon Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, and Acar Tamersoy. The many kinds of creepware used for interpersonal attacks. In *2020 IEEE Symposium on Security and Privacy (SP)*, 2020.
- [46] Aarti Shahani. Smartphones are used to stalk, control domestic abuse victims. *NPR*, Sep 2014.
- [47] Cynthia Southworth, Jerry Finn, Shawndell Dawson, Cynthia Fraser, and Sarah Tucker. Intimate partner violence, technology, and stalking. *Violence Against Women*, 13(8):842–856, Aug 2007.
- [48] Evan Stark. *Coercive Control: How Men Entrap Women in Personal Life*. Oxford University Press, Mar 2009.
- [49] Kateryna M. Sylaska and Katie M. Edwards. Disclosure of intimate partner violence to informal social support network members: A review of the literature. *Trauma, Violence & Abuse*, 15(1):3–21, Jan 2014.
- [50] Shashank Thakur. How to detect if an iOS device is jailbroken. <https://hackernoon.com/how-to-detect-if-an-ios-device-is-jailbroken-263u3tdj>, Oct 2020.
- [51] National Network to End Domestic Violence. App safety considerations for survivors of abuse. <https://www.techsafety.org/resources-survivors/app-safety-considerations>, 2014.
- [52] National Network to End Domestic Violence. Technology safety plan: A guide for survivors and advocates. <https://www.techsafety.org/resources-survivors/technology-safety-plan>, 2018.
- [53] Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. The tools and tactics used in intimate partner surveillance: An analysis of online infidelity forums. In *29th USENIX Security Symposium (USENIX Security 20)*, August 2020.
- [54] Sarah Turner, Jason Nurse, and Shujun Li. When googling it doesn’t work: The challenge of finding security advice for smart home devices. In *International Symposium on Human Aspects of Information Security and Assurance*, pages 115–126, 2021.
- [55] Jennifer Valentino-DeVries. Hundreds of apps can empower stalkers to track their victims. *The New York Times*, May 2018.
- [56] Jodie Valpied, Abigail Cini, Lorna O’Doherty, Ann Taket, and Kelsey Hegarty. “Sometimes cathartic. sometimes quite raw”: Benefit and harm in an intimate partner violence trial. *Aggression and Violent Behavior*, 19(6):673–685, 2014. Violence and Health: Current Perspectives from the World Health Organization (WHO) Violence Prevention Alliance.
- [57] Douglas M. Walls, Brandy Dieterle, and Jennifer Roth Miller. Safely social: User-centered design and difference feminism. *Composing Feminist Interventions*, page 391, 2018.
- [58] Kaitlyn Wells and Thorin Klosowski. Domestic abusers can control your devices. Here’s how to fight back. *The New York Times*, April 2020.
- [59] Rhiannon Williams. Google is failing to enforce its own ban on ads for stalkerware. *MIT Technology Review*, May 2022.
- [60] Taylor Paige Winfield. Vulnerable research: Competencies for trauma and justice-informed ethnography. *Journal of Contemporary Ethnography*, 51(2), 2022.
- [61] Delanie Woodlock. The abuse of technology in domestic violence and stalking. *Violence Against Women*, 23(5):584–602, 2017.
- [62] Delanie Woodlock, Mandy McKenzie, Deborah Western, and Bridget Harris. Technology as a weapon in domestic violence: Responding to digital coercive control. *Australian Social Work*, 73(3):368–380, Jul 2020.

A Appendix - Interview Questions

A.1 Scenario 1

In this first scenario, imagine your coworker tells you that they think someone is tracking their location through their phone. They show you their screen and ask you to help them find out if this is happening.

1. Could you guide us through how you would confirm whether someone is tracking your coworker's location?
 - HINT 1: Where in the phone settings could you go to find the source of location sharing?
 - HINT 2: If we go to the Privacy settings, is there anything here that we could do to find the potential source of location sharing, since your coworker wants to confirm whether or not it is happening? We see here a list of apps, and it looks like the app sharing location is Google Maps.
 - HINT 3: We see that Google Maps is tracking their location, but your coworker needs to use Google Maps to get places and doesn't want to turn location services off. Let's check the settings in Google Maps. What would you inspect here, in the app, to see if location is being shared with someone else?
 - HINT 4: Let's take a look at this drop-down menu when clicking the circle here. Is there anything here that might help?
 - SOLUTION: The last step is . . . to go to this option, Location sharing, and you can see that they're sharing their location with Mallory.
2. How easy or difficult was it to find the source of location sharing? Could you tell us why?
3. Your friend wants to stop sharing their location. What directions would you give to your friend to stop the location sharing?
4. How easy or difficult was it to stop the location sharing? Could you explain why?
5. Have you encountered the Google Maps app before?
6. After stopping location sharing, would you suggest your coworker do anything else or take any other actions?
7. If you thought your location was being tracked, like in this situation, what would you do?
 - (If they don't mention what steps they would take) Would you follow the same steps or do it differently?
 - Would you ask for advice? Who would you ask or where would you go for advice?

A.2 Scenario 2

Now, we'll move on to scenario two.

1. In your own words, what do you think spyware is?

Imagine your friend suspects that their significant other has "hacked" their phone. It seems like they have knowledge about messages, emails, downloaded apps, and other information. Even though they share passwords, your friend rarely leaves their phone out of reach, so they don't know when their partner would have had the time to look at this information. Your friend asks you for advice on finding out whether their significant other is remotely accessing their device.
2. What directions would you give your friend to see whether there is spyware or not on their phone?
 - HINT 1: While some spyware can be hidden, in this scenario, the spyware is an app that was downloaded from the app store. How can you tell if one of your friend's apps might be spyware?
 - HINT 2: We could review all your friend's apps. One way to do this is to look at the app library by swiping right on the home screen.
 - HINT 3: We are looking for an application that enables remote access to the device, not the one that has potential security vulnerabilities. Are you familiar with such remote access apps?
3. What are/were you looking for when you are/were looking for an app that can be used as spyware?

Now you and your friend search through every single app on the phone. Your friend points out this app, TeamViewer. Your friend has never used this app, and they remember that their significant other put this app on their phone, falsely claiming it was an antivirus. TeamViewer is what we call a dual-use app that can be used to share the screen, but also can be used as spyware. This app allows another person to temporarily control the device.
4. Were you familiar with the app, TeamViewer?
5. How easy or difficult was it to identify the spyware app? Could you tell us why?
6. After finding out that TeamViewer was being used to remotely access the device without your friend's permission, what precautions or advice would you suggest that your friend keep in mind?
7. Your friend decides that it is safe to remove the spyware. What steps would you take to remove the app?
8. How easy or difficult was it to remove the spyware app? Could you tell us why?

9. After removing the app, would you advise your friend to do anything else?
10. Let's say the phone settings had been changed or tampered with, what could your friend do to make sure they were changed back?
11. If you suspected someone "hacked" your phone, What would you do?
 - (If they don't mention what steps they would take) Would you follow the same steps or do it differently?
 - Would you ask for advice? Who would you ask or where would you go for advice?

A.3 Scenario 3

So now we will move on to scenario 3.

1. What do you think it means when your account is compromised? Imagine that you and a different friend suspect that someone has access to photos they have stored on their iPhone. Some of their photos disappear, and new photos they didn't take appear in their albums. Your friend also notices that new iMessages are appearing, which they never sent. Your friend asks you to help them figure out what is going on.
2. What are some steps you could take to figure out whether someone can see your friend's photos and messages? Could you walk me through this? (What are you looking for?)
 - HINT 1: Your friend says they used to sync their photos and messages onto other devices, but they're not sure how this works. Through what account might this be happening?
 - HINT 2: So we figure out it's probably happening through iCloud account syncing with another device. Where can we go on the phone to check on other devices logged into the iCloud account?
 - HINT 3: Let's go to the iCloud account settings by clicking on [Apple ID/iCloud account settings]. Is there somewhere here where you might find the list of devices?
 - SOLUTION: If we scroll down, we can see that another device is logged in. This is the source of the photo and message syncing.

You and your friend figure out that their iCloud account is synced with another device. Your friend says they don't recognize this device.

3. How easy or difficult was it to find the other device? Could you tell us why?

4. What are the steps you would tell your friend to take to remove the other person from the iCloud account?
5. How easy or difficult was it to remove the other device? Could you tell us why?
6. Would you recommend anything else to your friend?
7. If you thought your iCloud account was compromised, what would you do?
 - (If they don't mention what steps they would take) Would you follow the same steps or do it differently?
 - Would you ask for advice? Who would you ask or where would you go for advice?

A.4 Scenario 4 Version 1

1. What do you think "rooted" or "jailbroken" means? Imagine your friend tells you they think they are being stalked by a coworker, and they went to this FTC website: <https://www.consumer.ftc.gov/blog/2016/09/whos-stalking-what-know-about-mobile-spyware>.

Your friend shows you this part, titled "What can I do if I think a stalking app is installed on my phone?" (see Appendix B.1). They ask you about the second option, "Check to see if your phone has been rooted or jailbroken." Please let us know after you have read through that part the text.

Now, your friend tells you that they want to follow the website's advice and check whether their phone is rooted or jailbroken. They ask for your help finding the kind of app mentioned on the website.

- Is the meaning of rooted or jailbroken clear from the advice?
 - In your own words, what do you think it means when a person gets full control of the operating system?
 - Can you identify the website's recommendation for people who think a stalking app might be installed on their phone?
2. How would you help your friend find the kind of app mentioned on the website?
 3. (If they find apps) Would you recommend any of these apps to your friend? Why or why not?
 4. What would your criteria be for recommending a root checker app to your friend?
 5. How practical do you think the FTC website's advice to find an app, on a scale from 1 to 5, 1 being very easy to implement and 5 being very difficult to implement? Why did you give this rating?

A.5 Scenario 4 Version 2

1. What do you think “rooted” or “jailbroken” means? Now imagine that your friend tells you that they think their iPhone is jailbroken, and they went to this Avast website: <https://www.avast.com/c-jailbreaking#topic-6>. Please read this section called, “What does jailbreaking an iPhone do?” (see Appendix B.2.1) and let us know after you have read through that part of the text.

- Is the meaning of “jailbroken” clear from the text?
- In your own words, what do you think “modifies the operating system” means?
- In your own words, what do you think giving “unauthorized root access” means?

Your friend then shows you this part, starting with: “Nevertheless, if you think you have a jailbroken iPhone, there are some things you can do” (see Appendix B.2).

2. Can you identify the website’s recommendations for people who think their iPhone may be jailbroken?
3. Your friend wants to look for Cydia first. How would you do this?
- HINT: Search for apps using search bar in settings, search bar, or on home screen
4. It looks like Cydia is not on the phone. Your friend now wants to follow step two, which suggests performing a factory reset. Can you help your friend do this?
- HINT: The option may be in Settings
 - HINT: We could look in “General” settings.
 - HINT: At the bottom, there is a “Reset” option.
 - HINT: Restoring to factory settings means deleting all data and settings. Which option here would do that?
5. Have you encountered this [reset menu] screen before?
6. How easy or difficult was the process of looking for Cydia? Could you tell us why?
7. How easy or difficult was the process of doing a Factory Reset? Could you tell us why?
8. How straightforward or easy to follow were the instructions provided by the website? Could you tell us why?
9. If you thought your phone was jailbroken and wanted to check your device, what would you do?
- (If they don’t mention what steps they would take) Would you follow the same steps or do it differently?
 - Would you ask for advice? Who would you ask or where would you go for advice?

B Appendix - Online Advice for Jailbreak Detection (Scenario 4)

Below are excerpts from Scenario 4’s online advice articles.

B.1 Advice to Find a Root Checker App.

Our study used advice from a now unavailable FTC blog post, which directed people who suspect they are being stalked to download a “root checker app,” in order to detect whether their phone is rooted or jailbroken [15]:

Check to see if your phone has been “rooted” or “jailbroken.” Stalking apps aren’t sold through typical app stores. In addition, they usually can be installed only on a phone that has been “rooted” or “jailbroken,” which allows a person full control over the phone’s operating system. If your phone is rooted or jailbroken and you didn’t do it, a stalking app could be installed. “Root checker” apps can quickly tell you whether a phone has been rooted or jailbroken.

Since beginning this study, the FTC published an article about stalking apps containing similar advice [8]:

Check to see if your phone has been “rooted” or “jailbroken.” Stalking apps can be installed only on a phone that has been “rooted” or “jailbroken,” which gives a person full control over the phone’s operating system. “Root checker” apps can quickly tell you whether a phone has been rooted or jailbroken. But if there is stalkerware on the device, the abusive person may see this activity. If you find that your phone has not been rooted or jailbroken, but the person knows more than they should about your phone or online activities, it may be that they are getting that information from your phone another way.

B.2 Advice to Find Cydia and Reset

In the second version of the interview, advice from Avast [12] was presented to the participants. The article recommends some actions readers can take if they think their iPhone may be jailbroken. We showed the participants two parts: 1) the section called “What does jailbreaking an iPhone do?” B.2.1 that explains jailbreaking and what it allows the users to do on their device, and 2) the section called “Can you tell if a phone has been jailbroken?” that has instructions we asked participants to follow B.2.2.

B.2.1 What does Jailbreaking an iPhone do

Jailbreaking an iPhone **modifies the operating system**, giving you **unauthorized root access** to the jailbroken device’s

core software and structure. So, what can you do with a jailbroken iPhone? Besides slipping through the wormhole to the underground jailbreaking community, and potentially exposing your device to hackers and viruses, there are some reasons why jailbreaking an iPhone or other iOS device might be desirable. With a jailbroken phone, you can:

- Freely do as you please with your phone or tablet.
- Access third-party apps outside the official App Store.
- Customize and personalize your phone and its settings more deeply.
- Unlock carrier restrictions.

B.2.2 Can you tell if a phone has been jailbroken?

Nevertheless, if you think you have a jailbroken iPhone, there are some things you can do.

1. Find Cydia: On your iPhone, search for Cydia, the alternative app store. Even if it's hidden, this search will find the app. If Cydia is there, it's a jailbroken phone.

2. Restore factory settings: If you don't want to worry about whether or not your phone was jailbroken, an easy way around is to restore factory settings. Restoring factory settings brings back whatever may have been lost to jailbreaking.

C Appendix - Demographics

Our screening survey included questions about age, gender, race/ethnicity², education³, computer science and/or internet technology (CS/IT) education, CS/IT work experience, and income. Only P4 had CS/IT education and CS/IT work experience.

ID	Age	Gender	Race	Education	Income
P1	57	F	W	M	\$60-70k
P2	22	F	W	SC	\$20-30k
P3	34	F	W	M	\$20-30k
P4	75	M	AS	P	Prefer not to respond
P5	23	F	AS	B	\$10-20k
P6	36	M	AS	B	\$50-60k
P7	50	M	AS	B	\$60-70k
P8	26	M	H, W	B	\$70-80k
P9	36	F	H, NL	B	\$90-100k
P10	37	F	W	B	\$100-150k
P11	24	M	W	B	\$50-60k
P12	22	F	H, W	B	\$90-100k
P13	22	F	AS	B	\$90-100k
P14	24	F	AA, H	A	≤ \$10k
P15	27	M	AS	B	\$10-20k
P16	19	F	AA, W	SC	\$70-80k
P17	23	M	AA	B	\$40-50k
P18	30	M	H, NL	M	\$50-60k

²Race/ethnicity: AA = African American/Black, AS = Asian, H = Hispanic/Latino/Latina/Latinx, W = White, NL = Not Listed

³Education: A = Associate's degree (2-year), B = Bachelor's degree (4-year), M = Master's degree, P = Professional degree (JD, MD), SC = Some college, no degree

D Appendix - Codebook

Category	Codes	Definition	Scenario(s)	Participant Count
Advice Given by Participant	backup	Back up photos/messages	S3	1
	blocking	Block the person who was tracking the location	S1	1
	change passcode	Change password	S2, 3	3, 9
	confrontation	Directly confront the person responsible for stalking or surveillance, e.g. find the stalker, have a conversation with abusive partner	S1, 2	1, 5
	consult personal contact	Ask friends, family members, or personal contacts for advice; tech savvy personal contacts - could be in negative form (e.g. don't know anyone to consult)	S1, 2, 3, 4.2	2, 4, 2, 2
	consult police	Consult police or law enforcement	S1, 2	1
	delete app	Delete the app	S2	9
	delete more	Take more steps to delete the app, beyond just removing app (deleting app purchase history, deleting from cloud, deleting it from the other phone, etc)	S2	7
	do nothing	Would not advise the friend to do anything else	S1, 2, 3	4, 2, 1
	don't allow	Do not let other people take some action (e.g. download an app on your device)	S2	10
	internet	Use internet (search engine, forums, YouTube)	S1, 2, 3, 4.2	5, 5, 6, 6
	investigate	Find out how the situation happened and conduct test to look into the issue	S1, 2	1, 10
	log out	Log out other devices from the iCloud account	S3	1
	manual check	Go through settings manually and change them back	S2	9
	mfa	Use multi-factor authentication	S2, 3	1, 2
	monitoring	Regularly check and review apps, settings, details, and/or logged in devices	S1, 2, 3	6, 10, 7
	no advice	Wouldn't ask for advice	S1, 3, 4.2	2, 3, 3
	other apps	Check or change settings on other apps (e.g. social media or messaging apps)	S1	2
	remove device	Remove device from iCloud setting (suggested this before being prompted)	S3	2
	reset	Reset the device	S2, 3	10, 1
restart the phone	Turn off phone and turn it back on	S1, 2, 3	2, 2, 2	
tech support	Consult customer service	S1, 2, 3, 4.2	2, 4, 3, 3	
turn off location service & bluetooth	Set location service setting to "Not Allow" and turn off bluetooth	S3	1	
VPN	Use a VPN	S1, 2	1,4	
Criteria for recommending a root checker app to your friend?	ask device owner	Would ask device owner what they were ok with	S4.1	1
	internet	Would search online for an app	S4.1	8
	privacy	Apps should be clear about how personal informatio is being used	S4.1	1
	security	App itself should be secure	S4.1	1
	user review	Many views on a video, good feedback/review for the app	S4.1	2
	videos	Would watch videos to get information leading to a root checker app	S4.1	4
Difficulty	clear or simple	The advice was clear or simple to participants; easy to understand	S4.2	7
	confused	Participant stated they were confused	S2	3
	deleting is easy	All participants found removing the app to be easy	S2	18
	difficult	Resolving the problem was difficult.	S2, 3, 4.1	14, 8, 3
	don't know	Participant said they did not know or had no idea	S1, 2, 3	9, 14, 8
	don't know full capability	Participant did not know apps' full capabilities of e.g. assumed capabilities solely by the name/logo of the app	S2	6
	easy	Resolving the problem was easy.	S1, 2, 3, 4.1, 4.2	1, 18, 17, 2, 7
	familiar app search	Participant was familiar with the process of searching for an app	S4.2	2
	familiar reset	Participant was familiar with the process to reset a device	S4.2	3
	few steps	Process did not take many steps	S1, 3	2, 3
	find-source difficult	Finding the source of location sharing is the Google Maps was difficult.	S1	8
	find-source easy	Finding the source of location sharing is the Google Maps was easy.	S1	3
	find-source moderate	Finding the source of location sharing is the Google Maps was moderate.	S1	5
	frustrated	Ready to give up, struggled, expressed frustration or exasperation	S2	1
	hard to find	Participant appeared to struggle to find the right options/settings (observation)	S1, 2, 3	4, 9, 10
	if I knew	After getting the hint or being shown the solution, participants said they would have found it easy to resolve the problem, if they had known this	S1, 2, 3, 4.2	9, 5, 5, 2
	interview format limitation	Limitations caused by the interview formation: not being able to control screen, not recognizing emails or user IDs	S1, S3	2, 5
	intuitive	Process was self-explanatory, easy	S1	6
	many steps	Process took a lot of steps	S1	4
	moderate	Resolving the problem was neither easy or difficult.	S2, 4.1, 4.2	2, 3, 3

	not me	Problem is not personally relevant (e.g. never thought someone tracking me, i'm not that interesting, haven't experience this before)	S1, S2	2, 2
	obvious label	The settings menu/option is easily recognizable and comprehensive (e.g., red big bold "remove" [button])	S3	6
	order	Participant found the order in the instructions helpful	S4.2	1
	reset is easier	Participant considered the reset step to be easier than finding the Cydia app	S4.2	1
	stop-source difficult	Stopping sharing from the Google Maps location sharing screen was difficult.	S1	1
	stop-source easy	Stopping sharing from the Google Maps location sharing screen was easy.	S1	12
	stop-source moderate	Stopping sharing from the Google Maps location sharing screen was neither easy or difficult.	S1	1
	unfamiliar	Had not heard of it or wasn't aware of the process	S2, 3, 4.2	4, 6, 2
	unfamiliar with app	Unfamiliar with TeamViewer	S2	9
	unfamiliar with reset	Didn't know how to reset device, expressed lack of familiarity with reset	S2	2
	unintuitive	Unintuitive to navigate or solve problem	S3	2
	wording	Unsure which menu options/labels to choose; found wording confusing	S4.2	5
Other	conditional	Participant's mentioned technical/personal condition or preferences e.g. depends on how paranoid they were	S1,2	2,3
	escalation	Participant mentioned personal experience relating to scenario or to the strategy used in the scenario (e.g., sharing location with family members)	S1	1
	recommendation	Recommended or imagined a feature that does not currently exist	S1, 2, 3, 4.2	4, 3, 1, 5
	study design limitation	Participant was influenced or limited by study design	S2	2
	exception	Participant mentioned social condition for exception to privacy & security advice due to trust or consent, e.g. if you trust them	S2	3
	trusted source	Mentioned that they trust the information on the website or provided by us	S4.2	3
Practicality of FTC advice	more details	Need more clarification, more elaboration, better keywords for finding app	S4.1	3
	no example	No example app given	S4.1	2
	no results	Nothing came up from search	S4.1	2
Search Keywords Used by Participant	best app to tell if your iPhone is jailbroken		S4.1	1
	best root checker app		S4.1	1
	how to check if my iPhone has been jailbroken		S4.1	1
	how to diagnose jailbroken iPhone		S4.1	1
	jailbreak checker		S4.1	1
	jailbreaker		S4.1	1
	root		S4.1	1
	root checker		S4.1	5
	root checker app		S4.1	2
	root checking		S4.1	1
	rooted		S4.1	1
	security check		S4.1	1
	stalk		S4.1	1
stalker		S4.1	1	
	add a new share	Attempted to add a new person to share location with	S1	3
	antivirus	Suggested using antivirus to solve problem	S2	1
	app store	Went to Apple's App Store to review apps	S2, 4.1	2, 7
	change location services	Changed location service settings, e.g. from Always to Never, or Precise Location, only while using	S1	3
	change other google maps' setting	Incognito mode, block person, go to myaccount.google.com and click X	S1	4
	change privacy settings	Suggested changing privacy settings	S2	5
	change TeamViewer app setting	Changed TeamViewer app setting (background app refresh)	S2	1
	check device list	Checked the device list	S3	1
	concluded there's no app	Concluded that Cydia is not installed on the device	S4.2	5
	confused by other apps	Confused by many unfamiliar apps, also includes participants who mistakenly identified HackIt or other app as spyware	S2	4
	confused by other features in an app	Confused by other Google Maps features, e.g., click "fix" for notifications pop-up, clicked blue arrow	S1	6
	control center	Suggested looking at Control Center settings in iOS	S1	1
	download Cydia	Suggested downloading Cydia or using an alternate app store to find it	S4.2	2

Participant Strategy	engage with suspicious app	Engaged with (opened or tried to sign up for) a suspicious or unknown app	S2	5
	Erase All Content and Settings	Chosen reset option	S4.2	3
	experience	Mentioned personal experience relating to scenario or to the strategy used in the scenario (e.g., sharing location with family members)	S1, 2, 3	1, 8, 5
	factory reset	Suggested doing a factory reset	S2	1
	false alarm	Misunderstood or got suspicious on settings that are irrelevant	S4.1	1
	familiar	Familiar with the app, feature or process	S3	4
	get rid of phone	Suggested getting rid of the phone	S2	2
	guessing	Guessed	S2	1
	hidden	Mentioned possibility of hidden apps or files	S2, 3	3, 2
	home screen	Searched for Cydia on the home screen	S4.1, 4.2	2, 2
	internet	Use internet (search engine, forums, YouTube) for strategy or advice	S4.1	8
	messages	Went into "Messages" settings	S3	3
	not my phone	Mentioned how it might be different if it were their phone, or if they need to take a look at their own phone to figure it out	S2	3
	not possible	Mentioned that doing something was not possible (e.g., finding source of iCloud login)	S3	1
	other apps	Mentioned checking or changing settings on other apps, like social media apps or messaging apps	S3	1
	other iCloud settings	Searched in iCloud settings but could not easily find device list and looked into different settings	S3	9
	password & security setting	Went into "Password and Security" iOS iCloud account setting to try to find other device	S3	4
	photos	Went into "Photos" settings	S3	12
	recognition	Asked interviewers if the friend recognizes the app	S2	5
	reset all settings	Chose "Reset All Settings" in the reset menu, which doesn't delete the entire data	S4.2	7
	scroll down in settings	How participant reviewed the apps	S2	9
	search bar	Utilized iOS search bar (from settings or home screen)	S1, 2, 4.2	2, 1, 8
	search in settings	Searched Cydia from the search bar in the Settings	S4.2	1
	settings	Searched Settings to find a root checker app	S4.1	3
	Share my location	Looked at Share My Location or Find My settings	S1	8
	shared album	Checked or commented on "Shared Albums" setting in iOS Photos settings	S3	6
	tracking	Went to "Tracking" setting	S1, 2	5, 3
	turn off location service	Turned off the location service completely	S1	11
	unsure if app is installed	Unsure whether Cydia could still be on the phone, after not finding the app	S4.2	4
	What were participants looking for?	cross-app	Connection across multiple apps could be an indicator of spyware	S2
data analytics		Spyware may look similar to data analytics	S2	1
keywords		This could be an indicator of spyware	S2	5
location sharing		This could be an indicator of spyware	S2	1
power consumption		This could be an indicator of spyware	S2	1
spyware finder app		Finding an app to look for spyware apps	S2	1
weird name		Spyware app may have peculiar name	S2	2

If You Can't Get Them to the Lab: Evaluating a Virtual Study Environment with Security Information Workers

Nicolas Huaman ^C Alexander Krause ^C Dominik Wermke ^C Jan H. Klemmer ^{*}
Christian Stransky ^{*} Yasemin Acar [†] Sascha Fahl ^C

^C*CISPA Helmholtz Center for Information Security, Germany,*

{nicolas.huaman, alexander.krause, dominik.wermke, sascha.fahl}@cispa.de

^{*}*Leibniz University Hannover, Germany, {klemmer, stransky}@sec.uni-hannover.de*

[†]*George Washington University, USA, acar@gwu.edu*

Abstract

Usable security and privacy researchers use many study methodologies, including interviews, surveys, and laboratory studies. Of those, lab studies allow for particularly flexible setups, including programming experiments or usability evaluations of software. However, lab studies also come with challenges: Often, it is particularly challenging to recruit enough skilled participants for in-person studies. Especially researchers studying security information workers reported on similar recruitment challenges in the past. Additionally, situations like the COVID-19 pandemic can make in-person lab studies even more challenging. Finally, institutions with limited resources may not be able to conduct lab studies.

Therefore, we present and evaluate a novel virtual study environment prototype, called OLab, that allows researchers to conduct lab-like studies remotely using a commodity browser. Our environment overcomes lab-like study challenges and supports flexible setups and comprehensive data collection. In an iterative engineering process, we design and implement a prototype based on requirements we identified in previous work and conduct a comprehensive evaluation including a cognitive walkthrough with usable security experts, a guided and supervised online study with DevOps, and an unguided and unsupervised online study with computer science students. We can confirm that our prototype supports a wide variety of lab-like study setups and received positive feedback from all study participants.

1 Introduction

Laboratory studies are common in usable security and privacy research and find broad application in many experiments with end-users or expert users. Researchers can flexibly set up very

specific study environments and collect a wide variety of data, including video and audio recordings [31, 50, 41], think-aloud data [2, 16, 5, 37], or user behavior for particular software and tooling [4, 12, 10, 5]. While laboratory studies support flexible experimental setups and data collection, they come with the following challenges:

Participant Recruitment. It is often challenging to recruit sufficiently skilled participants for in-person lab studies. Due to the geographic location of a research lab or specific requirements for participants such as age [35], gender [19], or professional experience [2, 5, 30, 47], a laboratory study might not be feasible. In all scenarios, conducting expert studies with *security information workers* (SIWs) to test security development and system design [3, 11, 1, 37, 23, 10], system configuration and administration [16, 46, 45], or test and analyze those systems' security [31, 12, 5, 41, 32] is challenging. Local expert participant pools are usually too small and might lack diversity, representativeness, or statistical power. These challenges required researchers to be pragmatic in their study design, e. g., by recruiting computer science students as stand-ins for developers for lab studies [2, 24, 25, 16, 12, 46, 32] or by simplifying programming tasks to a few lines of code that can be studied online. Hence, researchers have started to conduct expert studies remotely [43, 38, 27, 5, 50, 37].

Complicated Circumstances. Laboratory studies in-person are feasible as long as no circumstances prohibit inviting participants to a research lab. However, events such as the COVID-19 pandemic make in-person lab studies even more challenging. Alternatively, researchers conduct studies online [52, 1, 4, 31, 10] dealing with the same restrictions as described above. Additionally, lab studies require certain resources, including space, personnel to supervise participants, and equipment, e. g., workstations to conduct studies.

To address the challenges above, we make the following contributions:

- **We use a literature-based requirements engineering process** to identify requirements for typical lab studies in usable security and privacy research based on previous

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

work. Therefore, we analyze 24 publications, including SIW studies, since they require particularly skilled participants who might be geographically widely distributed and usually hard to recruit.

- **We design and implement a virtual study environment prototype** that we call *OnlineLaboratory* (short OLab). OLab supports highly flexible lab-like online studies with extensive data collection. This virtual study environment allows researchers to recruit participants from anywhere and conduct highly customized studies in a commodity browser. Researchers can freely choose operating systems and tooling and collect a wide variety of data from participants, including edited files, copy and paste events, browser histories, and screen and audio recordings.
- **We evaluate our virtual study environment prototype in three studies** (Section 4) to illustrate its applicability to security studies with expert users. First, we conduct a cognitive walkthrough with four usability experts; second, a guided study with nine experienced DevOps; and third, an online programming experiment with 16 computer science students.
- Based on the evaluation results, we iteratively improved OLab’s usability and user experience.

We use a literature-based requirements engineering process to design and implement a virtual study environment prototype that we call *OnlineLaboratory* (short OLab). OLab supports highly flexible lab-like online studies with extensive data collection. This virtual study environment allows researchers to recruit participants from anywhere and conduct highly customized studies in a commodity browser. Researchers can freely choose operating systems and tooling and collect a wide variety of data from participants, including edited files, copy and paste events, browser histories, and screen and audio recordings.

Figure 1 depicts the overall structure of this work. We provide further information regarding OLab on an accompanying website.¹

While we designed and evaluated our prototype in the light of usable security experiments with SIWs, we are convinced that it can be seen as a blueprint for a general-purpose platform to conduct lab-like usable security and privacy user studies with expert users and end-users, during the COVID-19 pandemic and beyond.

2 Related Work

We discuss related work focusing on security information workers in three key areas: Laboratory experiments, remote studies that are not browser-based, and browser-based online studies. Finally, we aimed to identify the most recent and

¹<https://publications.teamusec.de/2022-soups-olab/>

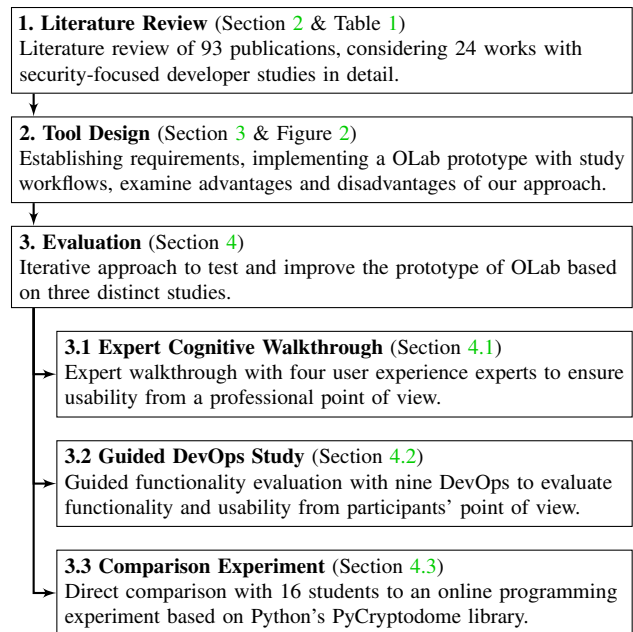


Figure 1: Diagram illustrating literature review, OLab’s design, and its evaluation.

relevant requirements for OLab based on task-based studies with SIWs, published at high-quality venues in the recent past (cf. Table 1). Therefore, we included work from USENIX SOUPS, ACM CHI, IEEE Security and Privacy, USENIX Security, NDSS, ACM CCS, and (Euro)USEC. We use them as a foundation to design, implement, and evaluate our OLab environment.

Lab Studies. Conventional lab studies often focus on participants performing a task on-site, either with a researcher or while being observed.

Acar et al. investigated in a lab study with 54 Android developers, how different information sources affect code security when solving security and privacy-related programming tasks [2]. Krombholz et al. conducted a lab study with 28 participants that had to securely configure TLS on a web server to explore and identify usability challenges in that process [16]. Follow-up work by Tiefenau et al. applied the original study methodology to *Let’s Encrypt* and *Certbot*, finding better usability leading to a higher number of secure TLS deployments [46]. Naiakshina et al. performed a qualitative usability study with 20 computer science students to understand better how developers handle secure password storage [24]. In 2018, Naiakshina et al. replicated a study examining ecological validity by priming participants using the deception of a real-world task; they concluded that it has a significant impact, resulting in more secure solutions [25]. Hänsch et al. examined the understanding of obfuscated source code in reverse engineering process in a lab study with 66 students [12]. Nosco et al. proposed a new search strategy for finding bugs

and software vulnerabilities. They grouped 12 participants into small teams within a ten-day lab experiment and had to discover vulnerabilities in several services [30]. Smith et al. conducted both a heuristic walkthrough and a lab study evaluating the usability of four security-focused static analysis tools, finding several usability issues. In the lab study, 12 developers had to fix warnings reported by those tools [41]. A similar study by Tupsamudre et al. identified several usability problems in two open-source *Static Application Security Testing* (SAST) tools; in a lab study, eight developers had to solve a password storage task in a web application while using the tools [47]. Plöger et al. conducted a lab study evaluating the usability of the Clang Static Analyzer and libFuzzer with 32 local CS Students and Capture-the-flag players, finding that libFuzzer performs a lot worse on usability compared to Clang Static Analyzer [32]. All in all, this research has in common the limitation of a local sample, often using students as stand-ins for developers or administrators.

Remote Studies. Besides conventional lab studies, some study setups work remotely over the internet, using online calls or self-reporting so participants can solve the tasks in any location using their own computers.

For example, Ruef et al. presented a novel cybersecurity contest that included breaking code and encouraging developers to build secure applications [38]. The authors hosted the contest and evaluated the solutions via an automated system regarding security. Based on the contests, Votipka et al. conducted an in-depth qualitative analysis to understand common security mistakes made by developers [49]. Nguyen et al. demonstrated in a remote experiment with 39 Android developers and students that IDE security plugins can be an effective measure helping developers with writing secure code [27]. Aksu et al. evaluated the open-source vulnerability scanner *OpenVAS* concerning its usability by employing heuristic walkthroughs and an experiment with 10 security experts at a single cybersecurity company. The participants had to solve six different tasks, ranging from scanning a system to choosing remediation actions [5]. In 2021, Roth et al. investigated the misconceptions web developers have with *Content Security Policies (CSP)* through a qualitative Zoom interview study with 12 participants [37].

Multiple papers replicated and extended studies by Naiakshina et al. [24, 25]: In 2019, Naiakshina et al. showed that online freelancers behave similarly as students [23]; in 2020, Danilova et al. replicated the original lab study with freelancers and the deception of a real-world project which had a negligible effect [7]; also in 2020, Naiakshina et al. demonstrated that professional developers perform better than students and freelancers [22]. Another study of Votipka et al. conducted semi-structured observational interviews with 16 reverse engineers to understand the reverse engineering process and to improve interactions with reverse engineering tools [50]. Several of these studies involved downloading

code and uploading solutions, with no possibility of observing intermediate attempts or behavior.

Browser-Based Studies. A specific type of remote studies utilizes browser-based environments that participants can access via a web browser developed explicitly for the respective study. These are more closely related to OLab.

For example, Yakdan et al. conducted an online experiment to evaluate the usability of different decompilers for reverse engineering with nine professional malware analysts and 21 students [52]. Oliveira et al. conducted an experiment with 109 developers who had to solve six programming puzzles in Java, which include so-called *API blind spots* [31]. The results underline the importance of well-designed APIs, as (security) blind spots reduced the number of functional and secure solutions.

Acar et al. evaluated the usability of different cryptography Python APIs with 256 GitHub developers, who had to solve basic cryptography tasks with the APIs in a web-based study environment [1]. Based on this setup, Gorski et al. conducted another experiment with 53 developers, examining the effect of integrated security advice and warning messages on code security [10]. Furthermore, Fischer et al. evaluated the effect of Google search ranking results on code security and functionality with 410 GitHub developers using the same setup [9]. In another paper, Acar et al. conducted an online experiment with four different security-critical programming tasks (e. g., encryption, password storage) with 307 developers from a GitHub convenience sample to assess the validity of experiments with GitHub users [4]. Based on this study and the previously mentioned one by Acar et al. [1], Stranisky et al. presented a browser-based virtual laboratory called *Developer Observatory* and experiences from using it for developer studies [43]. The main idea of *Developer Observatory* is similar to this paper's approach, but limited to languages supported by Jupyter Notebook kernels [15]. OLab follows a more holistic approach, combining multiple different steps (e. g., introduction, consent form, tasks, surveys, information pages) in a single integrated workflow.

To summarize, these publications made significant contributions to our research community, thus underlining the importance of controlled experiments in which software developers, operators, and others solve tasks. Therefore, we derive and evaluate requirements for a remote study platform to facilitate research with such methods.

3 OLab Design and Implementation

In this section, we describe the requirements we identified in previous work, illustrate constructed study workflows for OLab both from a researcher's and participant's point of view, and discuss key features of OLab.

Table 1: Overview over our categorization of related work in the field of developer security.

	Year	Type	Channel	N	Time	Recruitment	Environment	Tools (Explicit Mention)	Data Collection	In OLab
		Security Analysis Security Configuration Development Task	Lab Remote Browser Based	Number of participants	Duration in hours	Organisations Online Mass Recruitment Online Targeted Recruitment Local University Students	Windows Linux Variable Unspecified	Git/Gitlab Custom CA IDE Android Studio Plugin Static Analyzer/Fuzzer Browser-Based Platform Android Emulator Qualtrics Survey Programming Language ¹	Screen Recordings Transcripts Video Recording Audio Recording Observation Notes Participation Diaries Think Aloud Source Code Copy and Paste Events Eclipse Actions Config Files Bash History Gitlab Issues/Comments Copy of image Qualtrics responses Browser History	Implemented Possible Impossible
[38]	2016		●	156	3x2w	●	●	C		●
[2]	2016	●	●	54	01:00	●	●	J		●
[52]	2016	●	●	30	01:00	●	●	C		●
[1]	2017	●	●	256	01:00	●	●	PY		●
[24]	2017	●	●	20	08:00	●	●	J		●
[27]	2017	●	●	39	01:00	●	●	J		●
[4]	2017	●	●	307	01:00	●	●	PY		●
[16]	2017	●	●	35	02:00	●	●	J		●
[25]	2018	●	●	40	08:00	●	●	J		●
[31]	2018	●	●	109	>00:20	●	●	J		●
[12]	2018	●	●	66	00:47	●	●	J		●
[10]	2018	●	●	53	?	●	●	PY		●
[46]	2019	●	●	31	05:00	●	●			●
[5]	2019	●	●	10	?	●	●			●
[23]	2019	●	●	43	06:30	●	●	J		●
[50]	2020	●	●	16	01:10	●	●			●
[7]	2020	●	●	43	72:00	●	●	J		●
[22]	2020	●	●	36	08:00	●	●	J		●
[30]	2020	●	●	20	80:00	●	●	C,PY		●
[41]	2020	●	●	12	01:00	●	●	J,C,P		●
[47]	2020	●	●	8	00:30	●	●			●
[32]	2021	●	●	38	20:00	●	●			●
[37]	2021	●	●	12	01:33	●	●	J,PY,P		●
[9]	2021	●	●	410	?	●	●	-		●

¹ J=Java; PY=Python; C=C/C++; P=PHP

3.1 Identifying Requirements

To identify requirements for the OLab environment, we considered all identified previous work (cf. Section 2). We started with high-level categories (cf. Table 1), collecting prevalent study environments, tools, and approaches. Six researchers created, merged and revised categories jointly and then decided on definitions based on these categories, resulting in our final codebook. Two or more researchers used “iterative categorization” [26] and re-coded all publications using the final codebook, resolving any emerging conflicts immediately, so we refrain from reporting the inter-rater reliability (IRR) [20].

Diverse Study Setups. We identified three different types of tasks for SIWs. 13 studies (54.1%) included security development tasks, referring to the implementation or use of security relevant source code (e. g., studies investigating the use of cryptography libraries). This type of study was most common in our dataset. Less common were 9 security analyses (37.5%), which included tasks such as reverse engineering binaries or finding vulnerabilities in code. In these studies, researchers provided participants with example binaries and tools. Additionally, 2 papers (8.3%) included security configuration studies. They provided participants with a setup that they should configure to be secure, e. g., a server stack. To move such lab studies to an online environment, OLab needs

to be capable of handling diverse study setups. These setups include providing and editing source code, configuration files, network connections, and running arbitrary applications.

High Accessibility for Participants. The top 5 studies with most participants (between 156 and 410) all were either remote- or browser-based. Browser-based studies rely on online mass recruitment, using platforms like Amazon MTurk or emails to reach developers globally. 10 studies (41.6%), of that 7 Lab Studies (70% of Lab studies) relied on university students for their sampling, only two of which recruited additional non-student participants to improve their sample diversity and size. To address the limitations of Lab study recruitment and allow for more diverse sampling procedures, OLab should obtain the ease of browser-based studies. It needs to be easy to access using a commodity browser. Furthermore, it should scale to many concurrent participants.

Data Collection. In previous work, researchers collected a wide variety of data from participants, including source code (used by 16, 66.6%) and browser profiles (7 studies, 29.2%). They also tracked copy & paste events (6 studies, 25.0%) and more fine-grained browser or IDE behavior (1 study, 4.2%). Furthermore, they recorded screen and audio (3, 12.5% and 4, 16.7%, respectively). Hence, OLab needs to be able to collect all of the above information.

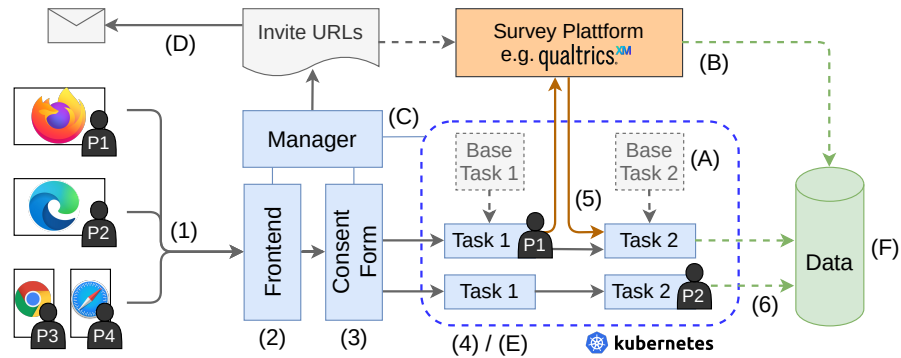


Figure 2: Overview of a typical setup with OLab. Walkthroughs labeled for researchers (A–F) and participants (1–6). See Section 3.2 for an in-depth label description.

3.2 Study Workflow

This section describes the interaction process between OLab, researchers, and study participants. Figure 2 illustrates an example study setup both from a researcher’s (A–F) and participant’s (1–6) perspective.

Participant Perspective. Steps 1–6 in Figure 2 illustrate the participants’ perspective.

- (1) **Receive Invite:** Invitees can participate in a study remotely by accessing a (unique) invite URL with a HTML5-capable commodity browser on a desktop or laptop computer, using a sufficiently stable internet connection (validated for 8.0 Mbit/s downlink and 0.8 Mbit/s uplink).
- (2) **Landing Page & Consent Form:** After clicking the invite URL, OLab forwards invitees to a landing page showing study information and a consent form (cf. Figure 7a).
- (3) **Briefing:** After giving consent, OLab presents participants a full study description, including an introduction to the study environment (cf. Figure 7b).
- (4) **Solving Tasks:** Participants are encouraged to work on tasks in full-screen mode, look up the study and task descriptions with a mouse click, skip a current task, or finish the entire study. OLab aims to provide a working experience as close to a regular desktop environment as possible (cf. Figure 7c).
- (5) **Survey Questionnaires:** At any point in a study, OLab allows researchers to forward participants to external websites, including surveys (e. g., using Qualtrics).
- (6) **Debriefing & Exit:** After solving all tasks, OLab allows researchers to forward participants to an exit survey and a debriefing website.

Overall, we designed and implemented OLab to be unobtrusive, engaging, and fail-safe for participants.

Researcher Perspective. Steps A–F in Figure 2 illustrate the researcher’s perspective.

- (A) **Setup Study Environment:** During the study setup, researchers can freely choose operating systems, applications, tools, file access, and connection control.
- (B) **Setup Tasks & Conditions:** OLab supports within-subjects, between-subjects, and mixed studies. Tasks and conditions can be randomized or arranged using the *Latin squares* method [42].
- (C) **Scaling:** OLab is based on a highly scalable *Kubernetes* cluster [17] and allows researchers to run studies in different geographical regions with many concurrent participants to optimize connection speeds and scale available environments.
- (D) **Generate Invites:** OLab supports individual invite tokens for participants, forgoing the need to save participants’ personally identifiable information (PII). Invite tokens can be used to track participants across other services (e. g., Amazon’s MTurk).
- (E) **Study Progress:** Researchers can track the study progress and modify and manage scaling options using a dashboard.
- (F) **Data Access:** After study completion, researchers can gather the collected data (e. g., specific study metrics, metadata, and questionnaire answers) with a mouse click.

3.3 Key Features

Below, we illustrate key features, and discuss their advantages and limitations.

Common Task Support. OLab supports the automation of common tasks. These include collecting informed consent before starting a study and integration with external tools that provide surveys during and after the tasks. The tool automatically stores collected data on a per task and participant base.

Scalability. The OLab prototype allows scaling resources up and down to adapt to the number of concurrent participants. Furthermore, the prototype allocates resources dynamically for all participants. This scaling is possible as OLab relies on a *Kubernetes* cluster [17] to spin up, secure, scale, and orchestrate study environments. For more technical details, we refer to Appendix A.

Supported Study Types and Tasks. OLab can cover numerous study types and tasks. To refer back to the related work evaluation in Table 1, we commonly identified developer studies utilizing programming tools like IDEs and git (8, 33.3%) which OLab supports. Secure configurations allow for apps requiring restricted or encrypted network access (e. g., git or web servers). Support for volumes by *Kubernetes* allows for persistent storage across tasks, and GPU support is available [18].

Operating System and Tool Support. Six studies (25.0%) from related work relied on Linux, and one on Windows. Therefore, we decided to support Linux containers mainly.² This setup enables customized virtual environments focusing on the applications relevant to a specific study. In addition, the containers can provide environments for all existing programming languages (e. g., Python, Java, and C/C++). OLab supports full desktop environments with pre-installed applications and configuration files for more complex studies.

Data Collection. OLab allows collecting a large variety of different data including source code, configuration files, or other files. Additionally, OLab observes user interactions by recording copy-and-paste events, mouse clicks, or keyboard strokes. These observations can be further complemented with screen recordings. If required, system events can be captured, e. g., kernel events and logs. Researchers can freely configure all of the above data types.

Internet Connectivity. Unsupervised participants using the internet on containers can theoretically access any resources reachable by the parent network, e. g. the university, or cloud infrastructure (depending on the hosting setup). Researchers can address this by using firewalls or proxies provided by *Kubernetes* for network access. These can be configured as part of the study environment. By default, OLab allows full access to the internet except to other study containers.

Access Control. Since internet connection and other security risks exist with the kinds of setup we provide, we describe measures that we took to allow researchers of OLab to identify individuals misbehaving in the infrastructure: By default, OLab generates secure random tokens. These tokens serve as personalized tickets for participants, which researchers can link to participant profiles (e. g., MTurk profiles). OLab assigns containers to these tokens, tracks timings and container addresses, and participant-specific data, which can serve as a

²For additional OS support, a Windows Docker image (requiring Windows Server with an appropriate license) is available [8].

chain of accountability. Hence, researchers can trace potential abuse back to individual participants.

Participants' User Experience. Overall, OLab aims to provide good usability for participants. First, the effort to participate in a study is low, as participants only need a commodity web browser. Second, OLab allows easy navigation through study parts by offering progress indicators (cf. Fig. 7b) and “Start” and “Continue” buttons. Third, participants can access study and task descriptions at any point. Finally, participants that re-access a study after interruption are by default redirected to their current step instead of having to restart or navigate themselves.

Lab Study Support. In addition to using OLab for online studies, researchers can use it in traditional in-person laboratory settings. In that case, the experiment computer can access the OLab frontend, so that OLab provides the automatic study setup and data collection.

4 Evaluation

Overall, we followed an iterative usability evaluation and engineering approach [29]. We focused on participants we could easily approach (e. g., researchers, local CS students) and stopped recruiting once an evaluation step detected no further usability problems. We conducted studies with smaller but increasing sample sizes, instead of one large-scale usability study, following best practices for usability engineering [48]. We think our approach is suitable to provide good usability for OLab.

We conducted three studies, including (1) a cognitive walkthrough with experienced usable security researchers, (2) an evaluation from the participant's perspective, and (3) a comparison to an alternative online study setup. The first study, a cognitive walkthrough with experienced usable security researchers, focused on gaining first insights into participant usability (Section 4.1). The second study focused on a qualitative usability evaluation from the participants' perspective (Section 4.2). Finally, the third study compared OLab to an online task-download study (Section 4.3). While the first two studies are formative, guided studies to collect feedback for an iterative improvement of the OLab, the third was summative and inspired by a study setup from previous work [1]. This study setup allowed us to construct a well-evaluated version of the participant view. We also chose this setup to evaluate two different sets of expert populations: developers and DevOps. Furthermore, the two setups demonstrate the flexibility of OLab regarding different study types and tasks (e. g., programming and system configuration), different requirements for data collection, and a diverse participant pool.

Below, we summarize the ethical aspects of all three studies and provide an overview of our goals for each study, recruitment, participants' demographics, and limitations. Finally, we

describe the three studies and their results in detail in the following subsections.

Ethics. Our institution did not require formal Institutional Review Board (IRB) approval for the types of studies we conducted in this work. However, compliance with standard IRB requirements is a focus of OLab. Participants agreed to a consent form modeled after IRB-approved consent forms in previous work [51].

We handled the collected data in our studies under strict German data and privacy protection laws and the European Union General Data Protection Regulation (GDPR). Furthermore, to prevent exposure of any data to third parties, the OLab infrastructure runs entirely self-hosted.

Evaluation Goals. During the studies, we aimed for the following evaluation goals:

1. **EG1: Usability.** How well does the OLab follow common usability guidelines?
2. **EG2: Perception.** How do participants perceive studies using the OLab prototype?
3. **EG3: Limitations.** What problems can occur during the study? What requirements do all participants need to fulfill to use our OLab?
4. **EG4: Comparison.** How do studies with OLab compare to other conventional online study approaches?

For EG1, we consider usability goals and rules to be generally unknown to participants. Therefore, we decided to evaluate EG1 by conducting expert walkthroughs (Section 4.1). EG2 and EG3 are the focus of a guided DevOps study (Section 4.2). This study measures physical requirements like hardware and Internet bandwidth, but also collects feedback on the perception of participants regarding the study and uncovers misconceptions. In the third study, we focus on the comparison to other study types as detailed in EG4 (Section 4.3). This unsupervised study identified a few more technical limitations and usability challenges that did not come up in the previous supervised studies.

Recruitment and Demographics. Below, we describe all three studies’ recruiting process and participant demographics. Table 2 provides an overview of the collected demographics. For most demographic questions, we allowed multiple answers (cf. replication package in Section 5).

For the cognitive walkthroughs, we recruited four experienced usable security researchers (Section 4.1). The experts were not involved in the development or previous test phases of OLab. All participants have a Master’s degree or Ph.D. in computer science. The average experience in usability and conducting studies was 3.88 years (median = 3.5). We consider them all experienced usability security researchers, as they actively research and conduct studies in usable security and privacy.

For the second study (Section 4.2), we recruited nine experienced DevOps. We chose three recruitment channels: stu-

Table 2: Demographics for valid participants from all three studies. Omitting “Don’t know”/“Don’t want to answer” answers.

	Expert Walkthrough	Guided Study	Comparison Experiment
Participants			
Started	4	9	23
Finished	4	9	19
Valid ($n =$)	4	9	16
Gender			
Male	50.0%	100.0%	93.8%
Female	50.0%	0.0%	6.2%
Not M/F (Free Text)	0.0%	0.0%	0.0%
Education			
Secondary	0.0%	33.3%	62.5%
Bachelor’s	0.0%	33.3%	37.5%
Master’s or higher	100.0%	33.3%	0.0%
Age (in years)			
Median	27.5	29.0	22.0
Mean (μ)	27.75	29.88	22.06
Std. dev. (σ)	2.5	6.33	1.57
Relevant Experience (in years)[†]			
Median	3.5	-	1.5
Mean (μ)	3.88	-	2.27
Std. dev. (σ)	2.32	-	2.05

[†] Conducting studies and Python programming respectively.

dents from our university that worked in small- and medium-sized enterprises (2 participants), an online forum for DevOps (1 participant), and posts on four Subreddits related to DevOps (6 participants). Three DevOps had secondary education, three had a Bachelor’s degree, and three had a Master’s degree or a Ph.D.

For the third study (Section 4.3), we recruited a sample of 23 computer science students from our university. The study took two hours, and we compensated participants with €100. Four participants dropped out during the experiment, and 19 participants completed the study. We excluded another three participants due to longer breaks. Hence, 16 valid participants completed the study overall. Most of them were male (93.8%; 15). The majority studied for a Bachelor’s degree (62.5%; 10), while the remaining strived for a Master’s degree (37.5%; 6). The average Python programming experience was 2.27 years (median = 1.5).

Limitations. Our studies share limitations common among qualitative and task-based studies, like an opt-in bias concerning participants’ voluntary participation. We recruited people in a snowball sampling from our network for the cognitive walkthroughs. While we believe these are appropriate professionals, they may be biased towards our team and tool. We, therefore, refrain from evaluating and including their usability ratings beyond the walkthroughs themselves and exclude them from further conclusions for EG1, the usability of our tool. We recruited students to perform tasks that might not represent real-world developers in the comparison study. However, students were used in previous studies and found

to be acceptable proxies for professional software developers [4, 39, 44] for the type of study tasks we performed [1]. We explicitly pointed participants to the fact that we aimed to collect self-reported usability assessments for the infrastructure and not single components of the studies (e. g., the IDE we provided). While this worked smoothly for the supervised cognitive walkthroughs to help participants, we could not intervene in the comparison study. Some participants might have misunderstood our framing or explanations, as is natural in meta-evaluation studies. We evaluated descriptions in our pilot run with students to minimize this risk.

4.1 Cognitive Walkthrough

After developing and piloting the OLab environment, we evaluated the usability from four usability experts' points of view via cognitive walkthroughs.

Methodology. We conducted four cognitive walkthroughs in July 2020 during the COVID-19 pandemic using an online meeting software with screen and audio sharing. Two researchers accompanied and recorded each walkthrough with the participants' consent for later transcription. The experts provided usability feedback using different operating systems (macOS, Windows, and Linux) and web browsers (Chrome/Chromium and Firefox). Before the cognitive walkthroughs, we asked the participants to watch an animated video that explained and reminded them about Nielsen's ten usability heuristics [28]. We also told the participants to write down bullet points for each heuristic to remember them during the walkthrough.

We asked the experts to perform a study in the role of participants, except they did not have to solve the provided programming tasks. Instead, the experts should focus on the usability of the OLab prototype. To guide the walkthrough, we identified ten typical workflows for participants within the study environment. The experts had to pass each workflow step to finish the walkthrough successfully. During the walkthrough, we collected usability feedback based on Nielsen's heuristics and further feedback on the user interface (UI) and experience (UX). After completing the walkthrough, we discussed the comments and feedback and implemented the required changes before the following walkthrough.

Results. Table 3 provides detailed background information of the recruited experts, including both their study background and experience within their research field. The experts were generally optimistic about OLab's usability. Each expert completed all walkthrough steps without any significant issues. As Brooklyn summarized it:³ *“Everything was running fine, without any problems. [...] I didn't have any lag, it was like I was on my own system. [...] I didn't even notice that I was not working on my own computer.”* (Brooklyn).

³We translated all quotes in this paper from German to English.

Most expert feedback was on UI and UX. For example, we received feedback to name the buttons and links clearer and more consistent. Brooklyn mentioned that clicks within a popup should not close it and that all UI elements should receive mouseover tooltips or have their text improved to enhance clarity for participants during a study. As a result, we also added a help button to the sidebar (cf. Figure 7). Further on, Charlie suggested better framing of the study process by initially displaying the number of tasks that participants are going to do and generally improving the wording for indicating the study progress. Moreover, Dakota suggested adding functionality for participants to review content from previous pages, e. g., the consent form or introduction videos, and the addition of an information graphic introducing participants to the study scenario.

The remaining feedback focused on the survey's content or structure. Here, some clarifications targeted the consent form (Brooklyn). We implemented a redesign for questions in Qualtrics, so they match the overall layout and design of OLab (Ash). Dakota further mentioned that the consent form should be simplified to reduce cognitive load on participants and to include missing information regarding speed tests we are running in the background. Charlie also noted that OLab should show the consent form as the first item within the study environment.

4.2 Guided DevOps Study

In the second study, we evaluated the usability and participant interaction of OLab in a study with nine DevOps from small and medium-sized enterprises (SMEs). The study structure derived from a different project with DevOps from a local meetup. We then conducted the study within OLab with an additional focus on the usability of OLab. We observed the participants in a think-aloud study. These requirements are ideal for a functionality test since we could ask about the participants' perception of specific OLab prototype features during the study and observe and assist with issues that occurred to improve the prototype iteratively. Therefore, we designated this study as “guided”.

Scenario & Task. In a hypothetical scenario, we asked participants to imagine they were leading a DevOps team in a company that experienced a customer data leak recently. We required them to investigate how the leaks happened and who was responsible for them. We asked participants to express their thoughts in a think-aloud setup during the study. Think-aloud included talking about their experiences in similar scenarios, questions they would ask colleagues in the imaginary company, tools they would typically use, their experience with the tools we provided, and their suspicions on what caused the data leaks. After identifying the leaks and their root causes, we asked the participants how they would resolve the found issues in their company. We also asked for general feedback regarding OLab.

Table 3: Detailed overview of experts, their background, experience of conducting studies, and their main operating system.

Alias [†]	OS	Study Background	Study Experience
Ash	Linux	Online developer studies with a focus on programming tasks.	4 years
Brooklyn	Linux	End-user studies with a focus on crowd-worker platforms.	3 years
Charlie	macOS	Both developer and end-user studies, with a focus on lab experiments and surveys.	7 years
Dakota	Windows	Developer studies with a focus on qualitative interviews.	1.5 years

[†] Gender-neutral aliases assigned alphabetical to all experts.

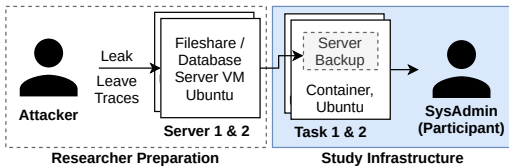


Figure 3: Overview over the guided usability study’s setup.

Figure 3 provides an overview of the task creation process. Since we aimed to test the participants’ abilities to manage security incidences in a company setting, we provided a virtual server backend within OLab. We set up two containers, a database, and a file server. We then simulated a company-internal attacker with access to the server that used social engineering to leak company-internal information. The resulting logs and system states were then backed up and included in a Docker image that provided a visual interface and tools to inspect the backups.

We added a hint regarding emails from the attacker to the admin using the second container, pointing towards the internal attackers, as an experiment condition. Hence, participants might have an easier time identifying the exact circumstances of the attack in this condition. OLab automatically assigned the condition to half of all participants. Within the study, conditions and the task order were randomized.

Two authors supervised the participants virtually during the study, took notes, and answered scenario-specific questions that participants asked. Participants were asked to screen-share the tab containing OLab, which we recorded to complement our notes (cf. replication package, Section 5).

Coding and Evaluation. Using the recorded videos and notes, two authors coded participants’ free-text responses in an “iterative categorization” [26] approach. The authors focused on the advantages and disadvantages the participants reported while interacting with OLab and their general survey sentiment. We focused on these general categories because a notable amount of feedback came up while participants were working on the tasks, not as a result of individual questions in the post-survey. After assigning initial codes to all feedback, both authors reviewed the resulting coding and resolved conflicts in a consensus discussion or introduced new codes. When new codes emerged, the already coded videos were revisited and re-coded. Since both researchers coded all

participant responses with immediate conflict resolution, we refrain from reporting the inter-rater reliability (IRR) [20].

Results. In general, OLab was well-received by all nine participants, while only some minor problems occurred that were related to OLab. Seven participants (P1, P3, P5–9) explicitly mentioned that they were impressed by OLab and its workflow. From our observations, the prototype was very fluid for all participants. Four of them (P1, P5, P8, P9) mentioned low latency, e. g., “*It worked flawlessly. I was very surprised that this works so well in the browser.*” (P5). Only P7 reported minor latency issues due to a low-quality mobile 4G/LTE connection. Other positive aspects mentioned by the participants were full functionality despite the use of ad-blockers (P5), the internet access with the possibility to install arbitrary additional software (P5), and that it works with non-German keyboard layouts (P7). Additionally, participants liked the visual appearance. To cite P1: “*The tools we are working with are modern, fast, looking good, I like that very much.*” (P1).

The most common limitation, mentioned by seven participants (P2, P3, P5–P9), were differences between the study’s infrastructure and the users’ typical environment. For example, as the environment in OLab cannot be customized for every user, the participants might miss any custom programs they like to use. As P6 put it: “*So I have some standard suite of programs that I have installed [. . .]. Well, you cannot take that for given. That would be [. . .] nice-to-have and not absolutely necessary.*” (P6).

Besides that, two participants encountered technical limitations. P2 tried to change the keyboard layout, but this is technically impossible during the study, and can only be changed when initializing the VNC connection. In addition, P2 and P3 noted the unavailability of `chroot`; this is disabled by default for security reasons. However, in researcher-supervised studies this could be enabled. Two participants (P1, P4) reported problems that were not related to OLab.

We queried participants on how they would solve the tasks in their everyday setup, i. e., not in a study within OLab. P2, P3, P4, and P8 reported differences that were not related to OLab. P5 explicitly mentioned that he would do the tasks as done in the study. The other four participants (P1, P6, P7, P9) highlighted that they would incorporate some form of social interaction during the tasks in a real-world scenario, e. g., contacting and talking to colleagues. We consider this to

be out of scope for OLab, as it is impossible to simulate this social interaction in a software tool.

We asked all participants for additional features they would appreciate. They mentioned missing tools that we could set up in future DevOps studies. P3, however, proposed that OLab should show the correct solution for self-evaluation after completing a task. We consider implementing this as an optional feature for future studies. The qualitative coding results can be found in the appendix (cf. Table 5).

4.3 Comparison Experiment

In this comparison experiment, we compare a study setup using OLab with a more conventional browser study regarding feasibility and usability. We based this experiment on the browser-based study setup of Acar et al., which consists of a developer study with two programming tasks that test cryptographic APIs and their documentation for usability [1]. This setup provides a good fit for a virtual study environment and a suitable starting point for a first unsupervised study with the OLab prototype.

Study Setup. We started with a Ubuntu 20.04 Docker container similar to the previous study setup. In that container, we installed Python including PyCryptodome [34] and the IDE PyCharm [33]. In PyCharm, we set up a Python project consisting of dependencies, a virtual environment, and a skeleton containing pre-written function names and comments with precise task descriptions. The original study relied on a browser-based approach using Jupyter Notebooks [14], likely due to the limitation that a fully virtualized setup containing an IDE was not available.

We decided to have each participant perform one task in a more conventional download setting for the comparison. We provided a website with the same structure, text, and study flow as in OLab. However, instead of redirecting to the virtualized environment, we provided them with a page to download the PyCharm project and upload their solution after completing the task on their computer.

We piloted the study internally and with students recruited in a snowball sample to evaluate task description clarity.

To ensure fair compensation and comparable internal validity, we instructed all participants to stop after at most 60 minutes per task and use PyCharm as a common development environment for the download condition. In the OLab prototype, we built the same setup based on an Ubuntu container image. It includes PyCharm with dependencies set up and the Python file containing the task opened in the IDE. Additionally, Chromium starts with the crypto API's documentation opened in a new tab. OLab collected the browser history and source code of the PyCharm project for our evaluation. An overview of the study setup can be found in Figure 4.

Task Setup. In our scenario, the developers had to implement (1) secure communication using an asymmetric encryption scheme of their choice and (2) encrypted storage using a

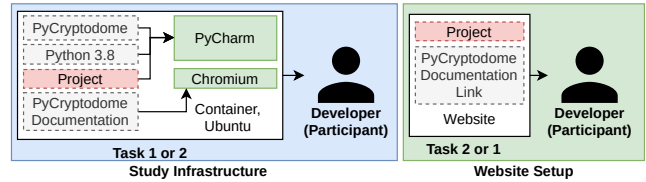


Figure 4: Overview over the comparison study's setup.

symmetric scheme. We required participants to implement this using the PyCryptodome library, which is a fork of the original PyCrypto library used by Acar et al. [3]. To combat the effects of learnability, we randomized which task our participants performed in which environment (download and OLab). We also randomized the order of study environments, i. e., half of the participants had to complete the download task first and use the OLab prototype second, with the other half completing those vice versa.

After each task, participants filled a short survey regarding feedback about the study environment and the cryptographic task. As mentioned in the limitations, we tried to differentiate between the environment and the task through diagrams and descriptions. The survey collected the *System Usability Scale* (SUS) score by Brooke [6] and the *Net Promoter Score* (NPS), a more industrially used usability score for product recommendation rates by Reichheld [36]. We decided to collect a self-assessment on security and functionality in line with the original study, but exclude other factors to prevent participant exhaustion.

Evaluation. To rate results for their security and functionality, we used an open-coding approach with two researchers to review the source code that participants submitted in OLab and the download environment. First, two authors executed the code to determine functionality. Then they rated all submissions for security, grading factors like usage of weak algorithms and insecure password generation. Finally, the coders discussed individual ratings and resolved conflicts to arrive at a complete security rating.

Since this study was unsupervised, we could not collect the degree of qualitative feedback obtained in the previous two studies. To alleviate that, we contacted participants individually after the study and asked a few follow-up questions regarding the issues and differences. For these responses, we used an “iterative categorization” [26] approach; two researchers classified the types of issues, advantages, and disadvantages participants reported on. As both researchers coded all source code and all participant responses and immediately resolved conflicts, we refrain from reporting the inter-rater reliability (IRR) [20].

Task Results. Overall, 24 out of 32 solutions were functional and secure according to our coding. Out of those, 9 of 16 solutions were secure when worked on in OLab, while 15 out of 16 were securely provided via our download environment.

We noticed that this difference comes from the second task (symmetric cryptography). The timings indicate that participants spend 60 minutes on that task using OLab. We asked the participants why they did not complete the second task. The participants who responded found it more complicated than the first task and noted that the clearly shown “skip”-button provided in OLab’s interface (cf. Figure 7c) in combination with the reminder to only spend 60 minutes on this task lead to submitting earlier when using OLab. These factors were not as present in the download condition, since there was no “skip”-button. Furthermore, participants might not see the time limit when using the website because they opened the documentation in a new tab on the same window. These responses indicated that the second task was infeasible within the provided time for students having no experience in cryptography. We think the qualitative results produced by this task are still valuable for the evaluation, as they indicate higher compliance with the study protocol of this unsupervised study when participants used the OLab. Studies within OLab need to set fair requirements and cannot rely on participants ignoring time constraints or the study conditions.

We also asked which editor participants used in both conditions in the survey, which revealed another compliance difference. All participants used PyCharm within OLab – likely because it was already set up and automatically started. However, 5 used other editors (4 VSCode, 1 VIM) in the download condition. We explicitly instructed participants to use PyCharm with our PyCharm project. Therefore, this difference represents a threat to internal validity not present in OLab. While this demonstrates the more challenging enforcement of requirements in download studies, researchers may offer multiple editors in OLab to accommodate for preferred software. However, this would increase the required time for study setup and piloting.

Usability Rating. We asked participants to rate each environment using the SUS score and the NPS. Regarding the SUS score, we had participants rate the environment after completing each task. The environment using OLab received an average SUS score of 80.0, which corresponds to a Grade A– according to Sauro and Lewis [40], with the download environment receiving a SUS score of 78.125, classified as B+. While the ratings are limited to the explanatory power of the study, they indicate at least comparable usability of OLab and the download environment. The NPS for OLab did not result in more promoters, but is equal to the download environment. We include an overview of both scores in Table 4.

This is further reflected in the participants’ preference for the environments. 9 preferred OLab while 7 preferred the download environment. As reasons for preferring the download study, three participants mentioned bandwidth limitations that lead to unresponsive or unstable experiences with OLab. While we found that OLab runs fine with typical desktop bandwidths (starting at around 0.5 Mbit/s), we assume higher

Table 4: SUS and NPS scores for both OLab and the download condition.

Score	OLab	Download Condition
SUS Score (mean)	80.0 (A-)	78.125 (B+)
NPS Promoters	10	10
NPS Passivers	3	3
NPS Detractors	3	3

round-trip times (RTTs)/pings cause a noticeably slower experience compared to a native interface.

The participants provided different reasons for preferring the download environment. To quote a participant: “*I can use my own IDE, which is adapted to my requirements. Also, I can open the documentation on a second screen, making research and reading easier.*” (P2). This preference is in line with the finding that 5 participants used a different IDE than specified for solving the task. Multiscreen support is currently impossible with OLab due to being limited to a single browser window. However, these advantages also affect the internal validity of the study results in the download study. Furthermore, they are only present for studies where the task can be downloaded to a participant’s machine, not in lab studies or studies using server environments like our previous setup.

When participants stated to prefer OLab over the download study, all 9 participants mentioned the much lower setup efforts as the main reason. A participant stated:

“On my desktop PC, I had problems importing crypto. Therefore, I had to switch to my laptop, on which working was much harder. This resulted in a lot of time spent on a problem that I didn’t have in the virtual desktop environment. In this environment, everything was prepared and I could immediately start working. There was also less distraction by open tabs or pop-up messages.” — P3

The virtualized environment within OLab can lower entry barriers for participants and reduce the time participants spend on tasks while still providing them with a fully-featured development environment that reflects their actual environment, even if customization might be missing.

To compare the timings for both studies, we asked participants about the perceived time spent on preparation, from 0 (very low) to 6 (very high). We found that people rated OLab 0.31 on average, indicating a lower setup time compared to the download environment that participants rated 2.13 on average. This confirms the suspected advantage in participants’ preparation time for studies using OLab.

5 Discussion

Below, we discuss our results in the context of the evaluation goals we presented in Section 4 and discuss how the study

results address them. We also elaborate on future directions for virtual study environments we derive from our results and how we plan to implement them.

EG1: Usability. After implementing the feedback we collected during the cognitive walkthrough, we could improve the usability of the OLab environment.

We found the OLab prototype to be easy to use for all participants (cf. the SUS scores in Section 4.3), with a low entry barrier, and safe to use in all scenarios we provided since it automatically stored participant results without storing results manually. The lower preparation time through pre-setup dependencies that participants reported in our comparison study demonstrated how this approach could be more efficient than conventional approaches. In our comparison study, participants were willing to spend more time with the tasks in their own environment, leading to more complete solutions. We believe this can be addressed through smaller tasks or more straightforward instructions.

In summary, we think that our approach can indeed fulfill the high accessibility requirements that we identified in Section 3.1.

EG2: Perception. Even when encountering minor latency issues or unknown setups in our study environments, participants remarked on the smooth study procedure possible through OLab. Participants also mentioned the low entry barrier through the provided tooling and setup as an advantage. In addition, the setup allowed us to test an unconventional setup in the form of servers that participants had to analyze for security issues. In supervised studies structured like interviews or remote think-alouds like our second study, we can even allow participants to use root access on the machines and install their own applications to complement the setups we provide them.

EG3: Limitations. We also encountered a few limitations, mostly related to security. One of these is the ability to use features like `chroot`, `KVM`, and `systemctl` that require privilege escalation beyond what is considered safe in a container. These can be ignored to some extent in supervised studies, where a researcher can ensure that participants do not abuse permissions on the container and therefore can declare the containers to be privileged. However, this poses a security risk for the entire infrastructure, including other participants and the host systems, when done without supervision.

Finally, latency is a significant limitation of the environment, and participants with a high connection latency reported difficulties using the OLab environment.

EG4: Comparison. From our previous findings, we conclude that in comparison to more conventional setups, the OLab environment provides the option to enforce higher internal validity at the cost of customization for participants. In our comparison study, we also found that the time spent on our OLab was lower on average. We assume that when providing participants the time to customize their setup in the OLab, this

advantage will vanish. In general, providing participants with a fully working setup in our OLab environment will always be faster than download tasks or tasks requiring setup time beyond reading the task description. We hope to capitalize on this advantage to conduct new types of studies that were previously hard or even infeasible to conduct online.

Replication. To allow for better replication of our work, we make the following items available as part of a replication package [13]: We provide the study protocols for the cognitive walkthrough, the guided and the comparison study including the study scenarios, the tasks descriptions, between-task surveys, and end surveys.⁴

Future Work. In future work, we plan to improve the usability of the researcher's web interface to illustrate the current state of a study, and to manage participants and study instances.

We plan to evaluate OLab in multiple large-scale studies, test more edge cases, and improve flexibility. Furthermore, support for complex features like interaction between participants or with researchers can expand the scope of possible studies for OLab.

6 Conclusion

In conclusion, we identified common requirements for lab-like studies with SIWs. We designed, implemented and evaluated the OLab environment as a novel approach to conduct lab-like studies online, and found that:

1. OLab can provide high usability for participants in online studies while enabling complex study setups such as programming and server administration studies, as evaluated through our expert walkthrough (cf. Section 4.1) and through the SUS scores (cf. Sections 4.2 & 4.3).
2. OLab handles typical research tasks like data and consent form collection and study parameters like task order, conditions, and the inclusion of external questionnaires, offering a flexible setup for complex studies to researchers (cf. Section 3.3)
3. OLab provides higher internal validity than approaches that involve external working environments, both regarding task compliance and regarding tools and environmental variables used (cf. Section 4.3).

Based on our results, we consider OLab to be a highly functional prototype that we plan to expand on for future real-world studies. Although it is not yet fit for a general release, we formally invite interested researchers to contact us regarding the collaboration and extension of OLab.

⁴The replication package is also available via this paper's accompanying website: <https://publications.teamusec.de/2022-soups-olab/>.

References

- [1] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. “Comparing the Usability of Cryptographic APIs”. In: *Proc. 38th IEEE Symposium on Security and Privacy (SP’17)*. IEEE, 2017.
- [2] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. “You Get Where You’re Looking For: The Impact of Information Sources on Code Security”. In: *Proc. 37th IEEE Symposium on Security and Privacy (SP’16)*. IEEE, 2016.
- [3] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. “You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users”. In: *Proc. 2016 IEEE Secure Development Conference (SecDev’16)*. IEEE, 2016.
- [4] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. “Security Developer Studies with GitHub Users: Exploring a Convenience Sample”. In: *Proc. 13th Symposium on Usable Privacy and Security (SOUPS’17)*. USENIX Association, 2017.
- [5] Muharrem Aksu, Enes Altuncu, and Kemal Bicakci. “A First Look at the Usability of OpenVAS Vulnerability Scanner”. In: *Proc. Workshop on Usable Security (USEC’19)*. The Internet Society, 2019.
- [6] John Brooke. “SUS: a retrospective”. In: *Journal of Usability Studies* 8.2 (2013), pp. 29–40.
- [7] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. “Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a Password-Storage Study with Freelancers”. In: *Proc. 16th Symposium on Usable Privacy and Security (SOUPS’20)*. USENIX Association, 2020.
- [8] *Docker images of Windows*. https://hub.docker.com/_/microsoft-windows.
- [9] Felix Fischer, Yannick Stachelscheid, and Jens Grossklags. “The Effect of Google Search on Software Security: Unobtrusive Security Interventions via Content Re-Ranking”. In: *Proc. 28th ACM Conference on Computer and Communication Security (CCS’21)*. ACM, 2021.
- [10] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. “Developers Deserve Security Warnings, Too: On the Effect of Integrated Security Advice on Cryptographic API Misuse”. In: *Proc. 14th Symposium on Usable Privacy and Security (SOUPS’18)*. USENIX Association, 2018.
- [11] Matthew Green and Matthew Smith. “Developers are Not the Enemy!: The Need for Usable Security APIs”. In: *IEEE Security & Privacy* 14.5 (2016), pp. 40–46.
- [12] Norman Hänsch, Andrea Schankin, Mykolai Protsenko, Felix Freiling, and Zinaida Benenson. “Programming Experience Might Not Help in Comprehending Obfuscated Source Code Efficiently”. In: *Proc. 14th Symposium on Usable Privacy and Security (SOUPS’18)*. USENIX Association, 2018.
- [13] Nicolas Huaman, Alexander Krause, Dominik Wermke, Jan H. Klemmer, Christian Stransky, Yasemin Acar, and Sascha Fahl. *Replication Package: “If You Can’t Get Them to the Lab: Evaluating a Virtual Study Environment with Security Information Workers”*. <https://doi.org/10.25835/spxeaic7>. 2022.
- [14] *Jupyter Notebook*. <http://jupyter.org/>. visited Nov. 2016.
- [15] Jupyter Project. *Jupyter Notebook Kernels*. <https://github.com/jupyter/jupyter/wiki/Jupyter-kernels>.
- [16] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. ““I Have No Idea What I’m Doing” - On the Usability of Deploying HTTPS”. In: *Proc. 26th Usenix Security Symposium (SEC’17)*. USENIX Association, 2017.
- [17] *Kubernetes*. <https://kubernetes.io/>.
- [18] *Experimental Kubernetes GPU support*. <https://kubernetes.io/docs/tasks/manage-gpus/scheduling-gpus/>.
- [19] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. “Designing Toxic Content Classification for a Diversity of Perspectives”. In: *Proc. 17th Symposium on Usable Privacy and Security (SOUPS’21)*. USENIX Association, 2021.
- [20] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. “Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice”. In: *ACM on Human-Computer Interaction* 3.CSCW (2019).
- [21] *minikube*. <https://minikube.sigs.k8s.io/>.
- [22] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. “On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers”. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI’20)*. ACM, 2020.

- [23] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. ““If You Want, I Can Store the Encrypted Password”: A Password-Storage Field Study with Freelance Developers”. In: *Proc. CHI Conference on Human Factors in Computing Systems (CHI’19)*. ACM, 2019.
- [24] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. “Why Do Developers Get Password Storage Wrong?: A Qualitative Usability Study”. In: *Proc. 24th ACM Conference on Computer and Communication Security (CCS’17)*. ACM, 2017.
- [25] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. “Deception Task Design in Developer Password Studies: Exploring a Student Sample”. In: *Proc. 14th Symposium on Usable Privacy and Security (SOUPS’18)*. USENIX Association, 2018.
- [26] Joanne Neale. “Iterative categorization (IC): a systematic technique for analysing qualitative data”. In: *Addiction* 111.6 (2016), pp. 1096–1106.
- [27] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. “A Stitch in Time: Supporting Android Developers in Writing Secure Code”. In: *Proc. 24th ACM Conference on Computer and Communication Security (CCS’17)*. ACM, 2017.
- [28] Jakob Nielsen. “Enhancing the Explanatory Power of Usability Heuristics”. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI’94)*. ACM, 1994.
- [29] Jakob Nielsen. “The usability engineering life cycle”. In: *Computer* 25.3 (1992), pp. 12–22.
- [30] Timothy Nosco, Jared Ziegler, Zechariah Clark, Davy Marrero, Todd Finkler, Andrew Barbarello, and W. Michael Petullo. “The Industrial Age of Hacking”. In: *Proc. 29th Usenix Security Symposium (SEC’20)*. USENIX Association, 2020.
- [31] Daniela Seabra Oliveira et al. “API Blindspots: Why Experienced Developers Write Vulnerable Code”. In: *Proc. 14th Symposium on Usable Privacy and Security (SOUPS’18)*. USENIX Association, 2018.
- [32] Stephan Plöger, Mischa Meier, and Matthew Smith. “A Qualitative Usability Evaluation of the Clang Static Analyzer and libFuzzer with CS Students and CTF Players”. In: *Proc. 17th Symposium on Usable Privacy and Security (SOUPS’21)*. USENIX Association, 2021.
- [33] *PyCharm*. <https://www.jetbrains.com/pycharm>.
- [34] *PyCryptodome*. <http://pycryptodome.readthedocs.io>.
- [35] Hirak Ray, Flynn Wolf, Ravi Kuber, and Adam J. Aviv. “Why Older Adults (Don’t) Use Password Managers”. In: *Proc. 30th Usenix Security Symposium (SEC’21)*. USENIX Association, 2021.
- [36] Frederick F. Reichheld. “The One Number You Need to Grow”. In: *Harvard Business Review Press* 81.12 (2003), pp. 46–55.
- [37] Sebastian Roth, Lea Gröber, Michael Backes, Katharina Krombholz, and Ben Stock. “12 Angry Developers - A Qualitative Study on Developers’ Struggles with CSP”. In: *Proc. 28th ACM Conference on Computer and Communication Security (CCS’21)*. ACM, 2021.
- [38] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L. Mazurek, and Piotr Mardziel. “Build It, Break It, Fix It: Contesting Secure Development”. In: *Proc. 23rd ACM Conference on Computer and Communication Security (CCS’16)*. ACM, 2016.
- [39] Iftaah Salman, Ayse Tosun Misirli, and Natalia Juristo. “Are Students Representatives of Professionals in Software Engineering Experiments?” In: *Proc. 37th IEEE/ACM International Conference on Software Engineering (ICSE’15)*. IEEE, 2015.
- [40] Jeff Sauro and James R Lewis. *Quantifying the User Experience: Practical Statistics for User Research*. Morgan Kaufmann Publishers, 2016.
- [41] Justin Smith, Lisa Nguyen Quang Do, and Emerson Murphy-Hill. “Why Can’t Johnny Fix Vulnerabilities: A Usability Evaluation of Static Analysis Tools for Security”. In: *Proc. 16th Symposium on Usable Privacy and Security (SOUPS’20)*. USENIX Association, 2020.
- [42] Springer Verlag GmbH, European Mathematical Society. *Latin Square*. http://encyclopediaofmath.org/index.php?title=Latin_square&oldid=47587. Accessed on 2022-02-18.
- [43] Christian Stransky et al. “Lessons Learned from Using an Online Platform to Conduct Large-Scale, Online Controlled Security Experiments with Software Developers”. In: *Proc. 10th USENIX Workshop on Cyber Security Experimentation and Test (CSET’17)*. USENIX Association, 2017.
- [44] Mikael Svahnberg, Aybüke Aurum, and Claes Wohlin. “Using Students as Subjects - an Empirical Evaluation”. In: *Proc. Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM’08)*. ACM.

- [45] Christian Tiefenau, Maximilian Häring, Katharina Krombholz, and Emanuel von Zezschwitz. “Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators”. In: *Proc. 16th Symposium on Usable Privacy and Security (SOUPS’20)*. USENIX Association, 2020.
- [46] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. “A Usability Evaluation of Let’s Encrypt and Certbot: Usable Security Done Right”. In: *Proc. 26th ACM Conference on Computer and Communication Security (CCS’19)*. ACM, 2019.
- [47] Harshal Tupsamudre, Monika Sahu, Kumar Vidhani, and Sachin Lodha. “Fixing the Fixes: Assessing the Solutions of SAST Tools for Securing Password Storage”. In: *Proc. 24th International Conference on Financial Cryptography and Data Security (FC’20)*. Springer, 2020.
- [48] Carl W. Turner, James R. Lewis, and Jakob Nielsen. “Determining Usability Test Sample Size”. In: *International Encyclopedia of Ergonomics and Human Factors*. 2nd ed. Vol. 3. CRC Press, 2006, pp. 3084–3088.
- [49] Daniel Votipka, Kelsey R. Fulton, James Parker, Matthew Hou, Michelle L. Mazurek, and Michael Hicks. “Understanding security mistakes developers make: Qualitative analysis from Build It, Break It, Fix It”. In: *Proc. 29th Usenix Security Symposium (SEC’20)*. USENIX Association, 2020.
- [50] Daniel Votipka, Seth Rabin, Kristopher Micinski, Jeffrey S. Foster, and Michelle L. Mazurek. “An Observational Investigation of Reverse Engineers’ Processes”. In: *Proc. 29th Usenix Security Symposium (SEC’20)*. USENIX Association, 2020.
- [51] Dominik Wermke, Nicolas Huaman, Yasemin Acar, Brad Reaves, Patrick Traynor, and Sascha Fahl. “A Large Scale Investigation of Obfuscation Use in Google Play”. In: *Proc. 34th Annual Computer Security Applications Conference (ACSAC’18)*. ACM, 2018.
- [52] K. Yakdan, S. Dechand, E. Gerhards-Padilla, and M. Smith. “Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study”. In: *Proc. 37th IEEE Symposium on Security and Privacy (SP’16)*. IEEE, 2016.

A Technical Details of OLab

OLab’s Kubernetes cluster runs entirely self-hosted on the researchers’ servers. This setup provides maximum security and data protection for participant data – without any third party involved. For a technical overview, see Figure 8.

Depending on the number of participants, OLab supports other deployment options. For minimal setups or testing purposes, *minikube* [21] requires only a single machine. In studies that exceed the researchers’ server resources, it is possible to host and operate OLab within a Kubernetes cloud environment, e. g., Amazon Web Services (AWS).

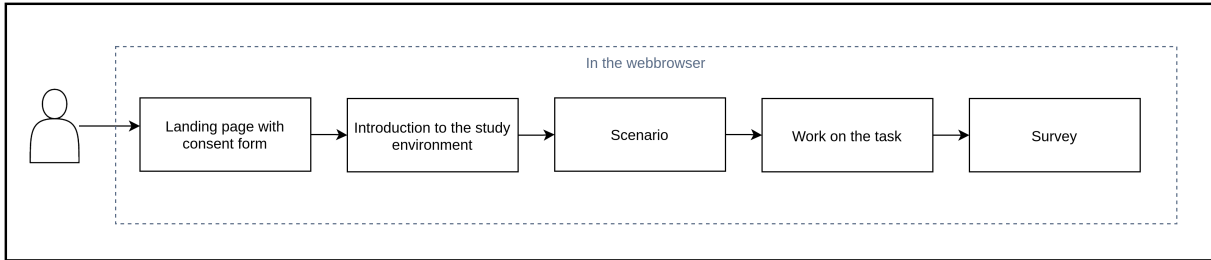


Figure 5: Comparison study (Section 4.3) setup for the condition that uses OLab.

Table 5: Qualitative coding of study results for the guided DevOps study (cf. Section 4.2).

Participants		Positive Points	Normal WE	Problems	Features
	Recruited at	OK			
		Works despite adblockers/addons			
		Not related to infrastructure			
		Low Latency			
		Visually Appealing			
		Very good			
		Internet Access, Software can be installed			
		translates us layout			
		Not related to infrastructure			
		Social Interaction			
		Similar to Infrastructure			
		Not related to infrastructure			
		Change keyboard layout			
		Not the usual approach/environment			
		No chroot available			
		some lags (LTE)			
		Not related to infrastructure			
		Show correct Solution			
P1	Admin-Forum				
P2	University				
P3	University				
P4	Reddit				
P5	Reddit				
P6	Reddit				
P7	Reddit				
P8	Reddit				
P9	Reddit				

Welcome to the study

Thank you for your interest in our study.

In this study, we aim to investigate, *****. As part of this study, we ask you to complete two tasks and fill out a questionnaire. You only need a web browser and there is no need to install any additional software.

Thank you in advance for your effort and time.

Consent Form

Project Title	<i>Example Study</i>
Principle Investigator	[REDACTED]
Student Researchers	[REDACTED]
Project Description	Example Study Description.
Risks & Benefits	The risks to your participation in this programming study are those associated with basic computer tasks, including boredom, fatigue, mild stress, or breach of confidentiality. The benefits are the learning experience from participating in a research study and a contribution to the state of scientific knowledge, and the reimbursement.
Compensation	For participation in the programming study, you will be paid [€XX]. You will only be paid if you meaningfully work on or complete all tasks, and complete the exit survey.
Confidentiality	To protect your privacy as related to this research, your study data will be pseudonymized. Identifiers like names and email addresses will be stored separately from study data. Data that will be shared with others about you will be pseudonymized. Any reports and presentations about the findings from this survey will not include your name or any other information that could identify you, except for the library/primitives you worked on. We may use pseudonymized quotes in publications and presentations. After the survey concludes, we will not retain any links between pseudonyms and identifying data like names and email addresses.
Subjects' Rights	Your participation is voluntary. You may stop participating at any time by exiting the study environment. For additional questions about this research, you may contact:

(a) Landing page with required consent form and further study information.

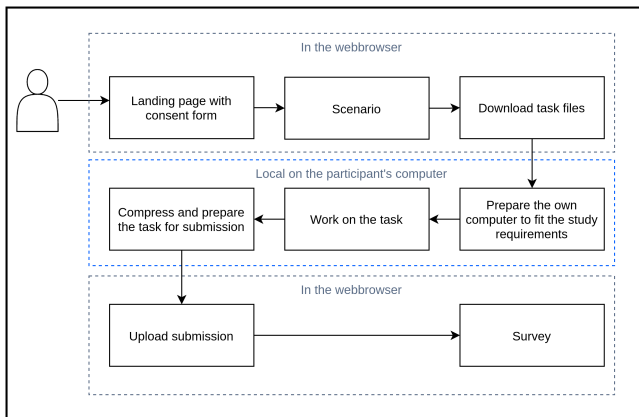


Figure 6: Comparison study (Section 4.3) setup for the download condition.

Scenario

For the study, please imagine that

Procedure of study

Overall task + image

For task processing

The study environment provides you with a full Linux operating system with a web browser and internet connection. [Study goals]

During the study, we kindly ask you to:

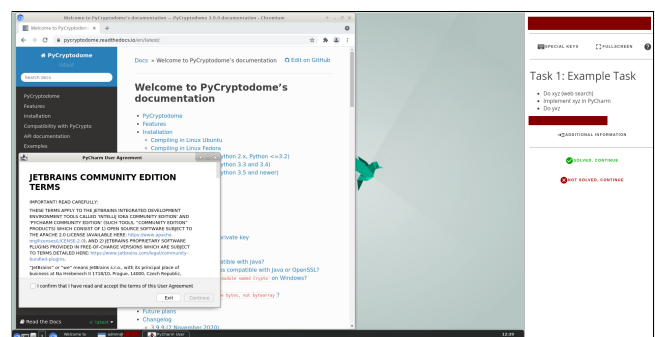
- XXX
- XXX
- XXX

1 2 3 4

Task 1: Example_Task_1 Survey Task 2: Example_Task2 Survey

START THE TASK

(b) In-between task progression status page, including survey steps.



(c) Virtual study environment running Chromium & PyCharm. The right sidebar includes task descriptions and control buttons.

Figure 7: Screenshots of the OLab prototype, during a generic programming study.

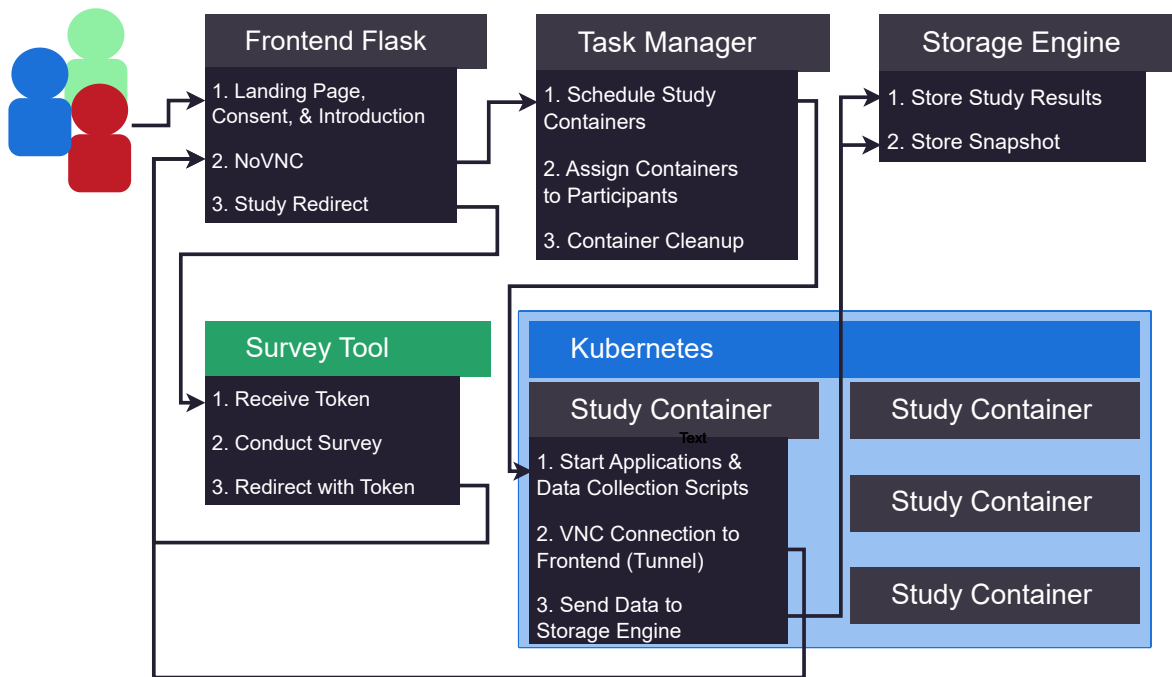


Figure 8: OLab architecture diagram.

Is it a concern or a preference? An investigation into the ability of privacy scales to capture and distinguish granular privacy constructs

Jessica Colnago
Google*

Lorrie Faith Cranor, Alessandro Acquisti
Carnegie Mellon University

Kate Hazel Stanton
University of Pittsburgh

Abstract

Privacy scales are frequently used to capture survey participants' perspectives on privacy, but their utility hangs on their ability to reliably measure constructs associated with privacy. We investigate a set of common constructs (the intended objects of measurement by privacy scales) used in privacy surveys: privacy attitude, privacy preference, privacy concern, privacy expectation, privacy decision, and privacy behavior. First, we explore expert understanding of these constructs. Next, we investigate survey participants' understanding of statements used in privacy scales aimed at measuring them. We ask a balanced sample of Prolific participants in the United States to identify the extent to which different constructs describe each of a set of 30 statements drawn from scales used commonly in the privacy literature and 39 that we developed. Our analysis reveals considerable misalignment between the constructs associated with the statements and participant understanding. Many statements used in scales or that we developed with the intention to measure constructs such as privacy concern, are seen by survey participants as describing other constructs, such as privacy preferences. We also find that no statement uniquely measured any one construct, though some more reliably track their target construct than others. Our findings constitute an epistemological problem for use of scales in the existing literature (are they capturing what we think they capture?) and a practical problem for construction of new scales (how to ensure construct validity in the face of ill-defined constructs and evolving privacy landscape?). We use methods from corpus linguistics to identify characteristics of those statements most reliably associated with their target con-

struct, and provide a set of provisional suggestions for future statement construction. Finally, we discuss the implication of our results for the privacy research community.

1 Introduction

Privacy scales are familiar instruments in privacy research [16]. These scales aim at measuring *constructs* — specific facets of participant privacy psychology, such as privacy concerns or privacy preferences — by soliciting degrees of agreement with statements believed to capture these constructs [13, 20]. A valid privacy scale can offer useful insight into public perspectives on privacy, but a scale that is not valid — that is, a scale that fails to measure its intended construct — presents a challenge for privacy research by yielding results that cannot sustain accurate generalisations and that lack predictive power [21]. Recent work has challenged the validity of existing scales [10, 18]. Here, we present evidence that problems with validity may be widespread — perhaps even intrinsic to the privacy scale as an instrument given the ill-defined and ever evolving nature of privacy — as thoroughly validated scales did not achieve conceptual clarity on the constructs they attempt to capture. We show that survey participants cannot identify *unique* constructs corresponding to statements used in scales, and that there is considerable variation in beliefs concerning which construct a statement corresponds to. There is little hope that a scale aimed at measuring, for example, *privacy concerns* can be trusted to do only that, when participants may have been understanding its constituent statements as expressing *privacy preferences*.

We investigate the following constructs, which are common in the privacy literature: *privacy attitude*, *privacy preference*, *privacy concern*, *privacy expectation*, *privacy decision*, and *privacy behavior*. Since there are no definitions of these constructs universally accepted by privacy scholars, we offered a set of definitions taken from a recent book chapter [5] to 22 privacy experts, and iteratively refined these definitions based on the experts' feedback. Next, as many privacy-related studies are performed using crowd-sourcing platforms, we

* The work was performed while Jessica Colnago was at Carnegie Mellon University.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022, August 7–9, 2022, Boston, MA, United States.

asked a sample of Prolific participants in the United States to identify the extent to which the different constructs, presented with our refined definitions, describe each of 30 statements from scales used commonly in the privacy literature. We also asked participants to perform this task for 39 statements that reflect commonly stated privacy opinions observed in qualitative privacy studies. We leveraged Prolific’s representative sample functionality to recruit a sample balanced using Census information on age, gender, and ethnicity. All studies were approved by our institution’s Internal Review Board.

Our analysis shows that many statements intended to measure certain constructs that commonly appear in the privacy literature and that are systematized in Cranor and Schaub’s framework [5] (for example, privacy concern) are, in fact, seen by survey participants as describing other constructs in the framework, such as privacy preferences.

We also found that no statement uniquely measured any construct. The results highlight the difficulty of using scales to measure privacy constructs uniquely and reliably. We observe that some statements were, however, more regularly matched with particular constructs. We use methods from corpus linguistics to identify features that these statements share and generalise over them to make provisional suggestions aimed at guiding future scale construction. Finally, we discuss the implication of our results for the privacy research community.

2 Background and related work

This paper builds on work in the privacy literature concerning privacy scales and privacy surveys, and on critical contributions that raise problems for those scales and surveys.

2.1 Privacy scales and privacy constructs

We focus on some of the most popular privacy scales: Westin’s Privacy Segmentation Index [12], Concern for Information Privacy (CFIP) [20], Global Information Privacy Concern (GIPC) [13], and Internet Users’ Information Privacy Concern (IUIPC) [13]. Some of these scales are validated—that is, carefully designed to ensure that the set of included statements consistently capture a construct. As we discuss, all of them appear to have been designed to measure *privacy concern*, as it was understood at the time of the scale’s creation. We present each scale discussed in this paper and discuss how it is used in our empirical analysis. All scales are reproduced in Figure 9 in the Appendix.

Westin’s Privacy Segmentation Index: Alan Westin created privacy indexes to track trends in privacy perspectives over time. Based on their answers, survey participants were classified into categories that “represent a continuum of privacy concern” [12]. To the best of our knowledge, these indexes did not form a validated scale. In particular, Westin’s Privacy Segmentation Index captured participants’ level of

agreement on a 4-point scale to three statements. Participants who agreed with the first statement and disagreed with the second and third statements were classified as *privacy fundamentalists*. Participants who presented the opposite pattern were classified as *privacy unconcerned*. Finally, all other participants were classified as *privacy pragmatists*.

Global Information Privacy Concern: The Global Information Privacy Concern (GIPC) scale was first mentioned by Malhotra et al. in 2004 [13] and considers six statements measured on a 7-point scale. An extensive literature search for mentions of GIPC did not yield results prior to 2004. Thus, we do not have information on how these statements were selected and whether this scale has been validated. In this paper, we consider that GIPC measures concern, given the presence of this construct in the scales’ name.

Concern For Information Privacy: In 1996, Smith et al. proposed the Concern For Information Privacy (CFIP) scale. This served as a first validated instrument for measuring concerns about organizational information privacy practices, but the paper does not provide a definition of concern. This scale followed a rigorous development methodology that included the generation of sample items and verification of content validity, followed by exploratory and confirmatory factor analysis, and assessments of internal validity, reliability, and generalizability. The CFIP scale includes 15 statements and four sub-scales that measure dimensions of individuals’ concerns about organizational privacy practices: collection, errors, unauthorized use, and improper access. Participants report their level of agreement with each of the above statements on a 7-point scale, which are then be converted into means for the sub-scales, as well as the overall scale [20].

Internet Users’ Information Privacy Concern: Malhotra et al. proposed the Internet Users’ Information Privacy Concern (IUIPC) scale “[t]o reflect Internet users’ concerns about information privacy” with a focus on “individuals’ perceptions of fairness/justice in the context of information privacy.” IUIPC was adapted from CFIP and included new items and dimensions. The authors proposed it to provide a theoretical framework on the nature of information privacy concerns for Internet users. As with CFIP, IUIPC was developed following a strict scale development methodology and results of a thorough validation process are presented in the paper. The IUIPC scale is composed of 10 statements and 3 dimensions: control, awareness, and collection (taken from CFIP). Participants report their level of agreement with each statement on a 7-point scale, and the means are calculated for each dimension [13].

2.2 Constructs and framework

We focus on a subset of constructs that have been identified to be of interest in the privacy literature: attitude, preference, con-

cern, expectation, decision, and behavior. These constructs are of long standing interest in the social sciences more broadly, where their importance and inter-relationships have been explored for decades [7]. Somewhat naturally, given such long-standing interest, we see variations in how these constructs are used across different fields, and even within the privacy literature [8]: different terms have been used to refer to the same underlying phenomenon and the same term has been used to describe slightly different phenomena over time. For example, psychologists use “worry” to refer to a “state of mental distress or agitation due to concern about an impending or anticipated event, threat, or danger” [22], while privacy scholars frequently use “concern” to refer to this state. As for “preference,” the term has been used to refer to different phenomena across psychology, social sciences, and economics [11].

We leverage the conceptual framework proposed by Cranor and Schaub [5] as a seed for our construct definitions. This framework covers privacy attitude, privacy preference, privacy concern, privacy expectation, privacy decision, and privacy behavior. We used this framework due to its simplicity and coverage of central constructs used in privacy research.

2.3 Lexical issues

Constructs are specified by terms that bear rich lexical relations that complicate unique construct measurement. As noted above (see Section 2.1), and in alignment with Smith et al. [19], the privacy scales being evaluated in this paper seem to have been meant to capture *privacy concerns*. However, *concern* is a subcategory (hyponym) of a broader class, *attitude* (hypernym) (cf. [6]). As such, any statement that falls under a subcategory (e.g. *privacy concern*) may also fall under the supercategory (*privacy attitude*), meaning that scales that claim to measure any subcategory may also be judged to measure the supercategory.

This inter-related nature of privacy constructs could explain the lack of construct validity found by previous work when investigating IUIPC [10, 18]. In particular, Gross notes that the sub-scales Control and Awareness had “unsatisfactory local fit for two items . . . calling the unidimensionality of these sub-scales into question” [10]. Our work builds on this past work, showing that statements used in privacy scales (as well as new statements we developed reflecting commonly stated privacy opinions) measure multiple privacy constructs, and frequently not the one originally intended.

Ambiguous or low-context statements, featured in many scales, also present problems. For example, a key difference between a *concern* and a *preference* is the affective valence of the attitude: concerns are negatively valenced whilst preferences are positively valenced. When unambiguous information about the intended affective valence is not available from the statement, this information must be supplemented by participants to determine whether the statement expresses a privacy preference or a privacy concern. For example, the

statement “To me it is the most important thing to keep my privacy intact from online companies” (GIPC) may be seen as describing *privacy concern* by someone who believes corporate data collection is harmful and as describing a *privacy preference* by someone who believes corporate data collection is benign or beneficial. This supplementation may be done differently depending on individuals’ priors [14].

Previous work has examined a related issue by exploring the framing of privacy-related questions [3, 10]. Findings indicate that use of priming words, such as privacy or autonomy, can lead to skewed results [10]. Furthermore, it was found that surveys introduced with privacy-related warnings elicited results significantly different from those without privacy warnings [3].

A further source of complication is that statements may possess features that are connected to multiple constructs—a statement may refer both to a behavior (and so judged to measure behavior) and to negative affect (and so judged to measure privacy concern). As a result of overlapping linguistic and conceptual structures in both constructs and statements, privacy scales may be by nature unsuitable for unique construct measurement.

3 Construct definitions study

We conducted two studies to investigate the extent to which various statements regarding privacy—many of which are employed in popular privacy concern scales—are described by distinct privacy constructs: a construct definitions study with experts (discussed in this section); and a statement classification study with a balanced sample of US respondents provided by the Prolific platform (discussed in Section 4). The construct definitions study leveraged experts’ opinions to define an initial set of privacy constructs and associated definitions, which we then refined through an iterative process and later provided to crowd worker participants in the statement classification study to reduce variation in interpretation of the constructs.

3.1 Methodology

In the construct definitions study, we iteratively vetted privacy constructs and definitions with privacy experts with the goal of defining a set of constructs and definitions to be used with Prolific participants in the statement classification study.

To navigate the observed variation in the literature, we first established working definitions for each construct. We started from a framework of privacy constructs and associated definitions proposed by Cranor and Schaub [5] that distinguishes privacy attitude, privacy preference, privacy concern, privacy expectation, privacy decision, and privacy behavior (Table 1). As the definitions associated with this framework had not been empirically tested, we engaged a set of privacy experts in a

Construct	Initial framework	Revised framework	Final framework
Privacy attitude	The data subject's predisposition regarding privacy, usually expressed in broad and non-actionable terms.	An individual's predisposition towards privacy (and technology) which influences their stance regarding different privacy-related situations.	An individual's predisposition towards privacy which influences their stance regarding different privacy-related situations.
Privacy preferences	What the data subject prefers to happen.	(Same as final)	An individual's preferred outcome for a specific privacy-related situation.
Privacy concern	What the data subject fears might happen.	(Same as final)	An expression of worry towards a specific privacy-related situation.
Privacy expectation	What the data subject thinks will happen.	An expression of what one views as the likely outcome of a specific privacy-related situation or behavior from the other parties involved.	An expression of what one views as the likely specific privacy-related outcome of a situation or behavior from the other parties involved.
Privacy decision	What the data subject decides or intends to do.	What an individual chooses to do in a specific situation given the resources available to support their decision making process.	What an individual chooses to do in a specific privacy-related situation among available options.
Privacy behavior	What the data subject does.	(Same as final)	What an individual actually does or has done in an attempt to achieve the level of privacy that they prefer.

Table 1: Evolution of the framework from its original format to the final version based on experts' feedback. Note that Cranor and Schaub's definition for privacy attitude was "The data subjects' general predisposition regarding privacy." We start with a modified version that the authors thought improved clarity.

process of refinement of the initial framework, so that the constructs and definitions would be generally well aligned with the experts' understanding. The refinement process took place until the feedback converged into agreement—this happened within two rounds.¹

In the first round, we presented 22 experts (described in Section 3.2) with a survey that introduced the constructs and the initial set of associated definitions. We asked the experts whether they agreed with the definitions, and offered an open-ended response field to elaborate on points of disagreement. We also presented experts with statements from privacy scales and asked them which constructs best applied. Based on the first-round results, we generated a revised framework of constructs and associated definitions.

In the second round, we presented the revised framework to the 19 experts who had agreed to be contacted again. We received nine responses, which led to several small changes in the definitions. The initial, revised, and final iteration of the framework are shown in Table 1. In Section 3.3 we present the comments that experts provided in both the first and the second rounds of Study 1.

¹The results of the statement classification study are robust to both providing and not providing participants with these definitions. See Section 4.

3.2 Expert selection and demographics

We selected privacy experts who worked in the areas of usable privacy, privacy law, or privacy policy; had authored at least five published papers in one of these areas in the past 10 years; and were located in the US.²

Two members of our research team generated an initial list of experts. We identified additional experts from the authors of papers retrieved with a search of the ACM Digital Library³ and equivalent queries using Web of Science. After compiling a list of 68 potential experts, we verified the requirements above through online publication lists. Nine did not fit the required criteria and we could not validate nine others. Seven were not located in the US. We contacted the remaining 43 experts via email. We obtained complete responses from 22 experts in round 1 and 9 experts in round 2.

In the first round, half of the experts self-identified as male and half as female. On average, the experts had 16 years of experience with privacy research (sd: 5.9 years). In the second round, three self-identified as male, and six as female. On average, the experts had 16 years of experience with privacy research (sd: 8.9 years). In the first round, 11 experts

²This was a requirement of our Internal Review Board due to concerns about the General Data Protection Regulation that they had not resolved at the time of our study.

³Search Queries: [All: "privacy policy"] OR [All: "privacy law"] OR [All: "usable privacy"] AND [Publication Date: (01/01/2010 TO 01/31/2020)]; and analogous searches with only one research area at a time

described their background as “Social Sciences,” nine “Computer Science,” five “Law,” and three “Other.” The majority of the experts reported working in academia, with two citing industry experience, and one mentioning policy and government. Only one expert listed industry and only one expert listed policy as their main area of focus. The second round had a mix of law, computer science, and social science in a similar proportion as the first round.

3.3 Expert feedback on definitions

The first round of feedback highlighted experts’ concerns over lack of clarity of some definitions. Some comments were targeted at the vagueness of the initial set of definitions: “The description lacks an indication of what the preference is about.” Others addressed specific word choices: “I am not sure that concern = fear. One can have legitimate concerns without being fearful.” Some experts suggested that we better tie the definitions to privacy: “The definition would need to be completed by indicating ‘what the data subject does with respect to privacy.’” This initial round of feedback led to significant changes to the initial set of definitions, as can be seen in Table 1. The revised set was presented again to experts in the second round of the study.

The second round of feedback was narrower and pointed, leading to the final framework presented in Table 1. Below, we summarize the feedback we received in the second round.

Privacy attitude: One expert pointed out that a parenthetical in “predisposition towards privacy (and technology) . . .” could be confusing. We agreed and removed the parenthetical. Another expert asked whether the definition only applied to attitudes about one’s self, or if it also applies for attitudes towards others (for example, “I think my kids should be more careful sharing information on Facebook”). We decided that the existing definition appropriately included both and did not revise further.

Privacy preference: In the second round we did not receive any feedback for this construct.

Privacy concern: One expert highlighted that there may be a fundamental difficulty with measuring concern, as concern is a combination of expectation and trust. One may not be concerned about an otherwise concerning issue because they trust the parties involved. While we agree, as our focus is not on sources of concern, we did not revise the definition.

Privacy expectation: One expert noted that the phrasing of the definition suggests that all outcomes of a privacy-related situation are privacy expectations, even if some are not related to privacy. We reworded so “privacy-related” modifies “outcome” rather than “situation.”

Privacy decision: An expert pointed out that our definition did not mention privacy. We revised our definition to refer to decisions in “privacy-related” situations and added that a decision can only be made from a set of available options.

Privacy behavior: This construct received the strongest negative review, with one expert stating:

This definition I disagree most with – I think privacy behaviors are often inconsistent with what people would prefer and many behaviors are in conflict with the level of privacy that people prefer. I think privacy behavior is what an individual does that has an impact on their privacy, regardless of whether it’s positive or negative or consistent with their attitudes, preferences, or concerns.

While we agree that privacy behaviors may not always achieve a person’s desired outcome and may even be counterproductive, we think it is important to limit this definition to behaviors that were intended to achieve a privacy-related outcome. For example, while closing curtains is a behavior that can increase privacy, people also close curtains for other reasons, such as reducing screen glare or darkening a room. For this reason, it is important that behavior-related statements specify the goal of said behavior.

4 Statement classification study

The statement classification study used data from online crowd worker participants—a typical population of focus for measuring privacy perspectives—to assess which constructs and definitions defined in the construct definitions study described a set of 69 privacy statements. We took 30 statements from existing privacy scales and developed 39 additional statements. For each of the new statements we developed, we classified it according to the authors’ expectations as to the construct with which it would best align.

We presented participants with the following prompt: “Imagine that you are talking to a friend, and your friend says the following sentence.” This was followed by a randomly selected statement. We asked participants to rate how well each of the constructs described what their friend was saying in that sentence. Participants rated each construct on a 5-point scale, from “Does not describe at all” (1) to “Describes very well” (5). Each participant was presented with a random selection of seven statements out of the 69 available. Each statement was rated by approximately 40 participants.

Since, in pilot studies, we did not identify differences in how participants classified statements between the group that was shown the constructs with the definitions and the one that only saw the constructs, and given our desire to normalize participants’ interpretations of the constructs to the maximum possible extent, we showed all participants the constructs and associated definition for each classification task.

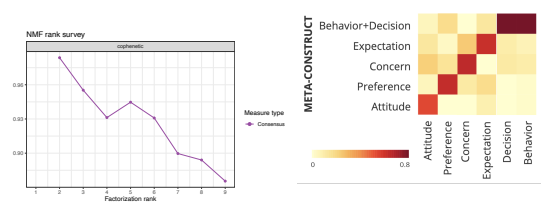


Figure 1: Left: Cophenetic correlation coefficient graph (ranks 2–10) showing a continuous drop for ranks >5. Right: NMF basis results showing the composition of the meta-constructs. Values were normalized to range from 0 to 1.

4.1 Participant Demographics

We recruited 400 participants from the Prolific platform. Prolific’s representative US sample provides a balanced sample in terms of gender, age, and ethnicity based on US Census data. Fifty percent in our sample self-identified as female, with one participant choosing non-binary. The mean age was 46.4 years, with a standard deviation of 16.3 years. When asked about their ethnicity, 71% of our participants self-identified as white, 14% as Black or African American, 8% as Asian, 6% as Other (which could encompass mixed race), and one participant self-identified as American Indian or Alaska Native. Furthermore, 7% of our participants self-identified as Hispanic or Latinx. Lastly, 16.5% of our participants reported working in or studying a technology related area.

4.2 Analyses Approach

We first binned participant responses for every statement into “high” (4 or 5) and “low” (1, 2 or 3) scores. To check the robustness of this approach, we compared results when binning the neutral option (3) with both the high and low categories. The differences observed did not impact the findings we present.

For each statement we determined whether there was a “primary construct” as follows. We identified the two constructs with the highest count of high scores (from approximately 40 responses) and compared their counts of high and low scores. We used Chi-square tests and Cramer’s V to determine whether the top construct was statistically different from the second highest one. The distributions were considered distinct if the p-value from the Chi-square was smaller than 0.05; otherwise, they were considered similar. For distinct distributions we report the effect size using Cramer’s V. The results are presented in Section 4.3.1.

The results of this analysis indicated that the majority of statements were not described by a single primary construct, and that those that were often had small effect sizes. Therefore, we turned next to an analysis approach that did not rely on distinct constructs and could provide insights into how the constructs related to one another. We used Non-negative Matrix Factorization (NMF) which automatically “extract[s]

sparse and easily interpretable factors” [9]. This method provides a better understanding on how the constructs relate to one another and how they relate to the statements. We ran the algorithm on a matrix composed of the six constructs and 69 statements. Each cell corresponded to the count of participants who selected a “high” level of agreement (Strongly agree (5) or agree (4)) for each construct statement pair.

Similar to cluster analysis, the first step in NMF is to identify how many ranks, similar to groups and clusters, will lead to stable and descriptive results. While there are many ways of selecting the rank [9], in this work we do so by examining the cophenetic correlation coefficient graph (Figure 1, left) obtained from the consensus matrix—the average connectivity matrix over many clustering runs [4].⁴

Following the rule of “select[ing] values of k where the magnitude of the cophenetic correlation coefficient begins to fall” [4], we selected five ranks, for which the algorithm outputs five basis components—we refer to these components as “meta-constructs.” These meta-constructs are a composition of the initial constructs and, as we can see in Figure 1 (right), they roughly break along the lines of the constructs, with privacy behavior and privacy decision being grouped in a single meta-construct.

By using the consensus output obtained from running the algorithm 100 times, the NMF algorithm associates each statement with a meta-construct. Thus we produced five groups of statements corresponding to our meta-constructs. We present our results in Section 4.3.2.

4.3 Statement classification results

We present our classification results based on primary constructs and meta-constructs, as well as broken down by scale.

4.3.1 Primary constructs

We see that only 33 of the statements (48%) had the top construct statistically different from the second highest one. This means that there was a primary construct that survey participants perceived as describing individual statements for roughly half of the statements. Even among those, none had a large effect size: 23 had a low effect size ([0.1, 0.3]) and ten had medium effect sizes ([0.3, 0.5]). For the rest, no primary construct was identified. The right side of Figures 2 through 6 show the percentage of high selections in green, highlighting those that had a primary construct with a dotted box.

4.3.2 Meta-constructs

Our findings for primary constructs seem to indicate a lack of independence between the constructs and definitions that we used. Therefore, we used NMF to identify composite

⁴The consensus matrix was obtained through 100 iterations of the algorithm.

Construct	CFIP	GIPC	IUIPC	Westin	New
Attitude	0	2	2	1	10
Preference	0	1	1	0	4
Concern	2	1	0	0	5
Expectation	0	0	0	0	1
Decision	0	0	0	0	3
Behavior	0	0	0	0	0

Table 2: Breakdown of the number of statements with primary constructs per source. For IUIPC we only consider the six statements unique to IUIPC, those related to Control and Awareness.

Construct	CFIP	GIPC	IUIPC	Westin	Self-gen
Attitude	0	2	1	0	8
Preference	6	1	3	0	6
Concern	6	2	1	1	6
Expectation	2	0	1	2	6
D & B	1	1	0	0	13

Table 3: Breakdown of the number of statements each meta-constructs per source. For IUIPC we only consider the six statements unique to IUIPC.

constructs. The NMF results show the weighted function of the identified meta-constructs that describes each statement (see left heatmap on Figures 2 through 6).

4.3.3 Results by scale

We present our results with statements grouped according to the scale in which they are used. Tables 2 and 3 summarize the breakdown of primary constructs and meta-constructs by source. Figures 2 through 6 also include the scale for each statement and the construct to which the scale authors expected or intended it to align.

CFIP: This scale was intended to capture the construct *privacy concern*. Out of the 15 statements that compose CFIP, we found that only six had *privacy concern* as their meta-construct (Figure 4), while six others had *privacy preference* as their meta-construct. Of note, “Companies should have better procedures to correct errors in personal information” and “Companies should take more steps to make sure that the personal information in their files is accurate” were associated with the meta-construct *privacy expectation*, though Figure 5 shows that none of the meta-constructs seem to be dominant.

GIPC: While we could not establish it with certainty, we consider that the underlying construct intended to be measured by GIPC’s statements is *privacy concern*. We

see a similar pattern to CFIP, where GIPC’s statements were infrequently associated with *privacy concern* as their meta-construct. Two out of the six GIPC statements had *privacy concern* as their meta-construct. Interestingly, the statements “I believe other people are too much concerned with online privacy issues” and “Compared with other subjects on my mind, personal privacy is very important” had *privacy attitude* as their meta-construct.

IUIPC: We consider that IUIPC had the intention to capture the construct *privacy concern*. For the six statements related to awareness and control, which were created for IUIPC, we see that *privacy concern* was the meta-construct for only one statement: “I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.” Instead, three statements had *privacy preference* as their meta-construct. “Consumer control of personal information lies at the heart of consumer privacy” had *privacy attitude* as its meta-construct while “It is very important to me that I am aware and knowledgeable about how my personal information will be used” had *privacy expectation*.

Westin: We consider that Westin’s Privacy Segmentation Index statements were created with the intent to measure *privacy concern*. However, what we found is a combination of concern and expectation. The statement “Consumers have lost all control over how personal information is collected and used by companies” had *privacy concern* as its meta-construct, though attitude was more frequently selected. The statements “Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today” and “Most businesses handle the personal information they collect about consumers in a proper and confidential way” had *privacy expectation* as their meta-construct.

Generated statements: We also examined the statements that we generated for the study, considering our specific constructs and definitions. Our expected construct matched the meta-construct predominantly selected as describing the statement for about 85% of the statements. As we can see in the heatmap figures, the statements that did not match were:

- I am not satisfied with my current level of privacy (Expected: attitude; classification: concern)
- I don’t care about privacy as long as I can use the service (Expected: preference; classification: behavior and decision)
- I don’t think there’s anything to worry about privacy (Expected: concern; classification: attitude)
- I will be able to achieve the level of privacy that I want to have (Expected: expectation; classification: preference)



Figure 2: Heatmap displaying the NMF coefficient results showing the composition of each statement based on the meta-constructs (left) and the percentages of high scores for each construct/statement pair (right) for statements under the “attitude” meta-construct. The primary construct identified is highlighted by a dotted box.

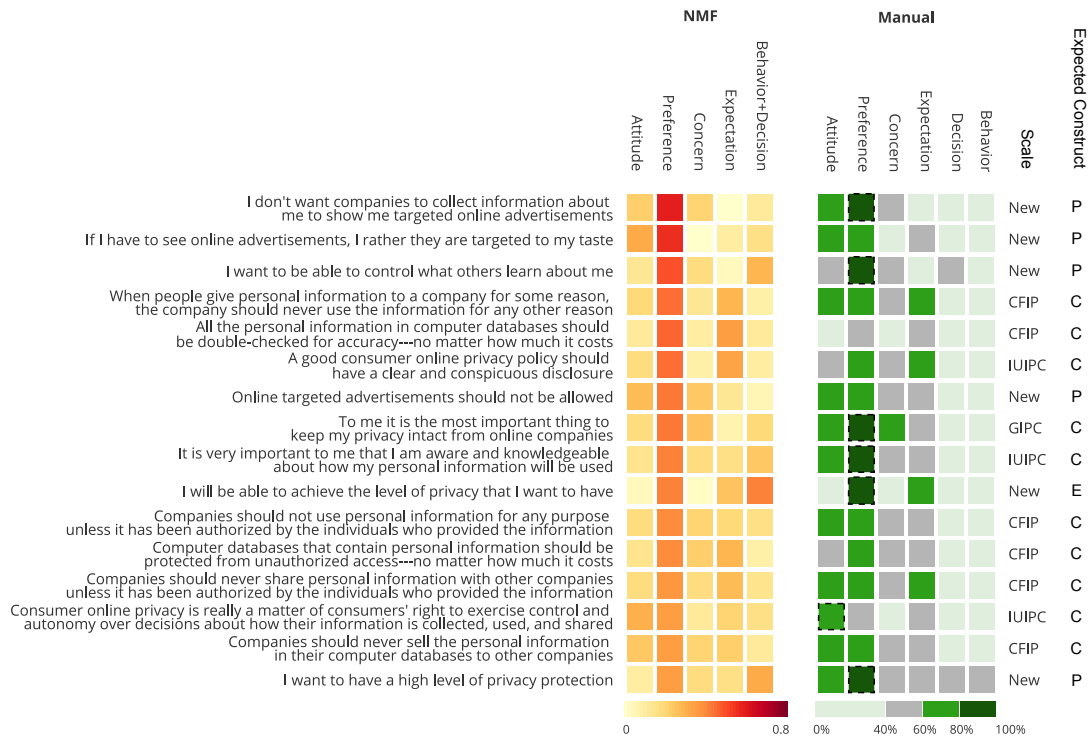


Figure 3: Heatmap displaying the NMF coefficient results showing the composition of each statement based on the meta-constructs (left) and the percentages of high scores for each construct/statement pair (right) for statements under the “preference” meta-construct. The primary construct identified is highlighted by a dotted box.

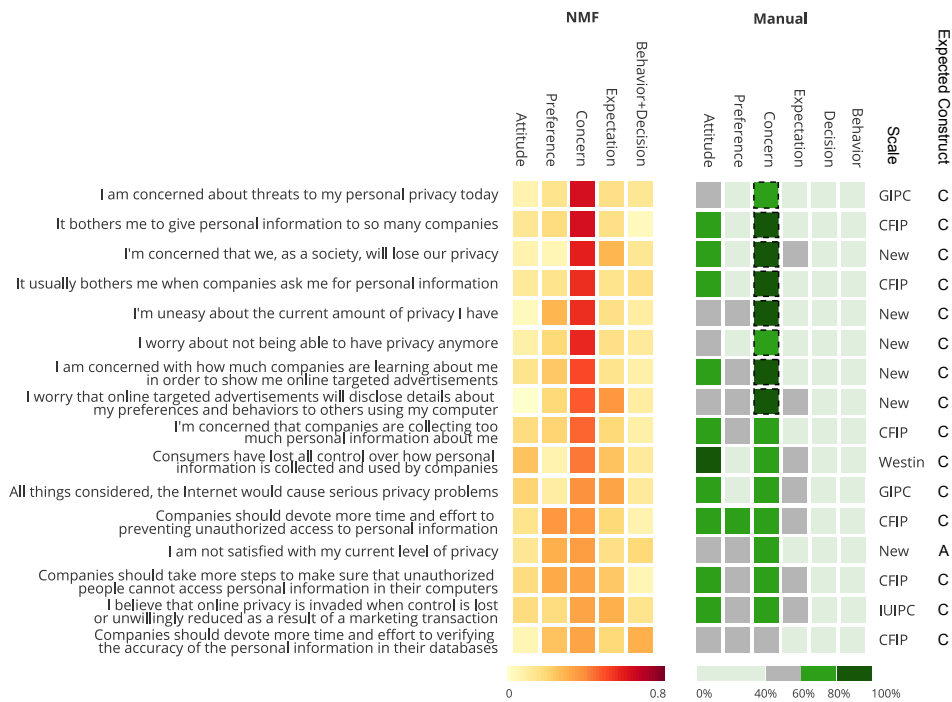


Figure 4: Heatmap displaying the NMF coefficient results showing the composition of each statement based on the meta-constructs (left) and the percentages of high scores for each construct/statement pair (right) for statements under the “concern” meta-construct. The primary construct identified is highlighted by a dotted box.

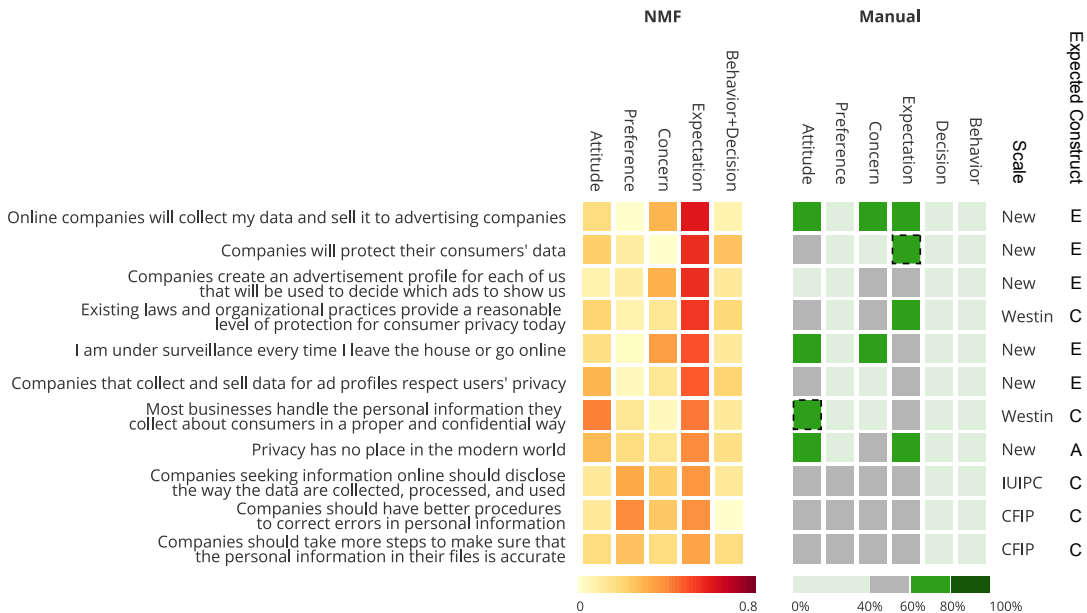


Figure 5: Heatmap displaying the NMF coefficient results showing the composition of each statement based on the meta-constructs (left) and the percentages of high scores for each construct/statement pair (right) for statements under the “expectation” meta-construct. The primary construct identified is highlighted by a dotted box.



Figure 6: Heatmap displaying the NMF coefficient results showing the composition of each statement based on the meta-constructs (left) and the percentages of high scores for each construct/statement pair (right) for statements under the “decision-behavior” meta-construct. The primary construct identified is highlighted by a dotted box.

- My life is an open book (Expected: attitude; classification: behavior and decision)
- Privacy has no place in the modern world (Expected: attitude; classification: expectation)

This suggests that even when building statements with specific constructs in mind, misalignment occurs between researchers’ goals and survey participants’ interpretations. In the next section we examine some of the linguistic patterns used in these statements that tend to be problematic or that tend to be associated with particular constructs. An awareness of these patterns may help researchers write statements that will be more likely to be interpreted as intended.

4.4 Corpus analyses on NMF groups

We conducted a corpus analysis to investigate whether linguistic patterns could be found that might help minimize problematic conceptual and lexical overlaps. The findings presented in Section 4 showed that some statements may be more strongly correlated with particular constructs; any regularities in the kinds of expression that occur in those cases could potentially be exploited in scale construction to improve researcher control over which constructs are being measured.

We constructed corpora (sets of statements) from the groups derived from NMF analysis. These were then analysed using Wmatrix [17]. WMatrix assigns broad semantic field categories and calculates overuse and under use of semantic

field categories between corpora. The software compares relative frequencies within the data and calculates log-likelihood and log ratio. We compared between construct corpora and the AMe06 corpus of written, published, American usage [15]. We discuss selected results of log likelihood analysis.⁵ High log likelihood ($p < 0.001 - p < 0.05$) represents statistically significant overuse of a semantic field in NMF corpus relative to AME06.⁶ Table 5 in the Appendix displays the binary log of the ratio of relative frequencies (log ratio) across statistically significant categories.

The following general patterns provide an instructive start. The *privacy attitude* corpus significantly overrepresented a range of semantic categories that unambiguously signal that the speaker is expressing an attitude or making an evaluation. Attitude verbs, nouns relating mental or conceptual objects, such as **thought**; **comparative judgements** and **judgments of importance** were prevalent in statements strongly correlated with *privacy attitude*. As noted above, ‘concern’, and ‘preference’ are sometimes considered subcategories of ‘attitude’ and so overlaps in overrepresentation were to be anticipated and were found; expressions signalling worry were overrepresented in both the *privacy attitude* corpus, and the

⁵See Appendix for full table of log ratio analyses. Log ratio is a metric of effect size, each point reflecting a doubling of the rate of occurrence in the NMF corpora relative to the AME06

⁶Unsurprisingly, given the context, certain categories (e.g. *Information technology and computing*; *business: generally*; *business: selling*) are over-represented across the corpora. These categories are common thematic topics across corpora.

privacy concern corpus, and value judgment categories occurred in both *privacy attitude* and *privacy preference*. However, the *privacy preference* corpus distinguished itself by overrepresentation of **verbs signalling desire and modals signalling desired outcomes**, including ‘want’ (under *Wanted*), ‘should’ (under *Strong obligation or necessity*) and ‘never’ (under *Time*). *Privacy concern* corpus distinguished itself with over-representation of a range of expressions signalling **negative attitudinal valence**, including attitude verbs and deverbal expressions, as seen under the categories *Worry* and *Failure*, along with negative morphemes (e.g. ‘un’ in ‘unauthorized’).

The *privacy expectation* corpus over-represented **future auxiliaries**, for example, ‘will’ under *Time: future* — a category also overrepresented in *privacy decision and behavior*. It was distinguished from the latter, however, by overrepresentation of **value judgments**. The decision and behavior corpus distinguished itself in over-representation of a range of **privacy-behaviour related verbs** (in categories: *Helping* (mainly populated by ‘protect’) and *Investigate, examine, test, search*) and verbs with privacy-related direct objects.

Perhaps the primary lesson to be extracted from this analysis is that statement interpretation is considerably more open-ended than has been previously accounted for. This open-endedness may be to some extent ineliminable due to close relations between the constructs.

Statements that saw least convergence between participants were long or syntactically complex — both factors increase the potential for participants to draw on distinct information sources leading to diverging interpretations. Shorter affectively ambiguous declaratives (i.e. declaratives with no clear indication of whether the content is intended to describe a positive or negative state of affairs) also led to high variation by participants, since lack of information leads to speakers supplementing background beliefs to extract an interpretation.

Those statements that saw greatest convergence between participants on a particular construct, suffered neither from excess length or brevity and bore features that encouraged participants to navigate the possibilities in similar ways. Statements aimed at measuring constructs signalling attitude types, for example, can be improved by including attitude verbs that clearly signal those types (for concern, ‘I worry/fear/am concerned that’ for preference ‘I like/prefer that/am comfortable with’). These provisional suggestions are not, however, programmatic, and should rather highlight work to be done in isolating linguistic factors that could help constrain participant interpretation.

5 Limitations

Our results are limited by a number of factors.

Sample: While we attempted to produce results that could be generalizable to the sample populations typically used in privacy studies by leveraging Prolific’s representative sample,

our results may still not generalize beyond that sample.

Analysis approach: While NMF is, to the best of our knowledge, the most well-suited method for the problem at hand, the algorithm may yield slightly different results in different executions. We minimized this by leveraging best practices, such as performing multiple executions and utilizing the consensus results. In our executions of the algorithm, these variations did not impact the findings presented here. Furthermore, our results are limited by the threshold selected for our analyses. We minimized potential issues with threshold selection by performing robustness checks, finding no significant impact to the findings.

Definitions: The definitions we proposed are a best-effort at an initial set to be used by the privacy community. However, they still need to be improved and more broadly vetted. Furthermore, while we tried to reduce the variation in interpretation of the constructs by providing participants with the associated definitions, there are no guarantees that the definitions were interpreted in the same manner by all participants.

6 Discussion

We presented the results from an investigation of constructs captured in privacy scales. First, we refined a set of definitions for commonly used privacy constructs with the aid of privacy experts. Next, we used these definitions to collect participants’ views on which constructs describe each of 69 statements. Those statements represent a collection of both newly generated statements and statements from privacy scales.

Our results suggest that statements from existing privacy scales measure multiple constructs simultaneously, and often represent constructs other than concern, which appears to be the intended construct. To a lesser degree, a similar phenomenon happens with statements that were designed with the constructs in mind. The observed lack of a one-to-one match between statement and construct is, arguably, a result of two separate factors: the inherent ambiguity of natural language and the overlap between privacy constructs. The observed mismatch between statements and constructs may be due in part to a lack of agreed upon definitions for different privacy constructs, and on the evolving understanding [1] and use of these terms since the scales’ creation.

We show that it is possible to leverage aspects of semantics and sentence structure to help participants identify a target construct. In general, simpler sentences that provide sufficient information to the reader, so that their range of interpretation is reduced, seem to be more successful at reducing variation in interpretation. Nevertheless, we must be mindful of how this information is framed to avoid eliciting an exaggerated emotional response [3, 10].

Nevertheless, it may be ultimately unlikely that we can create *statements* that *only* measure a specific construct. In this

paper, we show that the constructs considered in the privacy community are not perceived as fully independent—attitude, preference, concern, and expectation were frequently simultaneously selected, and behavior and decision were always simultaneously selected. This overlap between constructs likely explains why we, and previous work [10], observed how validated scales such as CFIP and IUIPC, which have shown high internal validity, contain statements that were described by multiple constructs: existing scales seem to be measuring a higher level construct, such as *privacy perspective*. Given that existing scales do not seem to uniquely measure the finer grained constructs the community commonly uses, as they are currently understood, moving forward we should acknowledge this issue and consider its impact on results.

Narrow interpretations based on the outputs of such scales and related statements have led to inconsistent findings such as the privacy paradox [1, 2, 8]. In addition to the many explanations already found for the paradox, fundamental issues may exist with the construct validity of our measuring tools.

7 Future work

There are different approaches that the privacy community can take in face of these results. Here we list a few possibilities, but they are not meant to be prescriptive or comprehensive.

Shared definitions: In this paper we present a set of definitions constructed with the aid of a diverse sample of privacy experts in the field. However, we acknowledge that this set does not necessarily have to be the one we agree to use as a community. Going forward *we need to discuss what these, and potentially other, constructs mean and develop a shared and consistent vocabulary*.

Scale development: The results presented under Section 4.4 could help in the creation of scale statements. Nevertheless, future efforts in developing scales *should take care in acknowledging the inherent and possibly systemic limitations of such tools within the privacy context*. In particular, these efforts should validate that the developed scale actually measures the construct it claims to measure and that, in all likelihood, the scale will measure a combination of related constructs. Furthermore, *we should conduct periodic assessments to ensure that scales are still in alignment with the contemporaneous understanding of these constructs*.

Measuring granular constructs: Given the overlap between more granular privacy-related constructs and the contextual nature of privacy, it is worth considering alternate methods of capturing these constructs beyond static, validated scales. *If a distinction between constructs is important to the research question at hand, using methods that allow researchers to follow up and tease apart the differences between constructs*

might be necessary. For example, to distinguish preferences, concerns, and expectations, participants might be given a description of a type of data collection and asked whether they would prefer to allow or restrict it from happening with their data (preference), whether they are worried about it happening (concern), and whether they believe it is happening (expectation).

8 Conclusion

We presented research meant to investigate our ability to uniquely and reliably capture people’s granular privacy perspectives. In particular, we focus on privacy attitude, preference, concern, expectation, decision, and behavior.

We found that existing, and newly developed, statements meant to capture specific privacy constructs frequently capture multiple constructs at once. This enmeshed nature of the explored privacy constructs could help explain why existing scales, while thoroughly validated when proposed, do not always succeed at providing predictive insights, for example, as to people’s engagement with privacy behaviors based on their privacy concerns. As an aid to future work developing privacy scales, we present key linguistic characteristics that could help in the creation of statements that more uniquely discern between constructs.

We further propose that future work create a well-accepted set of definitions for privacy constructs; take into account the limitations of existing privacy scales when leveraging them; periodically verify the alignment between scales and the contemporaneous understanding of what they are meant to capture; and, be mindful of the enmeshed nature of these privacy constructs, using appropriate research methods to tease them apart, when needed.

Acknowledgements

This work was supported in part by gifts from Norton-LifeLock, Google, Innovators Network Foundation, and the Carnegie Corporation of New York.

References

- [1] Alessandro Acquisti, Laura Brandimarte, and Jeff Hancock. How privacy’s past may shape its future. *Science*, 375(6578):270–272, 2022.
- [2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, 30(4):736–758, 2020.
- [3] Alex Braunstein, Laura Granka, and Jessica Staddon. Indirect content privacy surveys: Measuring privacy with-

out asking about it. *SOUPS 2011 - Proceedings of the 7th Symposium on Usable Privacy and Security*, 2011.

- [4] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.
- [5] Lorrie Faith Cranor and Florian Schaub. Usable and Useful Privacy Interfaces. In *An Introduction to Privacy for Technology Professionals, Second Edition*, chapter Chapter 5. IAPP, 2020.
- [6] D. Alan Cruse. *Hyponymy and Its Varieties*, pages 3–21. Springer Netherlands, Dordrecht, 2002.
- [7] Martin Fishbein and Icek Ajzen. Introduction. In *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*, chapter Chapter 1. Addison-Wesley, 1975.
- [8] Nina Gerber, Paul Gerber, and Melanie Volkamer. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers and Security*, 77:226–261, aug 2018.
- [9] Nicolas Gillis. The why and how of nonnegative matrix factorization, 2014.
- [10] Thomas Groß. Validity and reliability of the scale internet users’ information privacy concerns (iuipe). *Proceedings on Privacy Enhancing Technologies*, 2021(2):235–258, 2021.
- [11] Sven Ove Hansson and Till Grüne-Yanoff. Preferences. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.
- [12] Harris Interactive. Privacy on and off the internet: What consumers want. Technical report, Harris Interactive Inc, 2002.
- [13] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet Users’ Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, 15(4):336–355, dec 2004.
- [14] Eric McCready. Emotive equilibria. *Linguistics and Philosophy*, 35, 05 2012.
- [15] Amanda Potts and Paul Baker. Does semantic tagging identify cultural change in british and american english? *International Journal of Corpus Linguistics*, 17, 12 2012.
- [16] Sören Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. *International Journal of Human Computer Studies*, 71(12):1133–1143, 2013.
- [17] Paul Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549, 2008.
- [18] Janice C Sipior, Burke T Ward, and Regina Connolly. Empirically assessing the continued applicability of the iuipe construct. *Journal of Enterprise Information Management*, 2013.
- [19] H. Jeff Smith, Tamara Dinev, and Heng Xu. Information Privacy Research: An Interdisciplinary review. *MIS Quarterly*, 35(4):1689–989—1015, 2011.
- [20] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke. Information Privacy: Measuring Individuals’ Concerns about Organizational Practices. *MIS Quarterly*, 20(2):167, jun 1996.
- [21] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Laura Brandimarte, and Alessandro Acquisti. Would a privacy fundamentalist sell their DNA for \$1000... if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences. *SOUPS ’14: Proceedings of the Tenth Symposium On Usable Privacy and Security*, pages 1–18, 2014.
- [22] Worry. *APA Dictionary of Psychology*. American Psychological Association.

9 Appendix

Westin's Privacy Segmentation Index

- Consumers have lost all control over how personal information is collected and used by companies.
- Most businesses handle the personal information they collect about consumers in a proper and confidential way.
- Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today

GIPC

- To me it is the most important thing to keep my privacy intact from online companies.
- Compared with other subjects on my mind, personal privacy is very important
- Compared to others, I am more sensitive about the way online companies handle my personal information
- I believe other people are too much concerned with online privacy issues.
- I am concerned about threats to my personal privacy today.
- All things considered, the Internet would cause serious privacy problems

CFIP

Errors

- All the personal information in computer databases should be double-checked for accuracy--no matter how much it costs.
- Companies should have better procedures to correct errors in personal information.
- Companies should devote more time and effort to verifying the accuracy of the personal information in their databases.
- Companies should take more steps to make sure that the personal information in their files is accurate.

Unauthorized use

- Companies should not use personal information for any purpose unless it has been authorized by the individuals who provided the information.
- When people give personal information to a company for some reason, the company should never use the information for any other reason.
- Companies should never share personal information with other companies unless it has been authorized by the individuals who provided the information.
- Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared.

Improper access

- Companies should devote more time and effort to preventing unauthorized access to personal information.
- Computer databases that contain personal information should be protected from unauthorized access--no matter how much it costs.
- Companies should take more steps to make sure that unauthorized people cannot access personal information in their computers.

IUIPC

Awareness

- Companies seeking information online should disclose the way the data are collected, processed, and used.
- A good consumer online privacy policy should have a clear and conspicuous disclosure.
- Companies should never sell the personal information in their computer databases to other companies.
- It is very important to me that I am aware and knowledgeable about how my personal information will be used.

Control

- I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.
- Consumer control of personal information lies at the heart of consumer privacy.
- Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared.

Collection

(Used in both IUIPC and CFIP)

- It usually bothers me when (online) companies ask me for personal information.
- When (online) companies ask me for information, I sometimes think twice before providing it.
- It bothers me to give personal information to so many (online) companies.
- I'm concerned that (online) companies are collecting too much personal information about me.

Figure 7: Statements for each of the scales evaluated in this paper.

#	Statement
1	Companies create an advertisement profile for each of us that will be used to decide which ads to show us.
2	Companies that collect and sell data for ad profiles respect users' privacy.
3	Companies will protect their consumers' data
4	I already take steps to protect my privacy
5	I am concerned with how much companies are learning about me in order to show me online targeted advertisements.
6	I am not satisfied with my current level of privacy
7	I am under surveillance every time I leave the house or go online.
8	I don't care about privacy as long as I can use the service
9	I don't do anything to protect my privacy.
10	I don't mind that others know what I'm doing
11	I don't think that privacy is important to me
12	I don't think there's anything to worry related to privacy.
13	I don't want companies to collect information about me to show me targeted online advertisements.
14	I feel that society worries too much about privacy
15	I installed something on my browser to make it harder to track me online
16	I think that others worry too much about privacy
17	I think that privacy is important for society
18	I use private browsing for privacy reasons
19	I want to be able to control what others learn about me
20	I want to have a high level of privacy protection.
21	I will be able to achieve the level of privacy that I want to have.
22	I will be proactive about protecting my privacy.
23	I will install software to make it harder for my behavior to be tracked online.
24	I will take the privacy level that I am given.
25	I won't change any aspect of my online life to protect my privacy.
26	I worry about not being able to have privacy anymore.
27	I worry that online targeted advertisements will disclose details about my preferences and behaviors to others using my computer.
28	I would change how I use the internet to protect my privacy.
29	I'm concerned that we, as a society, will lose our privacy.
30	I'm uneasy about the current amount of privacy I have.
31	I've opted-out of online targeted advertisement through the NAI (Network Advertising Initiative) website.
32	If I have to see online advertisements, I rather they are targeted to my taste.
33	My life is an open book.
34	Online companies will collect my data and sell it to advertising companies.
35	Online targeted advertisements should not be allowed.
36	Only people who have something to hide need privacy.
37	Privacy has no place in the modern world.
38	Privacy is a fundamental human right.
39	Privacy is not enough of a reason for me to change how I use the Internet.

Table 4: List of candidate statements created for the purpose of this study.

Att	Pref	Conc	Exp	Beh+Dec	Category
-	2.81	-	-	-	Able/intelligent
-	-	-	-	3.53	Alive
3.38	3.42	-	-	-	Allowed
3.23	4.03	4.18	4.54	3.01	Business: generally
3.73	3.35	3.00	4.55	-	Business: selling
4.97	3.46	4.27	3.60	5.02	Closed; hiding/hidden
1.85	-	-	-	-	Comparing: different
4.62	-	-	-	4.81	Comparing: similar
-	10.21	-	-	-	Double-check
4.43	-	-	-	-	Exceed; waste
3.78	-	4.32	-	-	Failure
-	-	-	1.49	-	General actions / making
2.08	1.51	-	2.59	-	Getting and possession
-	-	-	-	3.01	Helping
2.75	-	-	-	-	Important
4.10	4.46	4.05	4.02	5.03	Information technology and computing
-	-	-	-	2.92	Investigate, examine, test, search
-	5.06	-	6.35	-	Knowledge
2.45	3.56	2.89	2.22	-	Knowledgeable
-	3.72	-	-	-	Learning
-	-	4.87	-	-	Like
2.26	-	-	-	-	Mental object: conceptual object
-	2.77	-	-	-	Money: cost and price
-	-	5.33	-	-	Not allowed
4.60	5.04	5.56	4.11	4.38	Not part of a group
-	-	-	-	1.08	Pronouns
-	3.32	-	-	-	Reciprocal
-	-	-	5.17	-	Sensible
-	3.00	-	-	-	Strong obligation or necessity
-	-	-	-	4.06	Texture
-	2.64	-	-	-	Time
-	-	-	2.49	2.59	Time:future
-	-	2.15	-	-	Time: present; simultaneous
2.11	-	-	-	-	Thought, belief
-	-	2.68	-	-	Trying hard
2.99	3.03	-	2.91	3.60	Using
-	2.87	-	-	-	Wanted
4.36	-	5.16	-	-	Worry

Table 5: Log ratio results across all statistically significant categories.

On recruiting and retaining users for security-sensitive longitudinal measurement panels

Akira Yamada^{1,3*}, Kyle Crichton^{2*}, Yukiko Sawaya¹, Jin-Dong Dong²,
Sarah Pearman², Ayumu Kubota¹, and Nicolas Christin²

¹ *KDDI Research, Inc.*

² *Carnegie Mellon University*

³ *National Institute of Information and Communications Technology*

Abstract

Many recent studies have turned to longitudinal measurement panels to characterize how people use their computing devices under realistic conditions. In these studies, participants' devices are instrumented, and their behavior is closely monitored over long time intervals. Because such monitoring can be highly intrusive, researchers face substantial challenges recruiting and retaining participants.

We present three case studies using medium- to large-scale longitudinal panels, which all collect privacy- and security-sensitive data. In evaluating factors related to recruitment, retention, and data collection, we provide a foundation to inform the design of future long-term panel studies.

Through these studies, we observe that monetary and non-monetary incentives can be effective in recruiting panel participants, although each presents trade-offs and potential biases. Contrary to our initial expectations, we find that users do not behave any differently in their first few weeks of participation than in the remainder of their time in the study. In terms of retention, we note that personalized enrollment follow-ups can lower initial dropout rates, but they are challenging and costly to scale. Communication, including following up with inactive users, is vital to retention. However, finding the right balance of communication is equally important. Interfering with a participant's everyday device use is a sure way to lose users. Finally, we present several findings, based on practical experience, to help inform the design of the data collection process in observational panels.

1 Introduction

Many recent studies have attempted to characterize how people use their computing devices under realistic conditions.

*Both authors contributed equally.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

Because of the limitations of user surveys and lab experiments, researchers have increasingly turned to longitudinal measurement panels, in which participant devices are instrumented, and their behavior extensively monitored over long time intervals [7, 14, 15, 20, 26, 29, 31, 33, 49, 57]. While these panels provide rich insights into real-world user behavior, they are difficult to conduct due to technical complexity, cost, and logistical challenges. As such, longitudinal panels remain relatively rare in the field despite the advantages they afford.

Central to the problem researchers face is the highly intrusive nature of longitudinal measurement studies. As users increasingly rely on computing devices—in particular smartphones—for all aspects of their life, measurements of device use become more and more privacy-invasive. This requires special attention be paid to data collection and storage security, further complicating cost and logistics. Equally important is that the privacy and security risks be properly communicated to potential participants. However, in presenting this information users may understandably be reluctant to participate. This leads to the fundamental challenge for researchers in conducting security-sensitive longitudinal measurement panels: recruitment and retention.

To better understand these challenges we present three case studies of recent large-scale longitudinal panels, featuring approximately 2 million, 2,000, and 600 users, respectively, and running for periods ranging from two to over four years. These studies were conducted in diverse geographical (Japan and the United States) and computing (personal computers and mobile devices) environments, using very different recruitment and retention techniques. For instance, one study used monetary incentives to recruit users, while another adopted a popular animation character; and the third study provided additional security functionality—in the form of an anti-phishing toolbar. Likewise, one of the studies features frequent interactions between the research team and the participants, while others only rely on minimal communication.

We aim to synthesize recommendations for recruiting and retaining participants in future privacy-intrusive panel studies. We selected these three studies because we were collectively

involved in various aspects of the design, conduct, and analysis of the research. Thus, we had direct access to the data, participants, and other researchers involved in each project. Our goal is not to provide a meta-analysis, but to assess recruitment and retention issues, based on (usually publicly unavailable) retention data and first-hand accounts. While our findings can apply to a broader set of studies relying on longitudinal panels, such as clinical health studies, we focus on security-sensitive panels where data collected are privacy-invasive and used to study security and privacy behavior.

We acknowledge that the differences between studies make direct, quantitative comparisons difficult, as does the relatively limited number of the panels considered. However, given the rarity of large-scale longitudinal measurement panels, we believe that there is great value in drawing what lessons can be learned from the few studies available. Acknowledging the aforementioned limitations, we employ a case study approach to qualitatively assess the three panel studies, supporting observations and findings with an appropriate level of quantitative evidence. We use a combination of measurements, research logs, surveys, and practical experience to compile a set of lessons learned regarding recruitment, retention, and data collection in long-term observational panels.

Overall, we find that both monetary and non-monetary incentives are effective in recruiting participants, although each may introduce its own potential bias. Contrary to our expectations, newly recruited users do not behave differently in their first few weeks than they do later on. As for participant retention, personalized enrollment and follow-ups can lower initial dropout rates, but are challenging and costly to scale. Communication, including following up with inactive users, is vital to retention, but finding the right balance of communication is equally important. Interfering with a participant's everyday device use is a sure way to lose users. Finally, we highlight the importance of monitoring for sensor outages and user dropouts, maintaining the order of observed events, establishing good measures for active user engagement, and handling multi-user devices and multi-device users.

2 Related work

We next discuss related studies by grouping them into three sets: recent user behavior measurement panels, work on participant retention in longitudinal studies, and inquiries in recruitment, motivation, and bias.

2.1 Measurement panels

Panels of personal computer users have been recruited to study a variety of behaviors related to human-computer interaction. These studies, which instrument the participant's computer with sensors, enable researchers to observe detailed information about the user's behavior over long periods of time. One major area of research using these panels has been

to study how users browse the internet and how that behavior changes over time [7, 29, 33, 49, 57].

In addition, numerous studies have used longitudinal panels to examine certain user security and privacy behaviors (e.g., password creation [35] or private browsing use [17]). Other work has examined behavior leading up, and in response, to encountering security threats such as cross-site scripting attacks and related scams [34] or drive-by-downloads [27, 28]. Some research has leveraged user behavior gleaned from these panels to predict exposure risk to malicious content [6, 25, 26, 42]. Besides characterizing user responses, several studies have used longitudinal panels to examine how users maintain their machines [38] and how accurately users perceive their own maintenance and security behavior [15, 51].

With users spending an increasing amount of time on their smartphones and tablets, researchers have recently taken to collecting data on mobile device use. Several early smartphone panels were created to enable researchers to deploy experiments related to smartphone use [20, 31]. These panels were used to compare a user's security intention to their actual behavior [8] and to develop a measure of users' information security awareness [4]. Other recent smartphone panels include investigations of smartphone lock use [50], and of how users evaluate requested permissions [53].

2.2 Recruitment motivations and bias

Previous work on recruitment incentives—predominantly focused on survey studies—has demonstrated that offering monetary incentives to participants improves recruitment rates and decreases non-response rates [23, 44, 46, 58]. Specific reward methods, such as lotteries, attract participants with psychologically-biased personalities and are highly effective in certain tasks [18]. Prior research on the use of non-monetary rewards suggests a similar, yet possibly weaker, effect [3, 58]. Alternatively, in volunteer-based platforms [1, 2, 37], the participants' motivation types highly affect attentions and dropouts [21]. However, relatively few studies have compared the effects of various recruitment incentives on sample composition or the quality of data collected [46]. What evidence exists suggests that monetary and non-monetary rewards do not equally appeal to all participants [58]. As a result, the use of different incentives can result in under- or over-representation of various demographic groups, especially related to education and income level [36, 40, 45]. Yet, previous studies have shown that incentives generally have no statistically significant effect on question non-response [43, 55].

2.3 Retention in longitudinal studies

Researchers conducting a measurement panel study must also retain user participation throughout a (often long) study. Maintaining contact with participants, recontacting participants who do not respond or show up, and using incentives have

been found to be key factors in user retention [52]. In their systematic review of 88 clinical studies, Robinson et al. identified 985 retention strategies and found a positive correlation between the number employed and retention rate. However, most clinical studies examined were descriptive, with only six of them designed to directly compare between strategies [39]. Of these studies, three found that cash payments and higher compensation led to higher retention [11, 12, 54], two reported higher retention rates for participants who received more contact and reminders from the research team [10, 13], and one found that small non-monetary rewards had no effect [5].

3 Methods

We next give an overview of the three measurement panel studies used in our analysis: the Security Behavior Observatory (SBO, [14, 15, 17, 35]), a Security Toolbar's trace data, and a Mobile Security Behavior Observatory (mSBO, [56]). We close with a discussion of the ethical review process and copyright licensing.

3.1 Security Behavior Observatory

The SBO was a longitudinal study of home computer use conducted between May 2015 and July 2019. As a part of the study, participants consented to have their home computers instrumented with a variety of sensors that collected, encrypted, and then transmitted data back to a central repository, in exchange for monthly payments. The study was limited to Windows desktop and laptop computers that were primarily used at home. The study received Institutional Review Board approval from Carnegie Mellon University.

Recruitment Over four years, the SBO project recruited a total of 623 participants who on average stayed in the study for just under two years ($\mu = 1.76, \sigma = 1.05$). Participants were predominantly recruited from one major U.S. metropolitan area, using a university research recruitment service as the primary recruitment source along with several secondary sources. Participants completed a pre-enrollment survey to confirm eligibility and provide consent, after which they received a phone call from a research team member to step them through the enrollment process in which consent was reconfirmed audibly. Individuals received \$30 upon enrollment and \$10 for each month they stayed in the study. If a participant encountered technical issues or data stopped being sent for an extended period of time, a member of the SBO research team would directly contact the participant via phone or email. Participants could discontinue their participation at any time.

Data collection The SBO was designed using a client-server architecture with several client-side sensors to collect different data types from participants' machines. Information

including the state of the user's machine, installed software, current processes, user interactions, and web browsing were sent whenever the participant's computer was powered on. We refer to Forget et al. [14] for a thorough discussion of the SBO architecture. Participants who reported issues with the sensors interfering with their daily use received a lightweight version of the sensor that only collected browsing data.

Upon completion of the study, participants were asked to complete an exit survey, described in Appendix B. The survey was distributed to the SBO email list to participants who had been in the study at any point. The survey was run on the Qualtrics online survey platform, where 203 responses were recorded. Those who completed the survey received a \$15 Amazon gift card as additional compensation.

3.2 Security Toolbar trace data

The second panel we look at is derived from data provided by a Japanese security company. This company offers a security tool to its customers which, as a part of its service, and with explicit customer agreement, collects web browsing information from the customer device.¹ This dataset contains more than four years of browsing data, ranging from December 2016 to February 2021. The data is limited to Microsoft Windows Internet Explorer (IE) users. However, this is less of a limitation than it may seem, as many Japanese administrations and businesses required IE until recently [30].

Recruitment The Security Toolbar is used as part of a specific type of web service used primarily in Japan. The web service partners distribute the toolbar on behalf of the security company as part of their services' security enhancement. Users can use the toolbar as long as they continue to subscribe to the web service and have the toolbar installed on their device. The data we have access to features over 2 million participants, with between 50,000–300,000 daily active users. Since Microsoft stopped IE support, the number of installations has declined over time. Prior to downloading the software, users are provided information about the data collected through the security tool, and are prompted to provide consent to continue. We obtained this data under a research agreement with the security company and the sharing of the data was approved by the Institutional Review Board at Carnegie Mellon University.

Data collection Data collection has been ongoing since December 2016. The collection software is installed as an add-on to the IE browser and sends encrypted data back to the company's servers. The data provided to us has been anonymized and does not include any demographic information. As such, we are unable to compare the sample composition with that of the other panel studies.

¹Due to a non-disclosure agreement with the company providing the tool and data, we cannot refer to the tool by name.

3.3 Mobile Security Behavior Observatory

The mSBO is an ongoing research project inspired by the SBO to observe user security behavior on mobile devices, and compare it to that of personal computer users. The application, which is free to download from the Google Play Store, collects data on how users interact with their mobile devices and periodically transmits the data to a central server when an Internet connection is available. Through a chat interface built in the app, users can report spam, phishing schemes, and malicious websites they encounter. Included in the application is a gamified animation character that appears on the user's home screen. Using "experience points" accumulated from interacting with the app and filling out periodic questionnaires (also provided in the app), users can customize the character's color, emotes, and vocabulary. Further details about the mSBO application, the system architecture, and the animation character can be found in Appendix C.

Recruitment The mSBO application was first distributed via the Google Play Store (Japan only) on March 16, 2020. The IARC generic rating was set to 18+ to prevent participants under the age of 18 from participating in the experiment. Upon downloading the app, users are asked to read and understand the terms and conditions to install. During installation, users are informed of the research project and are presented with information about the data collected through the app. Participants must separately consent to each type of data collected before they can start using the application. Participation can be discontinued at any time by uninstalling the application. In addition, users can withdraw consent at any time using a one-click option that leads to the deletion of all data collected from their device.

Coinciding with the launch of the app, recruitment was advertised on seven of our organization's websites and through our organization's Twitter accounts. An additional two-week Twitter recruitment campaign was run in June 2020. As of May 2021, 2,031 participants had installed the app, with approximately 400 daily active users.

Data collection Similar to the SBO, the mSBO relies on a client-server architecture. The mSBO application monitors the use of all other applications on the smartphone device as a background app. The sensor collects data on other installed applications, the use of those applications, web browsing, and network information. Within the app, a local heuristic filter purges email addresses, phone numbers, credit cards, SNS account names, and passwords from the collected data. In addition, the mSBO captures fuzzy hashes [24] of SMS messages that contain URLs, along with the plain text URL, to check for spam and malicious content. The data is then encrypted and sent back to a central server when the user's device has access to the Internet. Further details about the application architecture can be found in Appendix C.

Through the app, users can report security incidents and potential threats through a chat-based interface. In addition, short questionnaires are distributed twice a week which users can complete in exchange for experience points. The contents of the questionnaires vary widely and include topics such as security, information technology, and artificial intelligence.

Lastly, we distributed a 36-question survey through the mSBO application starting in December 2020. The survey asked users about their experience with prior research studies, their security behavior, and general demographic information. Included in the survey is a modified version of the 16-question Security Behavior Intentions Scale (SeBIS) developed by Egelman and Peer [9]. Since the survey was distributed to Japanese-speaking users, we utilized the revised RSeBIS scale which is more robust to language translation [41]. Because the SeBIS scale is geared toward personal computer users, we made slight modifications to several questions as follows. First, we replaced the phrase "computer screen" with "smartphone screen." Second, we combined two questions about device locking (F3 and F4) as they became essentially identical on smartphones. Third, we added a question about biometric authentication to better capture locking and unlocking behavior. Fourth, we removed a question regarding "mouse-over" use prior to clicking a link (F10) as that functionality does not exist on a mobile device. The full list of survey questions, including the modified SeBIS scale can be found in Appendix A. We will refer to this mobile-friendly version of the SeBIS instrument as the mRSeBIS scale. In total, we received 318 valid responses to the survey.

3.4 Ethics and copyright

Ethical review Data from the mSBO study and the Security Toolbar was collected in Japan by Japanese companies. In lieu of an academic Institutional Review Board (IRB), these studies were approved by an external ethics board which included privacy, legal, and ethics experts. All of the data collected as a part of these two studies was used for academic research purposes only and was not monetized in any way. U.S. researchers on the team did not collect any data related to these two studies, but received IRB approval from Carnegie Mellon University to receive and analyze it. The SBO study, which was conducted in the United States, received IRB approval from Carnegie Mellon University.

Copyright licensing To implement the mSBO mobile application, we adopted characters from a famous science fiction animated series. We obtained an official educational license from the copyright owner. The Android application is available on Google Play. We submitted additional license documents to Google for limited use of the characters when registering the app on Google Play. Users residing in Japan can download and install this smartphone application during the license period (currently ending in 2025).

4 Demographics

The key demographics from both SBO and mSBO studies are summarized in Table 1. Demographic information was not collected as a part of the Security Toolbar dataset. While the demographics in both samples are skewed in comparison to the general population, we find that the SBO sample is less representative. Most notably, participants in the SBO study had generally disproportionately lower incomes than the overall U.S. population. In the United States, 17.1% of the population have an income lower than \$24,000 [48], while as many as 32.1% of SBO participants reported an annual income lower than \$24,000. On the other hand, we do not observe significant income bias in the mSBO sample, which roughly aligns with the income distribution in Japan [47].

In addition, we observe a bi-modal age distribution in the SBO sample, skewed towards participants under 30 and over 60. This may be related to the income skew as the two largest subgroups in the SBO sample consist of university students and retirees, both which tend to have lower levels of income. Again, the mSBO sample does not present the same bias. However, the mSBO sample is strongly skewed toward men. We hypothesize this is because the sci-fi animation character in the mSBO app is based on Seinen manga, Japanese animation targeted toward younger adult men.

We do not observe substantial bias in the sample’s education levels. The mSBO sample slightly over-represents those with a high school degree or less, however this can plausibly relate to the animation character attracting younger male participants. On the other hand, the SBO sample is over-representative of participants with higher education.

5 Findings

We next present our findings and observations from the three panel studies. First, we examine various aspects of participant recruitment across the three studies. Second, we assess participant retention to identify factors that had positive and negative effects. Third, we draw upon these experiences to identify important practices for data collection and analysis.

5.1 Participant recruitment

Across the three panels we observe a range of different recruitment strategies, particularly in regards to the incentives offered to participants. We find that both monetary and non-monetary incentives are effective at recruiting panel participants. While we cannot draw causal conclusions about the effect of the incentives, based on survey responses from two of the panels we do observe key descriptive differences in participants’ motivation to join the study, privacy concerns, and security behavior. Despite these differences, and contrary to our own hypothesis, we do not find evidence to support the

Table 1: Demographics from SBO and mSBO studies

	Demographic	mSBO	SBO
Gender	Male	69.5%	40.2%
	Female	26.7%	59.3%
	Other/No response	3.5%	0.5%
Age	18-21	2.2%	5.3%
	22-30	10.4%	43.9%
	31-40	23.6%	16.0%
	40-50	36.8%	9.4%
	50-60	22.0%	8.9%
	Over 61	2.5%	16.0%
	No response	2.5%	0.5%
Education	No High School GED	3.8%	0.3%
	High School GED	28.0%	9.2%
	Some College	4.1%	24.4%
	Trade School Degree	18.9%	1.9%
	Bachelor’s Degree	29.9%	39.9%
	Master’s Degree	8.5%	20.1%
	Doctoral Degree	2.2%	4.2%
	Other/No response	4.7%	0.0%
Income	<2.5M JPY / <25K USD	21.1%	32.1%
	2.5-5M JPY / 25-50K USD	32.1%	22.0%
	5-7.5M JPY / 50-75K USD	19.8%	13.6%
	7.5-10M JPY / 75-100K USD	9.4%	8.2%
	10-15M JPY/100-200K USD	2.2%	8.7%
	>15M JPY / >200K USD	0.3%	2.1%
	No response	15.1%	13.2%
Occupation	Student	2.2%	35.9%
	Company employee	64.2%	40.2%
	Self-employed	5.3%	0.3%
	Public servant	8.2%	-
	Part-time job	6.3%	-
	Unemployed/Retired	-	22.0%
	Housewives and husbands	5.3%	0.5%
	Other/No response	8.5%	1.1%

theory of a more acute Hawthorne effect for users immediately after they are recruited into either study. Participants’ behavior and device use did not change between the period immediately following recruitment and the remainder of their time in the study.

5.1.1 Monetary and non-monetary incentives

Despite the use of a variety of incentives across the three studies, we observe that all of the incentives offered, both monetary and non-monetary, were effective at recruiting participants. Monetary incentives, like those offered in the SBO study, are a well-established form of compensation in research studies. In contrast, non-monetary incentives are infrequently used by the research community. However, longitudinal panels require incentives that can retain user participation over an often long period of time. This can be an expensive undertaking using monetary incentives. Looking towards alternative methods, the mSBO and Security Toolbar studies offered par-

ticipants a non-monetary incentive. The Security Toolbar, appropriately named, incentivized users by providing a security service as they browsed the web. In the mSBO study, users were offered a gamified, customizable in-app character from a popular sci-fi animation series.

The Security Toolbar, whose recruitment and distribution was done through a software company, was able to recruit and maintain several hundred thousand participants. The SBO and mSBO studies, whose recruitment channels were similar to that of a typical research study, both were able to recruit hundreds of participants and maintain over 300 daily active users despite very different incentives being offered. In fact, recruitment for the SBO study using monetary incentives was arguably more difficult, required advertisement through multiple channels, and took a longer period of time to ramp up to the same number of users as the mSBO study.

Although we find monetary and non-monetary incentives to work effectively, there are several tradeoffs for researchers to consider and potential bias, discussed in the following sections, to be aware of. First, experimental design can be simpler when using financial rewards as there are fewer variables and design decisions involved compared to using non-monetary incentives. In the case of monetary rewards, only the amount of time the user has to spend and the amount of the reward are considered. On the other hand, the types of non-monetary motivations are “boredom,” “comparison,” “fun,” “science,” and “self-learning,” which affect the attributes and behaviors of the participants [21]. Second, while non-monetary incentives can lower the direct costs of recruitment, the indirect costs stemming from the design and maintenance of the non-monetary reward should be considered. Third, while both sets of studies compete with other platforms for a limited pool of participants, the incentive design can affect the type of competing platform. With monetary incentives, we find that participants have often used a variety of crowd-sourcing platforms that compete for their time and attention. Although research projects must compete with these other platforms, simply offering higher monetary rewards is generally enough. On the other hand, with non-monetary incentives, researchers cannot easily control the many intangible factors that lead to the widespread adoption of some free apps but not others.

5.1.2 Research participation and motivation

From the surveys in Appendix A and B, we found that SBO participants had more prior experience with research and survey platforms, signed up for research studies more frequently, and were more financially motivated to participate in research than their mSBO counterparts. Two-thirds of SBO participants reported having used at least one crowd-working or survey platform outside of the university recruitment service the SBO study used. In fact, 23% of SBO participants had signed up for research studies at least once a month over the previous year. Conversely, less than 10% of mSBO partici-

pants had used a crowd-working service, and less than 30% had used a survey platform service. Fewer than 5% of participants had signed up for research studies at least once a month over the previous year.

Furthermore, when asked to select among eight factors that were important when deciding to participate in a study, SBO participants reported they would prioritize how much they will be paid (76%) and the amount of work required (67%). In contrast, mSBO participants reported that the study purpose (77%) and the security and privacy of the data collected (65%) were most important. Payment amount (16%) and the amount of work required (48%) ranked among the least important factors for mSBO participants. The full prioritized lists of user motivations are shown in Appendix D.

5.1.3 Privacy concerns

mSBO participants were more concerned about how their data was being collected and by whom than SBO participants. mSBO participants rated the “security or privacy of data collected in the study” (65%) as the second most important motivating factor for participation out of a total of eight. “Who is conducting the study” (57%), an indicator of trust and reputation, was the third highest-rated. However, in the SBO study, security and privacy (37%) rated fifth, and who is running the study (26%) rated sixth. While not definitive, these differences could also be related to the incentive being offered, as previous work has shown that people are willing to sell their privacy for minimal amounts of money [16].

5.1.4 Security behavior

Similar to the self-reported privacy concerns, mSBO participants also reported having greater security concerns than their SBO counterparts. The participants’ security concerns were measured using the SeBIS, RSeBIS, and mobile RSeBIS (described in Section 3.3) scales in the SBO and mSBO studies. In addition, because we cannot survey users of the Security Toolbar, we instead compare mSBO and SBO results to those obtained in the original RSeBIS work, that targeted Japanese PC users [41], which is the closest proxy for our Security Toolbar users we could find in the literature. The distribution of the SeBIS scores of participants in these three studies is reported in Table 2. Participants in the mSBO reported the highest level of security concerns, followed by SBO participants and then Security Toolbar participants. The difference in the distribution of scores between all three studies was statistically significant at the 95% confidence interval ($p < 0.001$).

To validate our comparison among different versions of the SeBIS scale, we evaluated the mobile-friendly version of SeBIS (mRSeBIS) using the same methodology in the original SeBIS [9] and the revised RSeBIS [41] papers. This method relies on confirmatory factor analysis (CFA) and Cronbach’s α to evaluate the validity and reliability of the proposed instru-

Table 2: **Distribution of SeBIS scores across PC users [41], mSBO, and SBO studies.** Scores are normalized by the number of questions (RSeBIS: 16, mRSeBIS: 15, SeBIS: 16)

	PC users [41]	mSBO	SBO
Scale	RSeBIS	mRSeBIS	SeBIS
Responses	500	318	399
Mean	2.572	3.739	3.406
Standard Deviation	0.931	0.763	0.523
Minimum	1.067	1.667	2.250
Maximum	5.000	5.000	5.000

ment. Confirmatory factor analysis measures the alignment between the scales’ items and a set of hypothesized latent factors, which, in this case, include proactive awareness, password selection, device locking, and software updating. A high level of alignment indicates that the scale measures the factors we expect them to measure, i.e., the scale is valid. Cronbach’s α measures the scale’s reliability; in other words, the items are measuring the same construct. This is important, as an unreliable scale cannot be valid. Our results in Table 3 show that the mRSeBIS scale has high reliability and a good fit, roughly equivalent to that of the original SeBIS scale.

5.1.5 Influence of monitoring on initial behavior

In analyzing usage data from the SBO and mSBO studies, we did not find any differences in behavior during the period immediately following user recruitment and their long-term behavior. This ran contrary to our hypothesis that users would change their behavior during their first few weeks in the study in response to being more aware that their device was being monitored. In other words, we expected the Hawthorne effect to be more acute during this initial period since participants were repeatedly made aware of the data collection and monitoring procedures during on-boarding. In particular, we expected that participants might use their devices less initially

Table 3: **mRSeBIS scale validation.**

Scale	mRSeBIS (JP)	Recommended
N	318	
Cronbach’s α	0.818	>0.60 [9]
RMSEA	0.055	<0.06 [19]
SRMR	0.058	<0.08 [19]
CFI	0.954	>0.90 [32]
TLI	0.942	>0.90 [32]

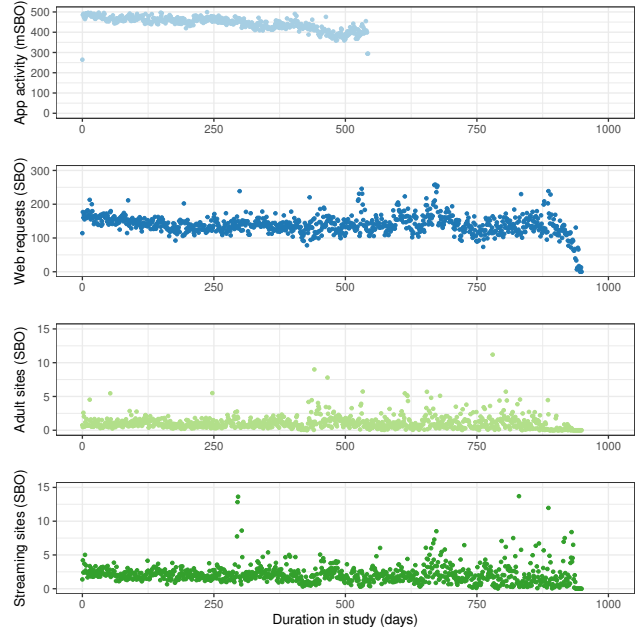


Figure 1: **Activity over time.** From top to bottom, average (1) foreground application records(mSBO), (2) user-initiated web requests(SBO), (3) visits to adult websites(SBO), and (4) visits to streaming websites(SBO), per user by the number of days elapsed since they joined the study.

and would refrain from engaging in privacy-sensitive activities like viewing pornography or visiting video streaming sites that frequently contain pirated content.

Figure 1 shows, relative to the number of days participants were in the study, the average application use for the mSBO; and user-initiated web requests, visits to adult websites, and visits to streaming sites for the SBO. As the figure shows, device use remained relatively constant regardless of the length of time a participant was in the study. We also observe SBO users visit adult and streaming websites from day zero onward. Thus, participants do *not* behave differently in an initial ramp-up period before reverting to usual device and browsing patterns. In other words, observed behavior in the period immediately following recruitment appears representative of true behavior. This is particularly important for short-term observational studies, which are much more common than longitudinal research panels.

5.1.6 Lessons learned on participant recruitment

- Both monetary and non-monetary incentives work effectively for recruiting panel participants.
- Indirect costs stemming from the design and maintenance of the non-monetary reward should be considered.
- Researchers compete for a limited pool of participants; incentives affect which platforms one is competing with.

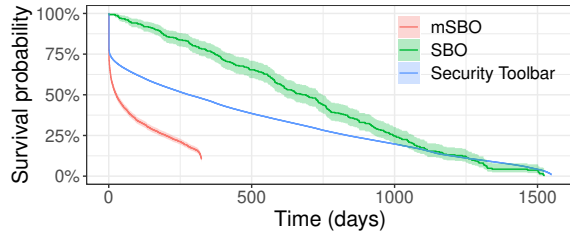


Figure 2: **Kaplan-Meier survival curves for all user panels.** Each point is the probability that a user participating at a time $t = 0$ will still participate at time $t = x$. The shaded area denotes the 95% confidence interval.

- Potential bias related to incentives should be considered, particularly related to privacy and security concerns.
- Newly recruited participants do not behavior differently in their first few days or weeks, than they do throughout the remainder of their time in the study.

5.2 Participant retention

Between the three studies, we observe markedly different retention rates among participants. Figure 2 shows the results of a Kaplan-Meier survival analysis [22] which illustrates the probability of a participant remaining in the study after a certain number of days. As shown, the survival curve for the SBO is relatively linear, with half of the participants dropping out after about 700 days in the study. In contrast, the Security Toolbar and mSBO study have high initial dropout rates, with participation stabilizing for users who stay in the study for at least a month. In fact, after a month, Security Toolbar users are more likely to maintain their participation compared to the SBO and mSBO, as indicated by the flatter downslope of the curve. The mSBO study suffers the highest initial dropout rate, losing about 60% of participants over the first month. After stabilizing, participants drop out at a rate similar (slightly steeper) as in the SBO study. We next identify four factors that help to explain the differences we observe between studies.

5.2.1 Minimizing interference

The first factor influencing retention is the stability of the sensor software and its interference with the participant’s use of the device. The mSBO application was tested prior to the initial roll out, however we could not cover the entire spectrum of possible Android devices and versions of the Android operating system that could run the mSBO. Our testing focused on the functionality and usability of the app, such as not interfering with the participant’s normal use of their device. Unfortunately, unanticipated compatibility issues and software bugs led to application instability and unexpected crashes during the first four months of the app’s release.

The initial version of the app also continuously displayed the character icon on the home screen and when using other

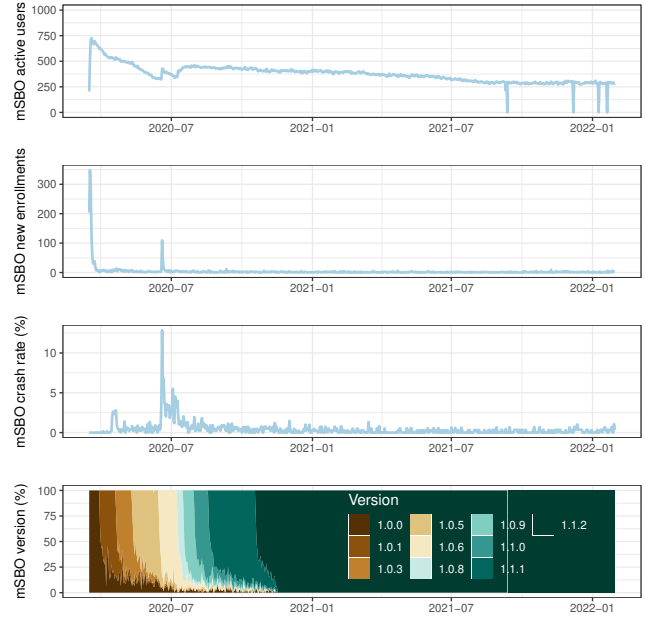


Figure 3: **mSBO panel evolution over time.** Each point is a computed over a one-day window.

applications. This was designed to remind participants of the app’s monitoring. However, we received feedback that this display feature severely interfered with user activities. The top panel of Figure 3, which shows the number of new and active mSBO users, demonstrates that during the period from March 2020 to July 2020, users left the study at a high rate. The second graph represents new installation, and the spike in late June 2020 reflects an additional Twitter recruitment campaign. The third graph illustrates crash encountering users per active users. After several bug fixes, a stable version of the app that disabled the constant character visibility was released on July 9, 2020. The fourth graph shows application version history. We released bug fixes and new features ten times during the first year. After the bug fixes released with version 1.0.5, the number of daily active users stabilized.

Similarly, one of the main complaints from participants in the early part of the SBO study was that the sensor software noticeably slowed down their device. The bottom plot in Figure 4 shows that opt-outs early in the study were primarily due to performance issues. This feedback led to the development of a lightweight version of the sensors. This was initially deployed only for impacted users before being rolled out to a broader set of users in December 2017. As the study continued, performance-related dropouts subsided, and a distinct decline in all dropouts occurred after the December 2017 deployment.

5.2.2 Communication balance

The second important retention factor is striking the right balance of communication with participants. In the mSBO study, we hypothesized that regular notifications would increase users' engagement with the app. Initially, users received three notifications per week with messages or questions for them to answer. However, it became clear that users found this level of communication too high, and many uninstalled the app. In the stable version of the app released in July 2020, we turned off the notifications and made them optional. Combined with fixing the bugs mentioned in Section 5.2.1, stopping the notifications helped to stabilize the number of active daily users. Neither the SBO nor the Security Toolbar studies employed notifications.

5.2.3 Following up with participants

As a corollary to communication, the third retention factor is the level of follow ups with participants. In the SBO study, new participants received an initial enrollment call from a member of the research team when they signed up. A member of the research team would also call participants to follow up whenever the participant had stopped sending data for an extended period of time. Figure 4 shows the number of active SBO users, new enrollments and calls, inactive users and follow-up calls, and opt-outs over the course of the study. Unlike the other panels, the SBO study maintained a positive increase in active daily users during the early phase of the study and relatively linear survival curve throughout. For comparison, of the 1,502 users who installed the mSBO app, during the first week, 25% effectively dropped out by either not opening the app, failing to provide consent, refusing the required permissions, or configuring the app settings incorrectly. We believe that the SBO enrollment calls helped alleviate this problem by addressing participant concerns upfront and resolving initial technical issues. In addition, the follow-up calls to inactive users likely helped achieve a lower attrition rate compared to the other panels: actual opt-outs were low.

5.2.4 Tangible benefits to participants

The fourth and final factor is providing a tangible benefit to participants. This effect is primarily observed among Security Toolbar users who, after an initial drop in participation, were the most likely to remain in the study long-term. While the mSBO app offers some utility through its reporting mechanism, the Security Toolbar provides everyday security benefits by helping prevent social engineering attempts. This benefit makes the toolbar quite popular among IE users and helps to explain the high retention rate.

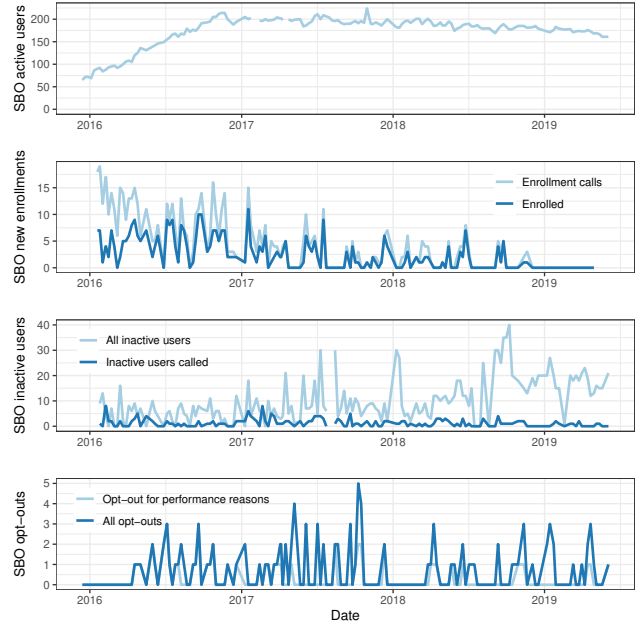


Figure 4: **SBO panel evolution over time.** Each point is a computed over a one-week window.

5.2.5 Device use and retention

One additional area of interest was the relationship between different types of users and their likelihood of remaining in the observational panel. In particular, we theorized that the frequency of device use might impact retention and, as a secondary effect, bias the sample. Based on an analysis of the mSBO and SBO studies, we find a mix of evidence. Table 4 shows the results of a series of regressions comparing the relationship between several metrics of device use and the length of time participant's remaining in the panel. The metrics for device use were log-transformed to obtain normal distributions and heteroskedastic robust standard errors were employed where Breusch-Pagan tests indicated heteroskedasticity. For additional details, see Appendix E.

While we observe a substantial amount of noise, we find a statistically significant positive relationship between how frequently a participant uses their device and how long they stay in the mSBO study. In the SBO study, we only find a significant relationship between average web requests and the duration in the study. It is possible that participants who used their device more frequently were more motivated to stay in the mSBO study due to the gamification of the animation character. The primary means that users leveled up their character, thereby unlocking additional features, was by filling out weekly surveys and reporting security issues they encountered. However, neither of these factors were strongly correlated with average web requests (surveys: $\rho = 0.287$, reports: $\rho = 0.171$) or average app use (surveys: $\rho = 0.356$, reports: $\rho = 0.250$).

Table 4: **Results of regression models.** These compare several metrics of average device use (independent variables) and the number of days in the study (dependent variable).

Study	<i>n</i>	Independent Variable	Coefficient	Intercept	<i>p</i> -value
mSBO	2229	Average web requests per active day (log)	11.526	112.879	0.020
		Average app use per active day (log)	64.359	-36.656	<0.001
		Network connections per active day (log)	126.923	-28.276	<0.001
SBO	307	Average web requests per active day (log)	35.227	239.782	0.0.031
		Average user-initiated web requests per active day (log)	31.571	326.858	0.081
		Average tab use per active day (log)	14.006	397.662	0.362

5.2.6 Lessons learned on participant retention

- Researchers should test the usability of the sensor software, which should not interfere with normal device use.
- The stability of the sensor software is vital for retention.
- Finding the right balance of communication between researchers and participants is critical.
- Researchers should monitor technical difficulties and follow up with participants quickly.
- Providing a tangible benefit to participants contributes to long-term retention.

5.3 Data collection

Planning for and designing the data collection process in longitudinal studies can be quite challenging. Often, unanticipated events arise during the course of the study that were not accounted for initially. This is particularly true for panels studies that span multiple years. In the following sections, we identify several data collection challenges and useful design decisions based on practical experience that can aid future researchers in creating observational panel studies.

5.3.1 Data collected per user

To give researchers a sense of how much data they can expect to collect, we analyzed the average amount of data collected per user in the SBO and mSBO studies. We find that on average researchers can expect to collect between 550–600 web requests on personal computers and between 50–100 web requests on mobile devices per user per day. Of the web requests made using personal computers, only about 12% of those are initiated in response to user-initiated navigation (e.g., link, bookmark, search, etc.). The remaining 88% were automatically generated by the browser or the web page. In addition, we find that personal computers users interact with browser tabs (e.g. create, switch, or close) about 120 times per day on average. On mobile devices, users interact with

and switch between different apps about 500 times per day on average.

OS limitations make it difficult to observe all web requests on mobile devices. VPN or web proxies could help alleviate this issue, but may degrade the user experience. Directly observing the URLs displayed in the web browser navigation bar is also challenging, as different smartphones frequently use different default web browsers (manufacturers often pre-install their own fork of, e.g., Chrome), with their own navigation bar. This, in turn, increases the complexity of the mobile sensor. Even more importantly, users spend more time on other applications than web browsers, and those applications may rely on internal browsers—using system HTML-rendering libraries, but with a different layout.

5.3.2 Identifying dropouts and technical difficulties

As mentioned in Section 5.2.3, following up with inactive users played an important role in retaining users. Therefore, the monitoring software should identify user inactivity. One way to accomplish this, which was used in several panels, was to create automated alerts or regular reports for users whose devices had stopped sending data for an extended period of time. In the SBO study, regular reports were used to follow up with participants manually. In the mSBO study, a “forget me” button was deployed so that participants could clearly signal their intention to dropout of the experiment.

However, a user device might also stop sending data due to a technical problem rather than the user intending to drop out of the study.² When it comes to data analysis, these kinds of gaps can be difficult to account for. The observation software used in the SBO study, which was designed with two sets of independent sensors, encountered many cases where one set of sensors would temporarily go down while the other would continue to send data. As a result, a large amount of analysis work was applied to detecting these gaps, and a substantial portion of the data collected had to be thrown out. One solution to this problem is to install an independent

²Often this was a result of a sensor failing until the device was restarted or the installation of other software that conflicted with the sensor software.

heartbeat sensor that regularly pings the home server. This can alert the research team when sensors go down as opposed to when a user is simply inactive.

5.3.3 Timestamps and order of events

For observational studies where data is being collected from a user device, timestamps alone are often not sufficient in maintaining the order of observed events. Many computational events can occur within the same (milli)second. In the SBO study, this led to significant post-hoc analysis to recreate the proper sequence of events, and in some ambiguous cases, data had to be discarded. A simple sequence counter enumerated by the sensor software would have alleviated this problem.

However, an ordering mechanism would not have fully solved all timestamping issues. Multiple studies observed skew in the timestamps recorded on participants' devices. This can occur if the user's device is not synchronized to a global time source, the user's device is defective, or the user manipulates their device's internal time intentionally. Furthermore, relying on the timestamp of data arrival at the server is insufficient. Users go offline often, even if they are connected to a mobile network, which delays the upload of sensor data. While not perfect, we find capturing the order of events, the number of seconds that have elapsed between events, and a combination of client and server timestamps to be most effective for data cleansing and analysis. Careful consideration and storage of a user's time zone, which may change throughout the course of the study, is also recommended.

5.3.4 Defining active user engagement

One limitation in using sensor data is that it provides the perspective of the device and only indirectly that of the participant. This can prove challenging when attempting to determine how long a user is actively interacting with their device, an important metric for many applications. For example, when a user navigates to a new web page, that information is logged by the sensors. However, if the user does not interact with their device for an extended period of time, it is unclear whether they are still engaged with that page or if they have left their device on but unattended.

In these studies, user engagement was roughly time-boxed using other recorded events, such as when the user navigated to the next web page or switched browser tabs. The mSBO study also used the foreground application history. However, some mobile apps, e.g., calendar and weather, always occupy the foreground of the screen, which makes it difficult to determine whether the user is active. The SBO study also relied on log in/out, power on/off, lock/unlock, and application change events. To refine this further, mouse movements were overlaid with the activity trace to determine active periods of user interaction. However, even this method is imperfect, as users could still be passively engaged with their device, like when

watching a movie, even if they are not actively using their mouse. We recommend that future studies explicitly capture events that indicate the end of a user interaction (e.g., closing a web page or application) if available. While the use of audio and video could provide precise measurements, they also raise substantial privacy concerns, and were avoided in these studies. Alternative measures of active engagement like mouse movements, keyboard use, touchscreen interaction, and resource usage are likely a better, albeit less precise, method.

5.3.5 Multiple users and devices

Over the course of a longitudinal panel study, there likely will be instances of multiple users sharing a given device (more so for personal computers than smartphones), and individual users with multiple devices. In the SBO study, in several cases more than one person was using a single personal computer. It would have been very useful to differentiate users, either by requiring separate logins or using some other identifier. In addition, over the course of the five-year study, most participants upgraded or replaced their computers at some point. A process for handling these cases was not originally in place, resulting in several substantial gaps in data coverage as users switched from one device to another. In the mSBO study, the multi-device problem typically occurred when one person owned more than one smartphone. The use of the primary smartphone differed greatly from that of secondary devices.

5.3.6 Survey distribution

In the mSBO study, the platform was designed such that researchers could distribute surveys to participants directly through the application. This provided a quick and easy way for researchers to interact with participants and gather supplemental data. Using this feature, researchers sent weekly questionnaires to which participants responded at an average rate of 30% of active users. In the SBO study, researchers had to distribute surveys to participants by email. Having to coordinate with participants, often individually, created significant overhead, so that surveys were distributed very infrequently. However, communicating by email rather than through the platform also enabled the SBO study to survey users who had previously participated in the study but had since left.

5.3.7 Lessons learned on data collection

- The monitoring software should identify user inactivity.
- Software should accurately record timestamps and event order, consider clock skew, and network disconnections.
- Researchers should capture metrics to define active user engagement.
- Monitoring should be designed with multiple users and devices in mind.

- Distributing surveys through the sensor software provides greater ease of use and flexibility.

6 Discussion

In the following sections we discuss the limitations of our work, and the implications our findings have for future studies.

6.1 Limitations

Since these studies were run independently, the main limitation of our analysis is that we cannot draw causal inference from any comparison across studies. These studies were also run in two different countries, each with distinct cultures, which may account for some of the observed differences. In addition, the sensors in these studies were device- and platform-specific. The mSBO application was limited to Android smartphone devices, the SBO platform was only available to PC users running Microsoft Windows, and the Security Toolbar was specific to Microsoft's Internet Explorer browser. These factors limit the generalizability of our findings.

6.2 Recommendations for future panel studies

In general, we find that retaining users in measurement panel studies is challenging, especially with new users. Personalized enrollment follow-ups can lower initial drop out rates but are also demanding and costly to scale. Communication, including following up with inactive users, is vitally important to retention. However, finding the right balance of communication is equally important and likely depends on the context of the study. Ideally, communication with participants should be enough to engage users without annoying them. In practice, making the sensors as invisible as possible may be best as interfering with everyday use of a device is a sure way to lose users. Conducting user testing early, with a variety of hardware and devices, is highly recommended.

There is no evidence of a ramp-up period for new users. Hence, participant data collected on initial device use are not as biased as we had originally hypothesized. This helps alleviate concerns over the results of short-term observation studies, and justifies including initial observations alongside long-term observations.

Foresight in designing data collection methods for long-term observation is quite difficult and unanticipated challenges are almost guaranteed. We propose five recommendations to help assuage these issues. First, design mechanisms within the observational software to identify user dropouts and sensor outages. Second, use a variety of sequences, time deltas, and timestamps to maintain the correct order and timing of observed events. Third, collect data that can help to clearly define when users are actively using the device such as start and end events, mouse movements, keyboard and touch-screen interactions, and device resource utilization. Fourth,

create a proactive process to handle cases where multiple participants use the same device and multiple devices are used by the same participant. Fifth, build in a mechanism, preferably within the observation platform, to easily follow up with participants, solicit feedback, and distribute surveys.

6.3 Future research

Our results indicate that monetary and non-monetary incentives provide viable means of recruiting participants for longitudinal measurement studies. However, both types of incentives have tradeoffs to consider and potential bias that they introduce. In presenting these three case studies, we are unable to draw causal conclusions about the effects of the various incentives offered. Further research, in a controlled setting, is needed to understand these effects with particular focus on participants' privacy concerns, security behavior, and motivation for participation. In addition, differences in privacy and security concerns may provide an opportunity for researchers to appeal to participants in recruitment and retention. Future work examining such methods would greatly benefit work on longitudinal panels.

7 Conclusions

This paper provides the first evaluation of factors that influence recruitment, retention, and data collection in longitudinal, security- and privacy-sensitive measurement panels. While substantial related work has been done in the context of surveys and clinical studies, privacy/security measurement panels are unique in the intrusive nature of the data collected. These types of studies are relatively rare, although are increasingly being used to observe behavior in a variety of research related to human-computer interaction.

We examined three medium- to large-scale panel studies, which all primarily collect privacy- and security-sensitive data (notably web browsing data). The three studies differed in origin (Japan vs. United States), recruitment incentives (monetary, gamification, added functionality), devices studied (personal computer vs. mobile), degree of interaction, and monitoring software visibility.

Our work provides new insight into recruitment efforts for longitudinal panels, including the effectiveness of monetary and non-monetary incentives, and into participant motivations, privacy concerns and security behavior. We show evidence that users do not act differently during their initial time in the study compared to their long-term behavior, alleviating concerns of potential bias. We identify key factors that affect user retention, including device interference, communication, follow-ups with potential dropouts, and tangible participant benefits. Finally, we derive recommendations to inform the design of the data collection process in future panel studies.

Acknowledgments

This work was partially funded by the National Security Agency (NSA) Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140), and through a Carnegie Bosch Institute (CBI) Fellowship. This work was also partially supported by WarpDrive: Web-based Attack Response with Practical and Deployable Research Initiative, the commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

- [1] Lab in the Wild. <http://www.labinthewild.org/>.
- [2] Project Implicit. <https://www.projectimplicit.net/>.
- [3] Johannes Abeler and Daniele Nosenzo. Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics*, 18(2):195–214, 06 2015.
- [4] Ron Bitton, Kobi Boyngold, Rami Puzis, and Asaf Shabtai. Evaluating the Information Security Awareness of Smartphone Users. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, April 2020.
- [5] Deborah Bowen, Mark Thornquist, Gary Goodman, Gilbert S. Omenn, Karen Anderson, Matt Barnett, and Barbara Valanis. Effects of incentive items on participation in a randomized chemoprevention trial. *Journal of Health Psychology*, 5(1):109–115, 2000.
- [6] Davide Canali, Leyla Bilge, and Davide Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, ASIA CCS '14, pages 171–182, Kyoto, Japan, June 2014.
- [7] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065 – 1073, 1995. Proceedings of the Third International World-Wide Web Conference.
- [8] Serge Egelman, Marian Harbach, and Eyal Peer. Behavior Ever Follows Intention? A Validation of the Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5257–5261, 2016.
- [9] Serge Egelman and Eyal Peer. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2873–2882, 2015.
- [10] M. Florencia Etcheverry, Jennifer L. Evans, Emilia Sanchez, Eva Mendez-Arancibia, Mercé Meroño, José M. Gatell, Kimberly Page, and Joan Joseph. Enhanced retention strategies and willingness to participate among hard-to-reach female sex workers in barcelona for hiv prevention and vaccine trials. *Human Vaccines & Immunotherapeutics*, 9(2):420–429, 2013.
- [11] David S. Festinger, Douglas B. Marlowe, Jason R. Croft, Karen L. Dugosh, Nicole K. Mastro, Patricia A. Lee, David S. DeMatteo, and Nicholas S. Patapis. Do research payments precipitate drug use or coerce participation? *Drug and Alcohol Dependence*, 78(3):275–281, 2005.
- [12] David S. Festinger, Douglas B. Marlowe, Karen L. Dugosh, Jason R. Croft, and Patricia L. Arabia. Higher magnitude cash payments improve research follow-up rates without increasing drug use or perceived coercion. *Drug and Alcohol Dependence*, 96(1):128–135, 2008.
- [13] Marvella E. Ford, Suzanne Havstad, Sally W. Vernon, Shawna D. Davis, David Kroll, Lois Lamerato, and G. Marie Swanson. Enhancing Adherence Among Older African American Men Enrolled in a Longitudinal Cancer Screening Trial. *The Gerontologist*, 46(4):545–550, 08 2006.
- [14] Alain Forget, Saranga Komanduri, Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, and Rahul Telang. Security Behavior Observatory: Infrastructure for Long-term Monitoring of Client Machines (CMU-CyLab-14-009). Jul 2014.
- [15] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. Do or do not, there is no try: User engagement may not improve security outcomes. In *Proceedings of the Tenth Symposium on Usable Privacy and Security (SOUPS'16)*, Denver, CO, July 2016.
- [16] Jens Grossklags and Alessssandro Acquisti. When 25 cents is too much: an experiment on willingness-to-sell and willingness-to-protect personal information. In *Proceedings (online) of the Sixth Workshop on Economics of Information Security (WEIS'07)*, Pittsburgh, PA, 2007.
- [17] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Cranor. Away from prying eyes: Analyzing usage and understanding of private browsing. In *Proceedings of the Twelfth Symposium on*

Usable Privacy and Security (SOUPS'18), Baltimore, MD, August 2018.

- [18] Gary Hsieh and Rafał Kocielnik. You Get Who You Pay for: The Impact of Incentives on Participation Bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 823–835, February 2016.
- [19] Li-tze Hu and Peter M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, January 1999.
- [20] Kasthuri Jayarajah, Rajesh Krishna Balan, Meera Radhakrishnan, Archan Misra, and Youngki Lee. LiveLabs: Building In-Situ Mobile Sensing & Behavioural Experimentation TestBeds. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, pages 1–15, June 2016.
- [21] Eunice Jun, Gary Hsieh, and Katharina Reinecke. Types of Motivation Affect Study Selection, Attention, and Dropouts in Online Experiments. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):56:1–56:15, December 2017.
- [22] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.
- [23] Ngo Manh Khoi, Sven Casteleyn, M. Mehdi Moradi, and Edzer Pebesma. Do Monetary Incentives Influence Users' Behavior in Participatory Sensing? *Sensors (Basel, Switzerland)*, 18(5), May 2018.
- [24] Jesse Kornblum and Tsukasa Oi. Ssdeep – Fuzzy hashing program, Apr 2018. <https://ssdeep-project.github.io/ssdeep/index.html>.
- [25] Fanny Lalonde Lévesque, Sonia Chiasson, Anil Somayaji, and José M. Fernandez. Technological and Human Factors of Malware Attacks: A Computer Security Clinical Trial Approach. *ACM Transactions on Privacy and Security*, 21(4):18:1–18:30, July 2018.
- [26] Fanny Lalonde Lévesque, José M. Fernandez, and Anil Somayaji. Risk prediction of malware victimization based on user behavior. In *2014 9th International Conference on Malicious and Unwanted Software: The Americas (MALWARE)*, pages 128–134, October 2014.
- [27] Takashi Matsunaka, Junpei Urakawa, and Ayumu Kubota. Detecting and Preventing Drive-By Download Attack via Participative Monitoring of the Web. In *2013 Eighth Asia Joint Conference on Information Security*, pages 48–55, 2013.
- [28] Takashi Matsunaka, Akira Yamada, Ayumu Kubota, and Takahiro Kasama. A User-participating Framework for Monitoring the Web with Privacy Guaranteed. *IPSJ Journal*, 57(12):2682–2695, 2016.
- [29] B. McKenzie and A. Cockburn. An empirical analysis of web page revisitation. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, 2001.
- [30] Shusuke Murai. Japan sticks with Internet Explorer as microsoft ends support for old versions. Japan Times, January 2016. <https://www.japantimes.co.jp/news/2016/01/12/business/tech/japan-sticks-internet-explorer-microsoft-ends-support-old-versions/>.
- [31] Anandatirtha Nandugudi, Anudipa Maiti, Taeyeon Ki, Fatih Bulut, Murat Demirbas, Tevfik Kosar, Chunming Qiao, Steven Y. Ko, and Geoffrey Challen. PhoneLab: A Large Programmable Smartphone Testbed. In *Proceedings of First International Workshop on Sensing and Big Data Mining - SENSEMINE'13*, pages 1–6, Roma, Italy, 2013. ACM Press.
- [32] Richard G. Netemeyer, William O. Bearden, and Subhash Sharma. *Scaling Procedures: Issues and Applications*. SAGE Publications, Inc, Thousand Oaks, Calif, 1st edition edition, March 2003.
- [33] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 597–606, 2007.
- [34] Kaan Onarlioglu, Utku Yilmaz, Engin Kirda, and Davide Balzarotti. Insights into User Behavior in Dealing with Internet Attacks. May 2012.
- [35] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 295–310.
- [36] Daniel Petrolia and Sanjoy Bhattacharjee. Revisiting incentive effects: Evidence from a random-sample mail survey on consumer preferences for fuel ethanol. *Public Opinion Quarterly*, 73, 08 2009.
- [37] Many Brain Project. TestMyBrain. <https://www.testmybrain.org>.

- [38] Elissa M. Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L. Mazurek. Asking for a friend: Evaluating response biases in security user studies. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 1238–1255, 2018.
- [39] Karen A. Robinson, Victor D. Dinglas, Vineeth Sukrithan, Ramakrishna Yalamanchilli, Pedro A. Mendez-Tellez, Cheryl Dennison-Himmelfarb, and Dale M. Needham. Updated systematic review identifies substantial number of retention strategies: using more strategies retains more study participants. *Journal of Clinical Epidemiology*, 68(12):1481–1487, 2015.
- [40] Erica Ryu, Mick P. Couper, and Robert W. Marans. Survey Incentives: Cash vs. In-Kind; Face-to-Face vs. Mail; Response Rate vs. Nonresponse Error. *International Journal of Public Opinion Research*, 18(1):89–106, 07 2005.
- [41] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-Confidence Trumps Knowledge: A Cross-Cultural Study of Security Behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2202–2214, May 2017.
- [42] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. Predicting Impending Exposure to Malicious Content from User Behavior. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pages 1487–1501, Toronto, Canada, January 2018.
- [43] Eleanor Singer, Robert M. Groves, and Amy D. Corning. Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *The Public Opinion Quarterly*, 63(2):251–260, 1999.
- [44] Eleanor Singer and Richard Kulka. Paying respondents for survey participation. *Studies of welfare populations: Data Collection and Research Issues*, 01 2002.
- [45] Eleanor Singer, John van Hoewyk, and Mary P. Maher. Experiments with incentives in telephone surveys. *The Public Opinion Quarterly*, 64(2):171–188, 2000.
- [46] Eleanor Singer and Cong Ye. The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1):112–141, 2013.
- [47] Statista. Distribution of annual household income in japan in 2019. <https://www.statista.com/statistics/614245/distribution-of-annual-household-income-japan/>.
- [48] Statista. Percentage distribution of household income in the u.s. in 2019. <https://www.statista.com/statistics/203183/percentage-distribution-of-household-income-in-the-us/>.
- [49] Linda Tauscher and Saul Greenberg. Revisitation patterns in world wide web navigation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '97, page 399–406, 1997.
- [50] Dirk Van Bruggen, Shu Liu, Mitch Kajzer, Aaron Striegel, Charles R. Crowell, and John D'Arcy. Modifying smartphone user locking behavior. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 1–14, Newcastle, United Kingdom, July 2013.
- [51] Rick Wash, Emilee Rader, and Chris Fennell. Can People Self-Report Security Accurately? Agreement Between Self-Report and Behavioral Measures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2228–2232, May 2017.
- [52] Nicole Watson, Eva Leissou, Heidi Guyer, and Mark Wooden. *Best Practices for Panel Maintenance and Retention*, chapter 29, pages 597–622. John Wiley & Sons, Ltd, 2018.
- [53] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android permissions remystified: A field study on contextual integrity. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 499–514, Washington, D.C., August 2015.
- [54] Claire E. Wilcox, Michael P. Bogenschutz, Masato Nakazawa, and George E. Woody. Compensation effects on clinical trial data collection in opioid-dependent young adults. *The American Journal of Drug and Alcohol Abuse*, 38(1):81–86, 2012.
- [55] Diane K. Willimack, Howard Schuman, Beth-Ellen Pennell, and James M. Lepkowski. Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *The Public Opinion Quarterly*, 59(1):78–92, 1995.
- [56] Akira Yamada, Shoma Tanaka, Yukiko Sawaya, Ayumu Kubota, So Matsuda, Reo Matsumura, Shun Umemoto, Jun Nakajima, Kyle Crichton, Jin-Dong Dong, and Nicolas Christin. Mobile security behavior observatory: Long-term monitoring of mobile user behavior. In *Proceedings of USENIX Symposium on Usable Privacy and Security (SOUPS)*, August 2020. Poster abstract.

- [57] Haimo Zhang and Shengdong Zhao. Measuring web page revisitation in tabbed browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 1831–1834, 2011.
- [58] Guili Zheng, Sona Oksuzyan, Shelly Hsu, Jennifer Cloud, Mirna Ponce Jewell, Nirvi Shah, Lisa V. Smith, Douglas Frye, and Tony Kuo. Self-Reported Interest to Participate in a Health Survey if Different Amounts of Cash or Non-Monetary Incentive Types Were Offered. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 95(6):837–849, December 2018.

A mSBO survey

(Note: The questions below are translated into English from the original Japanese survey.)

Demographics

1. Which gender do you most identify with?

- man
- woman
- non-binary
- self-describe: [free text form]

2. What is your age?

- I would prefer not to respond
- 18–21 years old
- 22–30 years old
- 31–40 years old
- 41–50 years old
- 50–60 years old
- Over 61 years old

3. What is your highest level of education?

- I would prefer not to respond
- No High School GED
- High School GED
- Some College/Current College Student
- Trade or Technical School Degree
- Bachelor's Degree
- Master's Degree
- Doctoral Degree or Equivalent

4. What is your occupation?

- I would prefer not to respond

- student (esp. a university student)
- company employee
- self-employed
- public servant
- part-time job
- Housewife/husband
- Other:

5. What is your income level?

- I would prefer not to respond
- Less than 2,500,000 yen
- 2,500,000–5,000,000 yen
- 5,000,000–7,500,000 yen
- 7,500,000–10,000,000 yen
- 10,000,000–15,000,000 yen
- More than 15,000,000 yen

Modified SeBIS (5-point Likert scale; from “never” to “always”)

- F3‡: I manually lock my smartphone screen when I step away from it.
- F4‡: I set my smartphone screen to automatically lock if I don't use it for a prolonged period of time.
- F5: I use a PIN or passcode to unlock my mobile phone.
- F6‡: I use biometrics (fingerprint scanner, facial recognition) to unlock my mobile phone
- F12‡: I change my passwords even if it is not needed.
- F13: I use different passwords for different accounts that I have.
- F14‡: I include special characters in my password even if it's not required.
- F15: When I create a new online account, I try to use a password that goes beyond the site's minimum requirements.
- F8': When someone sends me a link, I open it only after verifying where it goes.
- F11‡: I know what website I'm visiting by looking at the URL bar, rather than by the website's look and feel.
- F16‡: I verify that information will be sent securely (e.g., SSL, "https://", a lock icon) before I submit it to websites.
- F7‡: If I discover a security problem, I fix or report it rather than assuming somebody else will.

- F1: When I'm prompted about a software update, I install it right away.
- F2: I try to make sure that the programs I use are up-to-date.
- F9: I verify that my anti-virus software has been regularly updating itself.

The dagger (†) symbol represents questions modified in RSeBIS [41] from the original SeBIS [8, 9]. The double-dagger (‡) symbol denotes questions modified from RSeBIS [41]. F6‡ is introduced instead of F6 because F6 and F5 become identical in the smartphone context. A question related to using mouse-over as a strategy (F12 in the original SeBIS) is removed because smartphones do not offer this functionality.

Do you have any complaints about using the app? (5-point Likert scale; from “strongly disagree” to “strongly agree”)

1. The application slows down my device.
2. The application drains the battery on my phone.
3. The app shuts down unexpectedly.
4. I receive too many messages from the app.
5. The application interferes with the normal use of my phone.
6. I am concerned about the privacy of the data collected.
7. Other: [free text form]

What is your level of satisfaction? (5-point Likert scale; from “very dissatisfied” to “very satisfied”)

1. The character on the home screen
2. Experience / Level
3. Changing the emote color
4. Periodic questionnaire
5. Phishing/spam report
6. Profile (screen time)
7. Protocol (install/consent process)

About the app

1. Where did you hear about this app?
 - Press Release
 - news site
 - Twitter
 - Other social networking sites (e.g. Facebook)

- Friend/Colleague
- Other: [free text form]

2. Over the past year, how frequently have you signed up for a new research study? (Not including this study)?

- Never
- Less than one per month
- About one per month
- About one per week
- Several times a week
- Multiple times a day

3. What factors are important to you when deciding what studies to participate in?

- How much I will be compensated for participating
- Amount of effort or work
- Whether I can participate at home / online (versus going somewhere to participate in person)
- Purpose or topic of the study
- Security or privacy of data collected in the study
- Who is conducting the study
- How quickly I will be compensated get paid
- The study's consent form
- Other: [free text form]

4. Would you be more likely to continue participating if you received a small recurring payment or if we continued to add new customizations and features to the character?

- Small recurring payment
- The character's new customizations and features
- Neither

B SBO exit survey

1. Have you participated in any other research besides this study? (“Research,” includes academic research, like this study, or marketing research, like surveys for companies.)

- (a) Yes, both in person and remotely (e.g., online, by phone or mail).
- (b) Yes, in person only.
- (c) Yes, remotely only (e.g. online, by phone or mail).
- (d) No.
- (e) Not sure.

2. What research platforms have you used to sign up for studies? (Please select all that apply.)
 - (a) [University 1 platform]
 - (b) Other [University 1] platform (please describe).
 - (c) [University 2 platform]
 - (d) Other [University 2] platform (please describe).
 - (e) Amazon Mechanical Turk (MTurk).
 - (f) Qualtrics.
 - (g) Prolific.
 - (h) Other (please describe).
 - (i) None of the above.
3. Over the past year, how frequently have you signed up for a new research study? (Not including this study.)
 - (a) Never.
 - (b) Less than one per month.
 - (c) About one per month.
 - (d) About one per week.
 - (e) Several times a week.
 - (f) Multiple times a day.
4. What factors are important to you when deciding what studies to participate in? (You may select multiple factors.)
 - (a) How much I will get paid.
 - (b) Amount of effort or work.
 - (c) Whether I can participate at home / online (versus going somewhere to participate in person).
 - (d) Purpose or topic of the study.
 - (e) Security or privacy of data collected in the study.
 - (f) Who is conducting the study.
 - (g) How quickly I will get paid.
 - (h) The study's consent form.
 - (i) Other.
5. Have you participated in other studies that collected data about how you used computer(s), smartphone(s), or other internet-connected devices?
 - (a) Yes.
 - (b) No.
 - (c) Not sure.
6. How was your experience participating in this study, overall?
 - (a) Positive
 - (b) Negative
 - (c) Not sure
 - (d) Other
7. What did you like about the study, if anything?
8. What did you dislike about the study, if anything?
9. Did you have any concerns or reservations about enrolling in this study?
 - (a) Yes.
 - (b) No.
 - (c) I don't remember.
10. What were those concerns, and what caused you to enroll anyway?
11. Would you participate in a study that used software to collect data about your computer usage again?
 - (a) Yes.
 - (b) No.
 - (c) Not sure / Depends.
12. If you wish, you may elaborate on why you would or would not participate in this type of study again.
13. How did this study's payments compare to other studies you have participated in?
 - (a) Less generous.
 - (b) More generous.
 - (c) About the same.
14. Do you have any feedback about the payments for this study? This could include payment amounts, the payment method (Amazon gift cards), payment timing, or other payment details.
15. Do you think you used your computer differently than you normally would due to our research software being installed on it?
 - (a) Yes.
 - (b) No.
16. What caused you to use your computer differently when our research software was installed?
17. What was different about your computer usage while our software was installed?
18. Do you have any other feedback or comments about this study that you would like for us to know?



(a) The character is displayed on the homescreen of users' smartphone anytime, and informs messages to users

(b) When a user taps the character, the screen transitions to the chat interface, where the user responds to questions

Figure 5: **mSBO user interface** with (a) the animation character on the users' home screen and (b) the chat interface used for reporting potential security threats and answering short surveys.

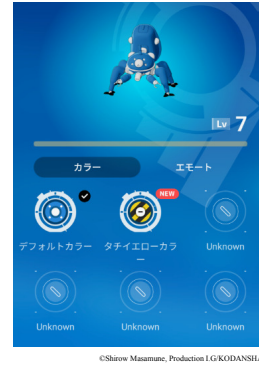
C mSBO app description

The mSBO app consists of a cartoon character agent on the smartphone's home screen and a chat-type interface. Fig. 5 shows a screenshot of the app. The animation character appears not just in the home screen but in any application. Tapping on the character icon launches the app and brings the user to the chat-type interface to interactively talk. We can send any message to the users as a pop-up attached to the character. We delivered questionnaire invitations through this pop-up.

We implement the character using Android's screen overlay functionality, enabling us to overlay Tachikoma on other applications. We do not implement the automatic location feature, so the icon and pop-up possibly cover other app displays or buttons. The users have to move the icon before tapping on something else. Although we use this design to clearly notify the users they were being observed, having the character constantly in the foreground admittedly may interfere with regular phone usage. Tapping on the character invokes the mSBO app, even if the participant is using another app.

The mSBO app provides experience points and stage levels to incentivize users. Figure 6 illustrates the color selection scene. Participants can earn points by answering regular questionnaires and reporting spam/phishing. As participants collect these points, they reach higher experience levels and can in turn further decorate the home screen character.

Figure 7 shows the mSBO sensor app configuration. The sensor app extensively relies on Android's Accessibility Ser-



©Shiro Masamune, Production I G/KODANSHA

Figure 6: **Color selection to decorate the home screen character.** This is based on the experience level reached.

vice, which is designed to provide alternative navigation feedback to applications installed on Android devices. For example, the Accessibility Service can be used to convert text to speech, or to warn of malicious web sites in addition to other tools (e.g., Google Safe Browsing). Most apps (e.g., Chrome, SMS, ...) fire `AccessibilityEvents` to communicate UI changes to the Accessibility Service.

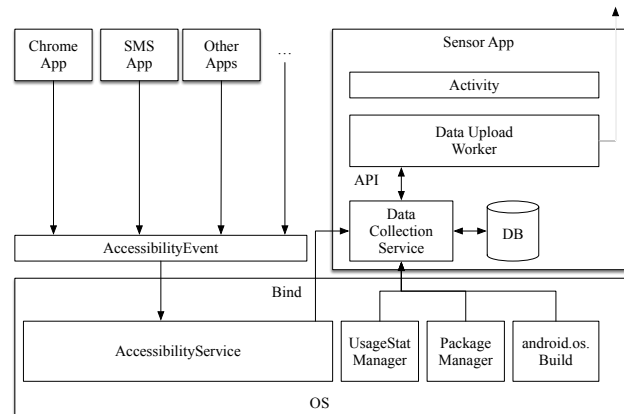


Figure 7: mSBO app configuration.

The mSBO app binds its own Data Collection Service to the Accessibility Service. That way, as long as the user grants Accessibility Service permission to the mSBO app, the Data Collection Service can capture whatever text is displayed in the app the user is running; e.g., the URL in the navigation bar, any anchor text in the browser, or any URL in an SMS.

The second major component of the mSBO is a `DataUpload Worker`. This worker, under Android's `WorkerManager`, uploads collected data as a background service. These uploads are scheduled, deferrable, asynchronous tasks, and are resilient to app crashes or device restarts.

D Participant motivations

In Section 5.1.2, we describe user motivations for engaging with the SBO or the mSBO. Below we present the full lists of motivations, ordered by decreasing priority, for both types of participants in Tables 5 and 6. SBO participants reported prioritizing how much they would be paid (76%) and the amount of work required (67%). In contrast, mSBO participants reported that the study purpose (77%) and the security and privacy of the data collected (65%) were most important. Payment amount (16%) and the amount of work required (48%) ranked among the least important factors for mSBO participants.

Table 5: Prioritized motivation list (SBO)

No.	Motivation	Rate
1	How much I will be paid	75.8%
2	Amount of effort or work	66.7%
3	Whether I can participate at home / online (versus going somewhere to participate in person)	64.1%
4	Purpose or topic of the study	52.0%
5	Security or privacy of data collected in the study	36.9%
6	Who is conducting the study	25.8%
7	How quickly I will get paid	20.7%
8	The study's consent form	13.1%
9	Other	4.0%

Table 6: Prioritized motivation list (mSBO)

No.	Motivation	Rate
1	Purpose or topic of the study	77.1%
2	Security or privacy of data collected in the study	64.9%
3	Who is conducting the study	56.7%
4	Amount of effort or work	48.0%
5	Whether I can participate at home / online (versus going somewhere to participate in person)	45.8%
6	The study's consent form	44.8%
7	How much I will will be compensated for participating	16.0%
8	How quickly I will be paid	4.1%
9	Other	3.1%

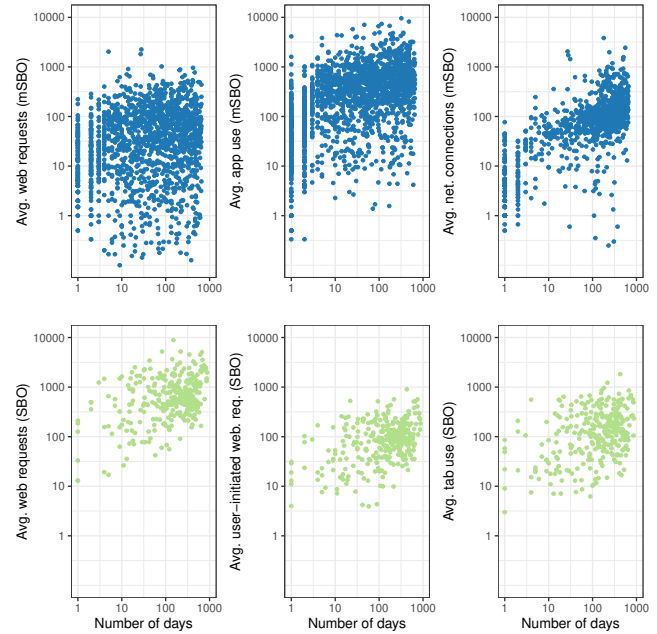


Figure 8: Device use and retention (scatter plots).

E Device use and retention graph

In Section 5.2.5, we ran several linear regressions to evaluate the relationship between device use and the length of participation in the study. As discussed, the dependent variables in the model were log-transformed to obtain normal distributions, an underlying assumption required for linear regressions. We tested for heteroskedasticity using Breusch-Pagan tests and found evidence of it in the mSBO sample. As such, we applied heteroskedastic robust standard errors in those regressions. All three metrics in the mSBO sample and the number of web requests in the SBO sample were found to have a statistically significant positive relationship with participants' duration in the study. We visually represent the relationship between the various metrics of device use and retention in both samples, in Figure 8. The figure presents scatter plots where the x -axis is the number of days users participated in the study, and the y -axes are the corresponding use of the device, according to various metrics. These scatter plots indicate that while a small positive relationship sometimes exists, the data are quite noisy.

Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys

Jenny Tang
Wellesley College

Eleanor Birrell
Pomona College

Ada Lerner*
Northeastern University

Abstract

Privacy and security researchers often rely on data collected through online crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) and Prolific. Prior work—which used data collected in the United States between 2013 and 2017—found that MTurk responses regarding security and privacy were generally representative for people under 50 or with some college education. However, the landscape of online crowdsourcing has changed significantly over the last five years, with the rise of Prolific as a major platform and the increasing presence of bots. This work attempts to replicate the prior results about the external validity of online privacy and security surveys. We conduct an online survey on MTurk ($n = 800$), a gender-balanced survey on Prolific ($n = 800$), and a representative survey on Prolific ($n = 800$) and compare the responses to a probabilistic survey conducted by the Pew Research Center ($n = 4272$). We find that MTurk response quality has degraded over the last five years, and our results do not replicate the earlier finding about the generalizability of MTurk responses. By contrast, we find that data collected through Prolific is generally representative for questions about user perceptions and experiences, but not for questions about security and privacy knowledge. We also evaluate the impact of Prolific settings, attention check questions, and statistical methods on the external validity of online surveys, and we develop recommendations about best practices for conducting online privacy and security surveys.

*Work was done while Lerner was at Wellesley College.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

1 Introduction

Over the last fifteen years, online surveys conducted through crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) [2] and Prolific [47] have become increasingly critical tools for conducting quantitative usable privacy and security research. Researchers often use these platforms to recruit participants for user studies. However, the external validity of these user studies depends on the extent to which the results of these online studies generalize to the overall population.

Prior work has investigated the validity of online surveys in various domains—such as social sciences [10, 11, 61], health behavior [53], and privacy [32, 52]—with somewhat mixed results. However, work by Redmiles et al.—based on surveys conducted between 2013 and 2017 [49]—made strong, positive claims about the external validity of privacy and security surveys conducted on MTurk. It found that (1) MTurk responses regarding privacy and security experiences, advice sources, and knowledge were more representative of the U.S. population compared to responses from a census-representative web panel and (2) MTurk responses regarding privacy and security experiences, advice sources, and knowledge were generally representative of the U.S. population for respondents who are younger than 50 or who have some college education.

However, the landscape of crowdsourcing platforms has changed significantly in the last five years. One key change is the rise of Prolific as a major crowdsourcing platform. Founded in 2014 specifically as a platform for conducting online user studies, Prolific was only rarely used to recruit participants in 2017. By contrast, we find that by 2021, Prolific was approximately twice as common as MTurk as a recruitment platform for usable privacy and security studies. A second key change is the increasing presence of sophisticated bots on MTurk, which can degrade data quality. While bots do not appear to have been a significant problem on MTurk in 2017, more recent work has estimated that 20-50% of MTurk accounts are actually bots, with significant bot levels dating back to approximately March 2018 [5, 39].

In light of those changes, this work attempts to replicate the key findings of Redmiles et al. [49]. We ask:

- (1) Are MTurk responses to privacy and security survey questions still representative of the U.S. population for respondents under 50 or with some college education?
- (2) To what extent do various classes of attention check question—reading-based attention checks, open text-response questions, and CAPTCHAs—and/or raking (i.e. demographic weighting) improve the generalizability of MTurk responses?
- (3) How well do Prolific responses to privacy and security questions generalize to the general U.S. population?
- (4) What are the current best practices for conducting and analyzing online user surveys in the domain of privacy and security?

Additionally, we investigate the limitations of online survey methods for surveying underrepresented demographic groups, reporting on ways that specific groups differ from the general population and how specific populations might be misrepresented by a focus on a representative sample.

To answer these research questions, we conduct an online survey on MTurk ($n = 800$), a gender-balanced survey on Prolific ($n = 800$), and a representative survey on Prolific ($n = 800$) and compare the responses to a probabilistic survey conducted through the Pew Research Center ($n = 4,272$). We find that MTurk response quality has degraded over the last five years, and our results do not replicate the finding that MTurk responses are representative of certain subsets of the U.S. population, even when we exclude the 39% of MTurk responses that fail attention checks and apply raking. We find that data collected through both representative and gender-balanced Prolific samples is generally representative for questions about user experiences, perceptions, and beliefs; however, responses to questions about knowledge of privacy and security concepts and about social media use differ heavily from the overall U.S. population. We also find that racial, age, and education subgroups from our Prolific representative sample are generally moderately representative of their respective subgroups in the American population.

Based on our results, we recommend that privacy and security researchers prefer Prolific to MTurk when recruiting participants for online user studies. Our results show that Prolific provides good quality, generalizable data for certain types of user studies about privacy and security (those that focus on experiences, perceptions, and beliefs), but that Prolific users are generally more technical than the overall population, resulting in different responses about knowledge and behavior. We do not recommend using attention check questions or CAPTCHAs on Prolific, as they lengthen surveys unnecessarily without improving external validity.

2 Related Works

Given the widespread use of crowdsourcing platforms as recruiting tools for user studies, the question of the data quality and external validity of online survey data has been extensively studied from a variety of different angles.

2.1 Generalizability of Online Platforms

Prior work has investigated the generalizability of online user studies conducted through MTurk and Prolific in a variety of different domains.

Amazon Mechanical Turk. Amazon Mechanical Turk has long been a platform favored by researchers across disciplines such as computer science and the social sciences to conduct user studies [11, 41, 42, 61], and thus the external validity of MTurk data has been investigated in various different research contexts [8–10, 25, 53] with varying results. Conclusions about the external validity of MTurk surveys about privacy and security have also been mixed: multiple studies [32, 52] have found significant differences between an MTurk study and a U.S.-representative survey, with MTurk users reporting more concerns about privacy and information use and higher levels of social media use, while Redmiles et al. [49] found that that for participants under 50 years of age or with at least some college education, responses to questions regarding privacy and security were similar to the general population within these demographics, and that MTurk appeared to be more representative overall than a census-representative web panel.

However, there have been noted concerns about demographic differences between the MTurk population and the over U.S. population. In particular, the MTurk population has been found to be younger and with higher levels of education than the overall U.S. population [32, 43, 49, 50]. Concerns have also arisen over the population on MTurk, particularly as highly active MTurk workers tend to complete many of the available tasks before others are able to, making the effective sample population on MTurk only 7000 [44, 56]. Furthermore, while a study conducted in 2014 found that MTurk workers with over an 95% approval rating provide high quality data and do not require attention checks [46], more recent research has shown that data quality on MTurk has decreased dramatically to be less reliable than that on Prolific, even when quality filters (at least 95% approval rating and 100 submitted tasks) were used [45].

Prolific. Prolific was launched in 2014, and was primarily designed for use by researchers [47]. In the past few years, we have seen an increase in the use of Prolific as an alternative to conducting surveys on Amazon Mechanical Turk. Both the number of users and the number of researchers on the platform have increased dramatically in recent years [41, 45], and studies have shown that it is a viable alternative to MTurk [44].

Prolific mandates a minimum hourly payment for studies, and compensation may be adjusted by researchers if the survey takes longer than originally intended. Furthermore, users on Prolific have an option to return their submissions, indicating they no longer wish to participate or that they do not wish their data to be used, making it easy for participants to withdraw consent at any time.

Although a study in 2017 found Prolific to be slower in gathering responses than MTurk and CrowdFlower (another online survey site) [44], we did not find such differences in our sample, perhaps due to the expansion of the Prolific worker pool over the last five years.

2.2 Data Quality and Attention Check Questions (ACQs)

Some prior work has found that MTurk workers performed well on attention check questions [30], but other work found that MTurk workers were less attentive than convenience-sampled college students [25]. Prior work comparing differing survey platforms in 2017 have found that almost half of MTurk and Prolific participants failed at least one attention check question, with MTurk users failing on average fewer attention checks than Prolific [44]. More recent work in 2021 saw Prolific users outperforming MTurk users on completing ACQs [45]. Excluding those based on passing attention checks had little effect for MTurk, and a small effect on Prolific [44].

Prolific specifically allows for Instructional Manipulation Checks (IMCs), which are questions that “explicitly instruct a participant to complete a task in a certain way” such as clicking a specific answer [47]. IMCs and other attention checks have been shown to increase the reliability of data, and have become relatively widely used [16, 27, 29, 40, 45]. However, IMCs might also influence participants to change interpretation and assessment of subsequent questions [29].

Some research has also investigated comprehension, which involves checking that participants are able to understand instructions and explain them back to the researchers. This can be conducted in formats such as IMCs, or through textboxes asking users to summarize the instructions. However, these might not function exactly the same as attention checks, as prior work suggests those who fail attention checks may not be the same as those who do not comprehend instructions [10]. Prior work has also found that Instructional Manipulation Checks making sure participants understood instructions improved data quality [20]. Prolific users also tend to outperform MTurk users on comprehension checks, and there appears to be a positive correlation between correctly passing ACQs and comprehension questions [45].

CAPTCHAs are commonly discussed as a mechanism for improving data quality by eliminating bots from a dataset, however prior work has found that bot accounts are able to reliably pass CAPTCHAs [39].

3 Methodology

To evaluate the generalizability of online privacy and security surveys, we compared survey responses from four sources: (1) responses to a U.S. nationally-representative probabilistic sample, (2) people recruited through Prolific using their representative sample option, (3) people recruited through Prolific using their gender-balanced option, and (4) people recruited through Amazon Mechanical Turk (MTurk).

3.1 Survey Questions

To decide what questions to ask on our survey, we started by identifying categories of topics in privacy and security that have been the subject of recent user studies. We identified 28 papers published in the Proceedings on Privacy Enhancing Technologies Symposium (PoPETs) and the Symposium on Usable Privacy and Security (SOUPS) in 2021 that included user surveys. Two papers [21, 23] exclusively surveyed specific technical populations (freelance developers and developers who have used Rust, respectively) about technical topics (security practices when developing code and experience with Rust), so we excluded them from our analysis. For the 26 papers that surveyed the public, we qualitatively coded the categories of questions asked in user surveys; we also determined what platform they used to recruit participants and how they handled attention check questions.

Our qualitative coding identified five classes of questions that characterize the space of recent usable privacy and security surveys:

1. **Behavioral.** Questions about what users do, would do, or have done in relation to technology, social media, and privacy and security tools. These questions refer to active behaviors undertaken by the user. For example, whether they use Twitter or whether they have recently decided not to use a service because of concerns about its data collection practices. 21 papers (80.8%) included behavioral questions in their user survey [1, 6, 7, 13, 17, 19, 22, 27, 28, 31, 35, 36, 48, 54, 57, 59, 60, 62, 64–66].
2. **Experience.** Questions about whether or how often participants had experienced a particular type of event. These questions refer to actions or circumstances that occur to the respondent without active action on the part of that person. For example, how often they had experienced someone taking over their social media or email account without their permission or how often they were asked to agree to a privacy policy. 17 papers (65.4%) included experience questions in their user survey [1, 6, 7, 19, 22, 27, 28, 31, 33–35, 48, 54, 57, 59, 64, 66].
3. **Knowledge.** Factual questions relating to privacy and security topics that test how much participants know about the topic. These questions have factually correct

answers. For example, what it means if a website uses cookies or what a privacy policy is. 11 papers (42.3%) included knowledge questions in their user survey [1, 6, 7, 12, 27, 31, 33, 34, 57, 58, 60].

4. **Perceptions.** Opinion questions about user perceptions of and attitudes towards practices and behaviors. These questions—which focus on what respondents believe a principal would do or the reasons why they believe the respondent would do something—include questions about trust, comfort, and mental models. For example, how confident they were that a company would follow what the privacy policy says it will do or how comfortable they are with companies using their data to help develop new products. 19 papers (73.1%) included perception questions in their user survey [1, 7, 13, 17, 22, 26–28, 31, 33–35, 54, 57–60, 64, 66].
5. **Beliefs.** Opinion questions about what security options or privacy rights people ought to have. Beliefs questions focus on what the respondent thinks should be true rather than asking about perceptions of the current world. For example, whether people should have the right to remove potentially embarrassing photos or criminal history from publicly-available search records. 9 papers (34.6%) included belief questions in their user survey [1, 7, 19, 26, 31, 34, 35, 54, 66].

For each of the five categories of questions, we selected 4–8 questions from a database of questions used in a past Pew Research Center survey [15] (a total of 30 questions). Drawing our questions from this source had two key advantages: (1) Pew questions are extensively validated before being deployed and (2) responses from a large-scale ($n = 4,272$), nationally-representative survey conducted by Pew in June 2019 are publically available [15], precluding the need to deploy our own nationally-representative panel survey. To enable intercomparison, our online surveys closely followed Pew’s methods: the phrasing of the questions were the same, the set of possible responses were the same, the order of the questions were the same—with randomization of question order or answer choices matching the Pew questionnaire—and there were no forced responses.

Since the Pew dataset includes demographic information for each participant, we also included basic demographic questions at the end of our survey. To facilitate comparisons with the Pew survey, we used demographic questions that matched the demographics released in the Pew dataset.¹

Finally, we identified three common techniques for excluding bots from online survey populations: reading-based attention check questions (i.e., questions that require participants

¹Note that these questions do not reflect current best-practices for asking about gender [55] or race [63]. Nonetheless, we believed that matching the Pew phrasing was critical in order to enable direct comparisons with responses to the Pew survey.

to select a particular answer, also known as IMCs) [27, 40], free-response text questions (survey responses are rejected if the answer is nonsensical, irrelevant, or clearly copy-pasted from the Internet), and CAPTCHAs. To allow us to evaluate the effect of these techniques on external validity, we added two additional questions to our survey: one reading-based attention check question and one free-response text question. We also required half of our participants (randomly selected) to successfully complete a CAPTCHA in order to submit the survey.

The full text of the survey can be found in Appendix A.

3.2 Datasets

We use four datasets: (1) a probabilistic dataset from the Pew Research Center panel [15], (2) a representative sample from Prolific (accurate to the US Census on age, sex, and race), (3) a gender-balanced sample from Prolific, and (4) a sample from Amazon Mechanical Turk (MTurk). We compensated online study participants \$1.50 for completing the survey, which we estimated to take 6 minutes. This was approved by the Institutional Review Boards of the authors’ institutions.

We additionally collected two filtered samples of under-represented populations that do not appear in the Pew demographic categories: Indigenous people and transgender people. We deployed surveys on Prolific using the platform’s prescreening filters to only allow participants in these demographics to take the study. Over a period of 8 days, we received responses from 79 Indigenous users on Prolific, and 197 transgender users.

1. **Pew American Trends Panel Wave 49.** This dataset ($n = 4,272$) was collected by Ipsos Public Affairs between June 3–17, 2019 on behalf of the Pew Research Center [15]. The weighted estimates for this sample are believed to accurate to ± 1.87 percentage points of the US population aged 18 and over. Pew Research Center typically makes survey data publicly available on their website two years after the data collection, so this dataset became publicly available in 2021.

Participants in this survey were a subset of Pew Research Center’s American Trends Panel (ATP) [14], a panel of more than 10,000 U.S. adults recruited and maintained by the Pew Research Center using state-of-the-art techniques.² This subset of the panel was chosen to be generally representative of the broader U.S. population; as this was a probabilistic survey, the resulting data was weighted to balance demographics to match

²Prior to 2018, panel participants were recruited at the end of a large, national, landline and cellphone random-digit-dial survey that was conducted in both English and Spanish. After 2018, ATP has relied on address-based recruitment to avoid the response-bias that has developed in telephone-based recruiting. It supports non-Internet connected participants by providing tablets that enable those people to take surveys.

the U.S. population (to compensate for any biases due to sampling and non-response).

Our analyses treat this dataset as the gold standard for U.S. responses to our survey questions.

- Prolific Representative Sample.** We sampled U.S. participants ($n = 800$) using Prolific’s representative sample feature. This sample is stratified on age, sex, and ethnicity based on the simplified U.S. census [47]. The median time to complete the survey was 6.1 minutes.
- Prolific Gender-Balanced Sample.** We sampled U.S. participants ($n = 800$) on Prolific, balanced on gender (50% male and 50% female). Prolific has been noted to have demographics that skew younger and more female [18]: currently within the U.S. sample space, there are over twice as many women on the platform as men. This survey took participants a median time of 5.3 minutes to complete. No participants are in both the Prolific representative and gender-balanced samples.
- MTurk Sample.** We collected a sample ($n = 800$) from Amazon Mechanical Turk, with participation restricted to people located in the U.S. who have completed over 50 HITs and have over 95% approval rate. We chose these filters as they are common practice for studies of this type deployed on the MTurk platform and believed to produce higher quality data [46, 49]. Participants took a median time of 5.2 minutes to complete the survey.

3.3 Analysis

We used chi-square proportion tests (χ^2) to compare response distributions. For each question, we ran χ^2 tests to compare the distribution of answers for each sample (Prolific representative, Prolific gender-Balanced, MTurk) pairwise against the Pew data. We also used Total Variation Distance (TVD) to quantify the distance between answer distributions in our surveys and the Pew data.

Total Variation Distance. Total Variation Distance (TVD), defined as $TVD(P, Q) = 1/2 \cdot \sum_i (P_i, Q_i)$, is a standard metric for quantifying the distance between two distributions [24]. Intuitively, it corresponds to the fraction of respondents who answer differently between the two samples. A TVD of 0 indicates that two distributions are identical; as the distributions become increasingly disjoint the TVD approaches 1.

To illustrate the concept of TVD, we show how to calculate the TVD between the distribution of responses to the first knowledge question for the Pew sample ($know1_{Pew}$) and responses for the Prolific gender-balanced sample ($know1_{Bal}$). In the Pew survey, .626 of the respondents answered correctly, .093 incorrectly, and .282 with “Not sure”; in our representative Prolific sample, the proportions for correct, incorrect, and

not sure were .866, .028, and .106 respectively. Therefore,

$$\begin{aligned} TVD(know1_{Pew}, know1_{Bal}) &= \frac{1}{2} (|.626 - .866| + |.093 - 0.028| + |.282 - .106|) \\ &= .2405 \end{aligned}$$

This TVD of .2405 shows that approximately one quarter of the responses were distributed differently between the Pew sample and the Prolific representative sample.

To show how we use TVD, consider a comparison between the survey questions `know1` and `exp5`. In both cases, a χ^2 test indicates a significant difference in answers between the Pew sample and the Gender-Balanced Prolific sample. To contextualize this result, we calculate TVD values for both pairs of distributions. Using the same definition as above, we find that $TVD(exp5_{Pew}, exp5_{Bal}) = .111$. The lower TVD for `exp5` provides evidence that the balanced Prolific sample may be closer to the Pew sample for the experience question (TVD = .111) than for the knowledge question (TVD = .2405).

We chose to use these two measures (χ^2 tests and TVD) as both have strengths and weaknesses in their ability to provide insights into the representativeness of these platforms’ participants. χ^2 tests with p -values provide a thresholded measure of sameness or difference, while TVD provides us with a limited but valuable continuous measure of distance. For example, when χ^2 tests show that answer distributions are statistically distinct for two question categories, TVD augments this analysis by providing a method of estimating whether the non-representativeness of one question category may be of larger magnitude than the other.

As prior work has found that online surveys were representative for populations under 50 years old or with at least a college level education [49], we further explored whether online survey platforms were more representative of certain demographic subsets in the general US population. In particular, we separately analyzed populations aged 18-29, 30-49, and 50+ from each of our samples against the corresponding demographic groups in the Pew dataset. We also conducted an analysis divided by education level, classified into one of three categories: high school graduate or less, some college, college graduate+.

Since the responses ranged from binary to multiple choice to Likert-scale, we did not attempt to code answers into binary variables. For most questions, we kept the response codings as presented to the user. For knowledge questions—which had one correct answer, 3-5 incorrect answers, and a “Not sure” option—we coded answers into three categories: Correct, Incorrect, Not sure. In the studies, participants were able to skip any question. As no more than 2.5% of any question had blank answers, we chose to impute the answers for non-response. Any skipped attention check was coded as a failed attention check. For knowledge questions, any skipped question was classified as “Not sure”. For all other questions, non-response

was classified into the most negative category (e.g. “Not at all confident”, “No, do not use this”). To validate this choice of imputation, we ran an analysis as well on imputation where non-response was classified as the most positive (e.g. “Very confident”, “Yes, use this”) and found that TVD remained ± 0.01 both within each category and overall.

To understand whether attention checks are effective in improving data quality, we ran an analysis wherein only responses of those who passed each of the three attention check questions were included. To determine efficacy of reading attention checks, we compared only those who passed the reading attention check (selected “Strongly agree”) against the Pew data. For the textbox attention check which asked users to define “digital privacy” in their own words, a researcher from our team coded all responses into accept, reject, and copy-paste. “Reject” responses referred to answers that either were nonsensical, unrelated to the question, or merely repeated the words “digital privacy”. “Copy-paste” indicated responses that were plausible definitions of “digital privacy”, but appeared verbatim 5 times or more throughout the sample or contained large chunks of text that were copied verbatim from these phrases. Under this coding, some other phrases were indeed often repeated, but were accepted if they appeared less than 5 times. A second researcher resolved cases where it was uncertain whether a response should be rejected. We removed all responses either coded as “Reject” or “Copy-paste” when analyzing samples that are said to have passed the textbox attention check. For the CAPTCHA analysis, for each of our samples, we ran analyses comparing those in the sample who saw and passed a CAPTCHA (around 50% for each sample) against the Pew dataset.

We conducted demographic raking using the R `anesrake` package, weighted by age, sex, education, and race to see if it would improve the generalizability of our samples. We used proportions from the 2017 American Community Survey [3] to match the demographic weighting used for the Pew dataset.

We were further interested in whether underrepresented groups had significantly different responses than the general population. We compared demographic groups from our Prolific representative sample to the same group from the Pew sample to investigate whether demographic groups on Prolific are representative of their respective group in the broader population. With our filtered samples of rare demographic subpopulations (Indigenous and transgender people), we compared our filtered sample to the Prolific representative sample.

3.4 Methodological Limitations

Due to the statistical constraints surrounding sample size and power, smaller sample sizes necessarily have less statistical power. Thus, for smaller samples (e.g., CAPTCHAs, underrepresented groups), we expect to find fewer instances of *statistical* significance (p -values < 0.05), implying that these samples more closely match the Pew dataset. However, this

Dem	Response	Pew (%)		Online Samples (%)		
		Raw	Wgt	Repr	Bal	MT
Age	18-29	16	20	23	45	22
	30-49	31	33	34	43	66
	50-64	31	26	30	9	10
	65+	23	20	12	2	2
Edu	HS or less	35	38	13	13	8
	Some College	28	31	34	34	26
	College grad+	38	30	53	52	66
Race	White	78	74	75	75	82
	Black	11	12	13	4	13
	Asian	3	4	7	12	3
	Mixed	4	5	4	7	2
	Other	4	5	2	3	1
Sex	Male	44	48	49	50	68
	Female	56	52	51	50	32

Table 1: Demographic characteristics of the four datasets. Since the Pew dataset was a probabilistic sample, the weighted dataset (Wgt) was analyzed. For the online samples—Prolific representative (Repr), Prolific gender-balanced (Bal) and MTurk (MT)—raw data was analyzed except where we explicitly state that raking was applied.

does not mean differences do not exist, but rather that they might be too slight to detect at lower sample sizes. Thus, we examine TVDs in conjunction with p -values in order to obtain a clearer picture rather than simply defaulting to p -values.

TVD is one of many measures that could be used to summarize our data and quantify distances between distributions. A primary limitation of TVD is that it does not account for whether the underlying data is categorical or ordinal, and thus on Likert-scale style questions, treats participant answers which differ by a small “amount” (e.g., from 1 to 2) identically from those that differ by a large “amount” (e.g., from 1 to 5). Similarly, like χ^2 tests, it cannot distinguish between different specific ways that categorical answer distributions differ. For example, TVDs for knowledge questions are large for both MTurk and Prolific representative (.30, .23), and χ^2 tests find that responses to all 8 knowledge questions are significantly different from Pew responses for both, yet the *direction* of these differences is opposite: MTurkers are more likely to be *incorrect* in their knowledge while Prolific respondents are more likely to be *correct*.

Other options for contextualizing results from surveys are possible, such as visualizations and tables of raw answer proportions, which are provided in our figures and in Appendix B. Qualitative work examining the reasons for and nuances of the differences we observe could provide another avenue of understanding how and why results between probabilistic surveys and online platforms differ and what implications these differences have for the validity of insights provided by usable privacy and security studies.

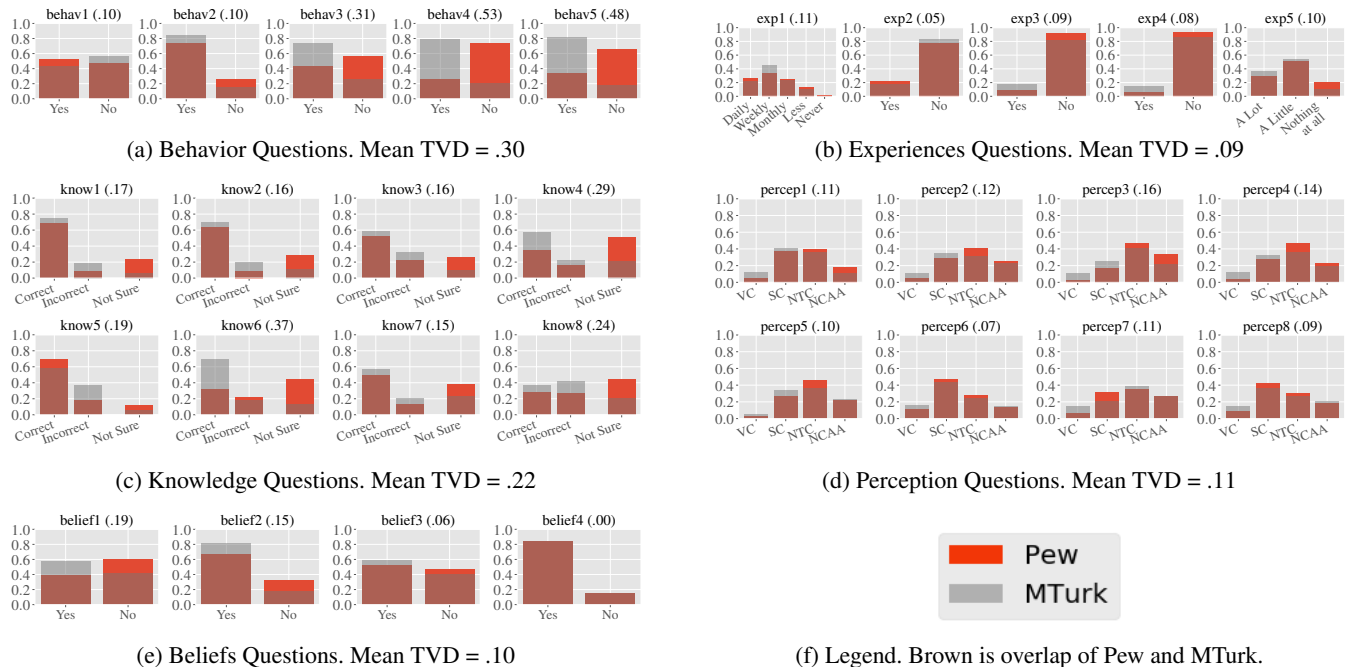


Figure 1: Distributions of responses to all questions for the Pew sample (weighted) and the MTurk sample (u50SC/Rak/freeAC). TVDs between the Pew sample and the MTurk sample are given in the captions.

We do not use corrections (e.g., Benjamini-Hochberg, Bonferroni) to analyze our results. These corrections control for Type I errors (false discovery rate) by limiting the number of erroneously statistically significant results (p -values < 0.05). However, we are attempting to find results that *do not* significantly differ between samples (i.e., $p > 0.05$), so these corrections could in fact overstate our results.

4 Results

We deployed three copies of our survey—Prolific representative sample, Prolific gender-balanced sample, and MTurk sample—in February 2022. We found that the Prolific representative survey took significantly longer to run; it took 49 hours to complete, compared to 2.5 hours for the Prolific gender-balanced survey and 2 hours for the MTurk survey.³ The Prolific representative survey also cost significantly more to deploy: collecting that sample cost \$2,784, compared to \$1,600.00 for the Prolific gender-balanced sample and \$1,682 for the MTurk sample. We then analyzed the responses to evaluate the external validity and data quality of the resulting samples. A summary of the demographics for each of the samples is provided in Table 1, and complete results are summarized in Appendix B.

³However, we note that since a significant percentage of our MTurk responses (39.1%) failed the free-response attention check, rejecting those responses and re-releasing those HITs would significantly increase the total deployment time; since less than 1% of responses failed that check for either Prolific sample, no extra time or effort would be required for those surveys.

4.1 The External Validity of MTurk

Our MTurk sample was heavily weighted toward younger participants (703/800 participants were under 50) and those with higher education (737/800 participants had at least some college education); this finding replicates prior work [32, 49]. However, while Redmiles et. al. [49] found that for the well-represented demographics (people under 50 or with at least some college) MTurker responses to privacy and security questions (about behavior, experiences, and knowledge) had high external validity, we were unable to replicate that result.

When we analyzed the raw MTurk sample, we found that responses collected through MTurk were extremely different from Pew (Table 2). We found statistically significant differences in responses for 29 of the 30 questions, and the overall average Total Variation Distance (TVD) was .29 (intuitively indicating that more than a quarter of the sample answered differently). We attempted to replicate prior work by also analyzing the sample that contained only people with under 50 or some college education and applied raking (i.e., demographic weighting). However, we still found statistically significant differences in responses for 29 questions, and the average TVD dropped only slightly (to .28).

Unlike the earlier work, we found that both raking and filtering out responses that failed a free-response text attention check question had significant effects on data quality. Combining these data quality measures with the demographic restrictions from prior work—i.e., restricting to people under 50 or with some college who passed the text attention check

Category	Raw Sample		u50SC/Rak/freeAC	
	TVD	$p < 0.05$	TVD	$p < 0.05$
Behavior	0.41	5/5	0.30	5/5
Experiences	0.27	5/5	0.09	5/5
Knowledge	0.30	8/8	0.22	8/8
Perceptions	0.30	8/8	0.11	8/8
Beliefs	0.15	3/4	0.10	3/4
Overall	0.29	29/30	0.17	29/30

Table 2: Measures of external validity for the MTurk sample. TVD indicates distance from the (weighted) probabilistic Pew sample. $p < .05$ shows the fraction of questions in each category for which the responses were statistically significantly different from the Pew sample. For both, lower is better.

and applying filtering, denoted u50SC/Rak/freeAC in Figure 1 and Table 2—produced the best-case results for the MTurk sample.

In this best-case scenario, 29 of the questions still had significantly different responses, but the average TVD went down to .17. In particular, responses about experiences, perceptions, and beliefs are somewhat generalizable: TVDs dropped to around .10, although χ^2 tests still showed that responses for most questions were still significantly different from Pew. However, knowledge questions were still very different: MTurkers were much more confident—and incorrect—on knowledge questions even after data quality measures were applied. Behavior questions—which focused on social media use—also remained very different: MTurkers were more likely to use Facebook (63% vs. 53%), Instagram (74% vs. 44%), Twitter (79% vs. 26%), and other social networks (82% vs 34%). Complete results for this best-case scenario are depicted in Figure 1 and measures of external validity are summarized in Table 2.

4.2 The External Validity of Prolific

Like the MTurk sample, the Prolific gender-balanced sample was heavily weighted towards younger participants and those with higher education; the age skew was more extreme and the education skew less. That sample also included significantly fewer Black participants and more Asian and mixed race participants. The Prolific representative sample was representative of the overall U.S. population for age and race, but showed the same skew towards higher education.

Overall, we found that both Prolific samples generalize better than the MTurk sample and that free-response attention checks were no longer critical for data quality. However, the external validity of the samples varied significantly depending on the type of question. Our results are shown in Figure 2, and measures of external validity are summarized in Table 3.

Behavior. Although responses about behavior from the Prolific representative sample were slightly more generaliz-

Samp.	Cat.	Raw Sample		u50SC/Rak/freeAC	
		TVD	$p < 0.05$	TVD	$p < 0.05$
Rep.	Behav.	0.22	4/5	0.19	3/5
Rep.	Exp.	0.07	5/5	0.06	5/5
Rep.	Know.	0.23	8/8	0.17	8/8
Rep.	Percep.	0.05	3/8	0.06	4/8
Rep.	Beliefs	0.07	3/4	0.06	2/4
Rep.	Overall	0.13	23/30	0.11	22/30
Bal.	Behav.	0.27	3/5	0.24	4/5
Bal.	Exp.	0.06	4/5	0.05	4/5
Bal.	Know.	0.24	8/8	0.16	7/8
Bal.	Percep.	0.05	4/8	0.07	7/8
Bal.	Beliefs	0.08	2/4	0.06	3/4
Bal.	Overall	0.14	21/30	0.12	25/30

Table 3: Measures of external validity for the Prolific samples. For both, lower is better. See Table 2 for more details.

able in terms of TVD (and both were more generalizable than the MTurk responses), neither of our Prolific samples demonstrated high external validity for behavior questions. Prolific participants were similarly likely to use Facebook compared to Pew participants but differed on other reported behavior, with the fraction of Prolific participants reporting that they use Instagram, Twitter, and “Other Social Media Sites” being 25%–54% higher compared to the Pew sample.

Experiences. Overall, both of our Prolific samples generalized well for questions about prior experiences. While most of those questions showed statistically significant differences, the magnitude of those difference was quite small (TVD = .07 for the representative sample and .06 for the gender-balanced sample).

Knowledge. Knowledge questions did not generalize well for either of our Prolific samples. Prolific respondents were more likely to provide *correct* answers and less likely to answer “Not sure”, suggesting that Prolific users are significantly more knowledgeable about privacy and security matters than the overall U.S. population.

Perceptions. Both Prolific samples had relatively high external validity for questions about perceptions of privacy and security. The Prolific representative sample had statistically different responses compared to the Pew sample for only 3/8 questions (4/8 for the gender-balanced sample), and TVDs between each sample and the Pew sample were small (about .05).

Beliefs. Both Prolific samples also had high external validity for questions about beliefs about privacy and security. While some of the questions were statistically distinguishable, the TVDs were low suggesting that the effect size was small.

We also analyzed the Prolific samples using the best-case data quality measures from our MTurk analysis—restriction to people under 50 or with some college who passed the text attention check and applying filtering, denoted

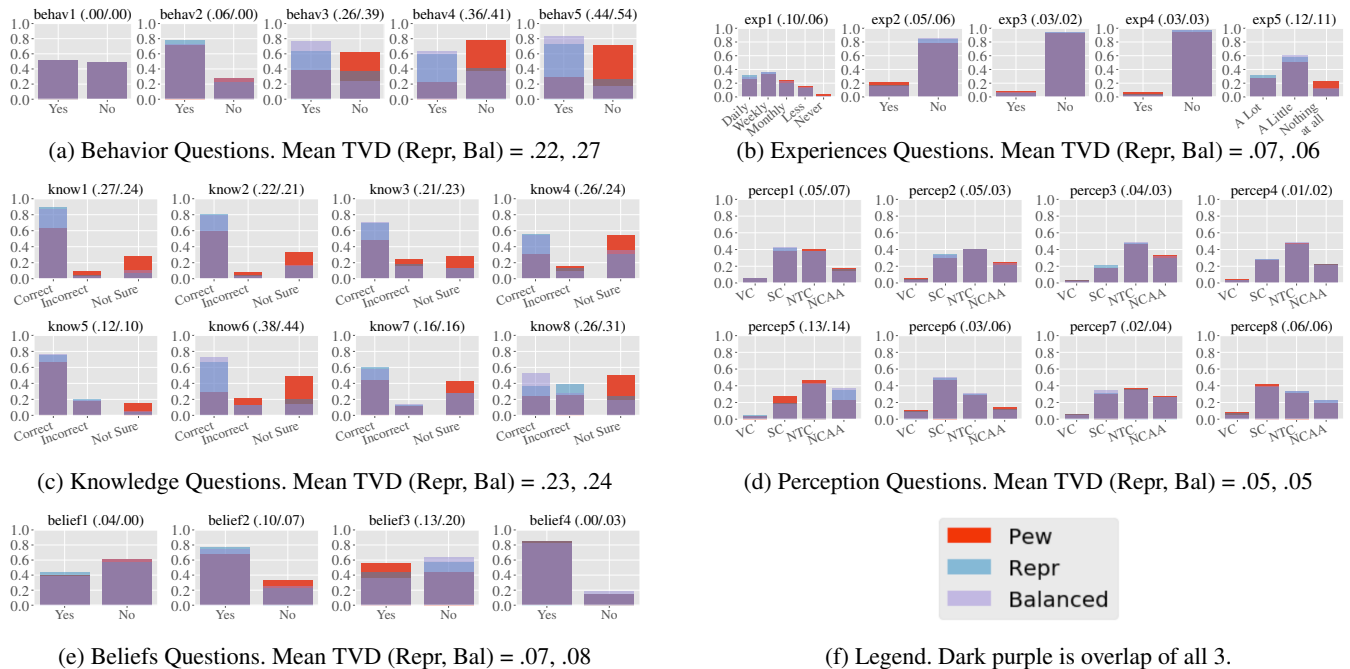


Figure 2: Distributions of responses to all questions for the Pew sample (weighted), Prolific representative sample (raw), and Prolific gender-balanced sample (raw). TVDs between the Pew sample and the Prolific samples are given in the captions.

u50SC/Rak/freeAC in Table 3). Overall, TVDs decreased slightly. However, unlike for the MTurk sample, this analysis did not dramatically improve the generalizability of the Prolific samples.

4.3 Data Quality Measures

We evaluated four data quality measures: reading-based attention checks, free-response text attention checks, CAPTCHAs, and raking. For the MTurk sample, we found that a free-response attention check (which 39.1% of responses failed) and raking both significantly improved data quality for the MTurk sample. Despite the data quality issues with MTurk, neither reading-based attention checks nor CAPTCHAs (which no respondents failed) significantly improved data quality. Although Prolific respondents did slightly less well than MTurkers on our reading attention check question (7.75%-8.25% failed), none of our data quality measures significantly improved data quality for the Prolific samples.

4.4 Beyond the “Average” User

While the standard metric of external validity is the extent to which results generalize to the overall population, overall generalizability does not necessarily imply that results are valid across all subgroups. We therefore also examine the question of how well our results generalize for various demographics subpopulations. We apply two analysis techniques: (1) we compare demographic slices from our online

samples to the corresponding demographic slices of the Pew sample, an approach that parallels the investigations of MTurk generalizability by Redmiles et al. [49] and (2) we compare demographic slices from rarer, traditionally understudied subpopulations to the overall population.

4.4.1 Prolific vs. Pew Demographic Subpopulations

Since Pew is our gold standard, we can perform this analysis only for demographic variables reported by Pew, and only for values of those variables which occur sufficiently frequently in the population to enable meaningful comparison. Based on these limitations, we choose to analyze two racial subpopulations (Black and Asian American), educational attainment, and age. Table 1 presents the numbers of people in each of these categories. Because the sample sizes are inherently smaller for these subpopulations than for the overall population, we focus our analysis exclusively on distance between the distribution of responses provided in the online surveys and the (weighted) distribution of responses to the Pew survey instead of considering p -values or the number of questions with statistically different responses.

Overall, we found that the Prolific representative sample tends to be the best of all three collected samples for each of these demographic brackets (although the Prolific gender-balanced sample is often nearly as good) and that the Prolific samples generalize better for younger and for more highly-educated subpopulations.

Race. Overall, the Prolific representative sample is almost as representative of the Black and Asian subpopulations as it is for the overall population. Average TVDs measuring the representativeness Black and Asian subpopulations are about .01–.02 higher on average than TVDs for the full dataset.

Age. For Prolific, we found that as ages increase, the samples become less generalizable. For people age 18-29, both Prolific samples are fairly representative of the general U.S. population, with particular improvement for knowledge questions (TVD = .15) and behavior questions (TVD = .16). Within both 18-29 and 30-40 age brackets, the Prolific samples actually generalize to the Pew dataset better than comparing the full datasets. By contrast, generalizability for people over 50 is worse, particularly for knowledge questions (TVD: Repr = .28, Bal = .32) and behavior questions as older Prolific users demonstrate significantly more knowledge of privacy and security and significantly higher levels of technology use. For our MTurk sample, both the 18-29 and the over 50 subpopulations were more generalizable than the full dataset, although they still had lower data quality than the corresponding slices of the Prolific samples.

Education. For our Prolific samples, we found that as education increases the samples become more generalizable. For respondents with a high school education or less, Prolific samples are less generalizable for this demographic slice than for the overall population, with participants reporting particularly higher levels of technology use. For respondents who are college graduates, both Prolific samples are reasonably representative of the overall population of U.S. college graduates (TVD: Repr = .10, Bal = .12), with more representative responses to knowledge questions (TVD: Repr = .13, Bal = .14). Conversely, the generalizability of the MTurk sample for people with high school education or less did not decrease (although the data quality remained worse than the Prolific samples) while data quality does decrease slightly for the subpopulation with Bachelor’s degrees (TVD = .30).

4.4.2 Rare Demographic Subpopulations

Finally we identified two populations with relatively low representation on Prolific—Indigenous people and transgender people—and explored (1) how effective Prolific’s filters are at producing large samples of rare (and frequently understudied) subpopulations and (2) to what extent generalizable results for the overall population apply to these subpopulations.

Indigenous People. Our filtered sample of Indigenous people ran for 8 days and obtained 79 responses during that time, an average of about 10 people per day. For context, at the time that we launched this filtered sample, Prolific reported that there were 294 Indigenous respondents who had been active in the past 90 days.

Comparing the distributions of responses of these 79 respondents to our full Prolific representative sample, we found that variations were relatively small, with Indigenous peo-

ple on Prolific being generally somewhat similar to other people completing surveys on Prolific. Comparable to the difference between our Prolific representative sample and Pew on the most representative question categories, mean TVDs comparing Indigenous respondents and the general Prolific population were under .10 for all question categories.

We emphasize that given the small size of this sample, we are unable to make conclusive statements about trends among Indigenous people on Prolific. Generally speaking, our data supports the claim that Indigenous people are more similar than different to other Prolific users, with TVDs between .04 and .05 for 3 question categories (Experiences, Perceptions, Beliefs). In terms of behaviors, they are more likely to use all social networks, including especially Facebook (TVD = .16) and other social networks (TVD = .12). Indigenous people in our sample appear to be slightly more likely to answer “Not Sure” to knowledge questions. No other clear trends emerge in how Indigenous respondents on Prolific are different from other respondents on Prolific in terms of experiences, perceptions, or beliefs.

Transgender People. Our filtered sample of transgender people ran for 8 days and obtained 197 responses during that time, an average of about 25 people per day. At the time that we launched this filtered sample, Prolific reported that there were 1231 transgender respondents who had been active in the past 90 days.

Comparing the distributions of responses of these 197 respondents to our full Prolific representative sample, we find that variations are small to moderate, with transgender people on Prolific being generally somewhat similar to other people completing surveys on Prolific. Mean TVDs comparing transgender Prolific respondents to the overall Prolific representative sample were under .12 for all question categories.

Although the low sample size precludes definitive findings, our data for this subpopulation provides preliminary evidence of some potential interesting trends. In terms of behavior, transgender people were less likely to use Facebook, and more likely to use Instagram, Twitter, and other social networks, than the Prolific Representative population. Transgender people were also more knowledgeable than Prolific participants overall, answering 6/8 knowledge questions correctly more often. Notably, transgender people were particularly more likely to understand how private browsing works, with a very large TVD of .26 distinguishing them as much more likely to answer `know8` correctly and much less likely to answer incorrectly or with “Not Sure”. This result might be due to the need for transgender people to use private browsing mode to protect themselves and their browsing habits from local adversaries, such as family, while seeking community, gathering information, and engaging in activism online [37].

Transgender people were also consistently more likely (TVD .07-.13) to select “Not Confident At All” in response to perception questions that asked about confidence that companies will follow their privacy policies, promptly notify about

data breaches, publicly admit mistakes that lead to privacy breaches, use personal information in appropriate ways, and be held accountable by the government for privacy missteps. Finally, they were slightly more likely to believe that people should have the right to have various personal information removed from public search results, with a particularly large (TVD = .16) increase compared to the Prolific representative sample in the likelihood that transgender people would say that people should be able to have “Negative media coverage” about themselves removed from public search results. We hypothesize that this might emerge from the likelihood of transgender people to experience media coverage about them as negative, for example if articles misgender them or include out-of-date personal details such as deadnames.

5 Discussion

While our results quantify the external validity of online surveys about privacy and security, they also provide insight into best practices for conducting online studies in this space.

Recommendation 1: *We recommend preferring Prolific to MTurk when recruiting participants for privacy and security surveys.*

Overall, we found major degradation of MTurk data quality and external validity since 2017 [49]. If MTurk samples are used, applying the data quality measures studied in this paper—including demographic raking and a stringent open textbox attention check—is critical to enhance external validity. However, even when applying these data quality measures to MTurk data, Prolific gender-balanced samples provide better validity and their use is recommended at the current time. It is important to remember, however, that both Prolific and MTurk samples better represent younger and more educated populations. Additionally, online samples appear to be differently representative for different types of questions, as we discuss below in the Recommendation 3.

Future work should examine the validity of samples from other platforms, which could be comparable to or better than Prolific. For example, we note that CloudResearch, which uses the MTurk population, has been found to provide similar data quality to Prolific when the default data quality filters are applied [38]. Although our literature review did not find any papers that used CloudResearch, it provides an alternative platform for recruiting participants in the future. Future work should also continue to monitor the external validity of MTurk and Prolific, as population demographics and data quality may continue to change over time.

Recommendation 2: *Determinations about whether to use Prolific’s representative sample feature can make trade-offs between generalizability and logistical constraints without significantly impacting data quality for most studies.*

We find that Prolific’s representative sample feature produces data that most closely matches the results from the nationally representative sample from the Pew dataset. However, the representative sample takes much longer (49 hours vs. 2.5 hours for 800 responses) and is significantly more expensive (\$2784 vs. \$1600) to deploy than collecting a gender-balanced sample of the same size from Prolific. In most cases, a Prolific gender-balanced sample performs nearly as well as a representative sample, with less than .02 difference in average TVD across all question categories. The largest gains for representative over gender-balanced were for behavior questions, for which neither was representative. All other question categories had very small differences (TVD < .01) between representative and gender-balanced samples.

Recommendation 3: *Be cautious when drawing conclusions from online studies about privacy and security knowledge or social media use, as these results might not be representative of the overall U.S. population.*

None of our online samples were representative of the overall population for knowledge questions—which posed factual questions about privacy and security topics—or behavior questions—which were dominated by questions about rates of social media use. We recommend that researchers take care in designing studies and interpreting data which depend on these properties of respondents. Similar to prior work, we do find that the younger and more highly educated the population, the more Prolific is representative, particularly for knowledge questions, which drop from TVDs of .28 for those over 50 to .15 for those 18-29, and from .27 for those with high school degrees or less education to .13 for those with college degrees. Even these TVDs are quite high, however, indicating that 15-13% of responses are different than they would be for that actual age or education range in an census-representative sample, and so we still urge caution in relying on such data.

Our results show that participants recruited through MTurker and Prolific are more confident about privacy and security knowledge compared to the overall U.S. population, with fewer respondents answering “Not sure” to most questions. We observe that this is a similar phenomenon to past results which found that MTurk workers are more certain about what information is available about them online [32]. This confidence gap also raises questions about the generalizability of non-survey studies that recruit participants through these online platforms, since prior work has found that confidence is a better predictor of security behaviors than actual knowledge [51].

One factor that might have contributed to the drastically higher reported use of social media in our online samples is response bias from participants who worry that they may be excluded from a survey (and thus not be paid) if they don’t use certain products, leading them inaccurately claim that

they use social networks which they actually do not. Another possibility is that the population on these platforms have different behaviors regarding social media than the general U.S. population, leading to higher adoption and use of social media platforms. Indeed, prior work on MTurk has also found that U.S. MTurk workers have higher reported social media use than the general US population [32], which is supported by our findings in our Prolific and MTurk samples.

While it is possible that some of the difference in responses to behavior questions might have been due to actual differences in social media use between 2019 and 2022, a 2021 survey [4] conducted by Pew about a year into the pandemic shows that social media adoption has not drastically increased since mid-2019, when the American Trends Panel Wave 49 was conducted. That survey found that in 2021, 69% of Americans used Facebook, 40% used Instagram, and 23% used Twitter, numbers which are very closely compatible with the 71%, 38%, and 23% found in 2019. This suggests that the higher usage numbers we find in our online samples are genuine symptoms that users of online crowdsourcing platforms are not representative of the overall population in terms of their social media use.

Recommendation 4: *Attention Check Questions and CAPTCHAs are not recommended for online surveys conducted on Prolific.*

We do not recommend reading attention checks (Instructional Manipulation Checks), textbox attention checks, and CAPTCHAs when collecting survey responses on Prolific. Our reading attention check was failed by 66/800 (representative sample) and 62/800 (gender-balanced sample) participants, but data quality was not improved by analyzing the data with these responses removed (see Section 4.3). Prolific users almost never fail textbox attention checks (7/1600 failures) or CAPTCHAs (0/1600 failures). Based on our results, using such attention check questions lengthens surveys unnecessarily. Using IMCs might also change participants interpretation of subsequent questions [29].

Recommendation 5: *Raking is not currently necessary when analyzing the results of online privacy and security studies.*

Raking is often used in survey methods that intend to be representative of the general population since perfect response rates from demographic groups cannot be ensured by any sampling approach. Although we found raking had little effect on the representativeness of either of our Prolific samples (Section 4.3), studies in other disciplines have seen success in using raking for MTurk survey data to better generalize to the US population [53], and we would recommend researchers consider it. However, we note that raking also requires decisions on which demographic variables to weight on and might differ for different questions and fields of study.

Recommendation 6: *Special care should be taken to include a diverse population of study participants, particularly for demographics that are rare or underrepresented on crowdsourcing platforms.*

Prior work has noted that online platforms tend to be younger, more highly educated, and more white than the general U.S. population [49]. This demographic imbalance could then lead to a fallacy of the “average” user on such platforms not being at all representative of the general population. Groups that do not make up the majority might also have significantly different preferences than the “general population”. For example, participants from racial minorities were more unsure about their security knowledge than the general population, and transgender people had lower confidence in companies taking responsible action regarding privacy issues than the general Prolific population. Therefore, we encourage researchers to consider specifically sampling underrepresented populations to understand possibly divergent privacy and security perceptions and backgrounds to avoid over-general interventions and claims that could contribute to further marginalizing already marginalized populations.

Study Limitations. As we limited our studies to participants located in the United States, we cannot make claims as to whether Prolific is similarly representative of other jurisdictions or of the overall global population. Indeed, prior work has also found differences between privacy attitudes between MTurk workers located in the U.S. and in India [32]. Additionally, while the Pew dataset was weighted on myriad strata of demographics to best represent the adult U.S. population, we still recognize that it is not perfectly representative of the general US population. As with all surveys, there might be non-response bias, even with the most carefully selected probabilistic sample. In other words, just as those who choose to take surveys on online platforms differ from the general population in terms of tech familiarity and use, so too might probabilistic studies vary from exact national preferences. Therefore, though we use Pew as our gold standard, we cannot guarantee that it is a perfect representation of the preferences of all Americans.

6 Conclusion

Online crowdsourced samples are an important source of data for usable privacy and security survey research today. Understanding the external validity of these samples is critical to ensuring that the results from such research generalize and can appropriately guide individuals, technologists, lawmakers, and regulators. Our work evaluates the external validity two popular crowdsourcing sites—MTurk and Prolific—and provides recommendations about best practices for conducting security and privacy surveys on these platforms.

Acknowledgments

We are grateful to Cassandra Pattanayak and Elissa Redmiles for their advice and support on this work. This work was supported by the National Science Foundation (CNS Award #1948344) and Wellesley College.

References

- [1] Desiree Abrokwa, Shruti Das, Omer Akgul, and Michelle L. Mazurek. Comparing Security and Privacy Attitudes Among U.S. Users of Different Smartphone and Smart-Speaker Platforms. In *17th Symposium on Usable Privacy and Security*, pages 139–158, 2021.
- [2] Amazon mechanical turk. <https://www.mturk.com>.
- [3] American communities survey. <https://www.census.gov/programs-surveys/acs/data.html>, 2017.
- [4] Brooke Auxier and Monica Anderson. Social media use in 2021. *Pew Research Center*, 1:1–4, 2021.
- [5] Hui Bai. Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. <https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random>, 2018.
- [6] Daniel V. Bailey, Philipp Markert, and Adam J. Aviv. "I have no idea what they're trying to accomplish:" Enthusiastic and Casual Signal Users' Understanding of Signal PINs. In *17th Symposium on Usable Privacy and Security*, pages 417–436, 2021.
- [7] David G. Balash, Dongkun Kim, Darika Shaibekova, Rahel A. Fainchtein, Micah Sherr, and Adam J. Aviv. Examining the Examiners: Students' Privacy and Security Perceptions of Online Proctoring Services. In *17th Symposium on Usable Privacy and Security*, pages 633–652, 2021.
- [8] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. Comparing the similarity of responses received from studies in amazon's mechanical turk to studies conducted online and with direct recruitment. *PLoS one*, 10(4):e0121595, 2015.
- [9] Tara Behrend, David Sharek, Adam Meade, and Eric Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800–813, 2011.
- [10] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3):739–753, 2014.
- [11] Michael D. Buhrmester, Sanaz Talaifar, and Samuel D. Gosling. An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13:149–154, 2018.
- [12] Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. Automated Extraction and Presentation of Data Practices in Privacy Policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110, April 2021.
- [13] Jason Ceci, Hassan Khan, Urs Hengartner, and Daniel Vogel. Concerned but Ineffective: User Perceptions, Methods, and Challenges when Sanitizing Old Devices for Disposal. In *17th Symposium on Usable Privacy and Security*, pages 455–474, 2021.
- [14] Pew Research Center. American trends panel. <https://www.pewresearch.org/our-methods/u-s-surveys/the-american-trends-panel/>.
- [15] Pew Research Center. American trends panel wave 49. <https://www.pewresearch.org/internet/dataset/american-trends-panel-wave-49/>, June 2019.
- [16] Jesse Chandler, Cheskie Rosenzweig, Aaron J. Moss, Jonathan Robinson, and Leib Litman. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods*, 51(5):2022–2038, 2019.
- [17] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, Somesh Jha, and Suman Banerjee. Face-Off: Adversarial Face Obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2021(2):369–390, April 2021.
- [18] Nick Charalambides. We recently went viral on TikTok - here's what we learned. <https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned>, August 2021.
- [19] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. "I would have to evaluate their objections": Privacy tensions between smart home device owners and incidental users. *Proceedings on Privacy Enhancing Technologies*, 2021(4):54–75, October 2021.
- [20] Matthew J.C. Crump, John V. McDonnell, and Todd M. Gureckis. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS one*, 8(3):e57410, 2013.

- [21] Anastasia Danilova, Alena Naiakshina, Anna Rasgauski, and Matthew Smith. Code Reviewing as Methodology for Online Security Studies with Developers - A Case Study with Freelancers on Password Storage. In *17th Symposium on Usable Privacy and Security*, pages 397–416, 2021.
- [22] Pardis Emami-Naeini, Tiona Francisco, Tadayoshi Kohno, and Franziska Roesner. Understanding Privacy Attitudes and Concerns Towards Remote Communications During the COVID-19 Pandemic. In *17th Symposium on Usable Privacy and Security*, pages 695–714, 2021.
- [23] Kelsey R. Fulton, Anna Chan, Daniel Votipka, Michael Hicks, and Michelle L. Mazurek. Benefits and Drawbacks of Adopting a Secure Programming Language: Rust as a Case Study. In *17th Symposium on Usable Privacy and Security*, pages 597–616, 2021.
- [24] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.
- [25] Joseph K Goodman, Cynthia E Cryder, and Amar Cheema. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.
- [26] Thomas Groß. Validity and Reliability of the Scale Internet Users’ Information Privacy Concerns (IUIPC). *Proceedings on Privacy Enhancing Technologies*, 2021(2):235–258, April 2021.
- [27] David Harborth and Alisa Frik. Evaluating and Redefining Smartphone Permissions with Contextualized Justifications for Mobile Augmented Reality Apps. In *17th Symposium on Usable Privacy and Security*, pages 513–534, 2021.
- [28] Ayako A. Hasegawa, Naomi Yamashita, Mitsuaki Akiyama, and Tatsuya Mori. Why they ignore English emails: The challenges of non-native speakers in identifying phishing emails. In *17th Symposium on Usable Privacy and Security*, pages 319–338, 2021.
- [29] David J Hauser and Norbert Schwarz. It’s a trap! Instructional manipulation checks prompt systematic thinking on “tricky” tasks. *Sage Open*, 5(2), 2015.
- [30] David J Hauser and Norbert Schwarz. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48(1):400–407, 2016.
- [31] Maximilian Häring, Eva Gerlitz, Christian Tiefenau, Matthew Smith, Dominik Wermke, Sascha Fahl, and Yasemin Acar. Never ever or no matter what: Investigating adoption intentions and misconceptions about the Corona-Warn-App in Germany. In *17th Symposium on Usable Privacy and Security*, pages 77–98, 2021.
- [32] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy attitudes of mechanical turk workers and the U.S. public. In *10th Symposium On Usable Privacy and Security*, pages 37–49, 2014.
- [33] Ankit Kariryaa, Gian-Luca Savino, Carolin Stellmacher, and Johannes Schöning. Understanding Users’ Knowledge about the Privacy and Security of Browser Extensions. In *17th Symposium on Usable Privacy and Security*, pages 99–118, 2021.
- [34] Smirity Kaushik, Yaxing Yao, Pierre Dewitte, and Yang Wang. "How I know for sure": People’s perspectives on Solely Automated Decision-Making (SADM). In *17th Symposium on Usable Privacy and Security*, pages 159–180, 2021.
- [35] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *17th Symposium on Usable Privacy and Security*, pages 299–318, 2021.
- [36] Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. Digital Inequality Through the Lens of Self-Disclosure. *Proceedings on Privacy Enhancing Technologies*, 2021(3):373–393, July 2021.
- [37] Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. Privacy and activism in the transgender community. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
- [38] Leib Litman, Aaron Moss, Cheskie Rosenzweig, and Jonathan Robinson. Reply to MTurk, Prolific or panels? Choosing the right audience for online research. <https://ssrn.com/abstract=3775075>, 2021.
- [39] Aaron Moss and Leib Litman. After the bot scare: Understanding what’s been happening with data collection on MTurk and how to stop it. <https://www.cloudrsearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/>, 2018.
- [40] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4):867–872, 2009.

- [41] Stefan Palan and Christian Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, Mar 2018.
- [42] Gabriele Paolacci and Jesse Chandler. Inside the turk: Understanding mechanical turk as a participant pool. *Current directions in psychological science*, 23(3):184–188, 2014.
- [43] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.
- [44] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.
- [45] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evenden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, pages 1–20, 2021.
- [46] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4):1023–1031, 2014.
- [47] Prolific. <https://www.prolific.co>.
- [48] HIRAK Ray, Flynn Wolf, Ravi Kuber, and Adam J. Aviv. “Warn Them” or “Just Block Them”? Investigating Privacy Concerns Among Older and Working Age Adults. *Proceedings on Privacy Enhancing Technologies*, 2021(2):27–47, April 2021.
- [49] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *IEEE Symposium on Security and Privacy*, pages 1326–1343, 2019.
- [50] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. 2010.
- [51] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *CHI Conference on Human Factors in Computing Systems*, pages 2202–2214, 2017.
- [52] Sebastian Schnorf and Aaron Sedley. A comparison of six sample providers regarding online privacy benchmarks. In *Symposium on Usable Privacy and Security*, 2014.
- [53] Daniel J Simons and Christopher F Chabris. Common (mis) beliefs about memory: A replication and comparison of telephone and mechanical turk survey methods. *PloS one*, 7(12):e51876, 2012.
- [54] Daniel Smullen, Yaxing Yao, Yuanyuan Feng, Norman Sadeh, Arthur Edelstein, and Rebecca Weiss. Managing Potentially Intrusive Practices in the Browser: A User-Centered Perspective. *Proceedings on Privacy Enhancing Technologies*, 2021(4):500–527, October 2021.
- [55] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. How to do better with gender on surveys: A guide for HCI researchers. *Interactions*, 26(4):62–65, 2019.
- [56] Neil Stewart, Christoph Ungemach, Adam JL Harris, Daniel M Bartels, Ben R Newell, Gabriele Paolacci, and Jesse Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making*, 10(5):479–491, 2015.
- [57] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. Awareness, Adoption, and Misconceptions of Web Privacy Tools. *Proceedings on Privacy Enhancing Technologies*, 2021(3):308–333, July 2021.
- [58] Christian Stransky, Dominik Wermke, Johanna Schrader, Nicolas Huaman, Yasemin Acar, Anna Lena Fehlhaber, Miranda Wei, Blase Ur, and Sascha Fahl. On the Limited Impact of Visualizing Encryption: Perceptions of E2E Messaging Security. In *17th Symposium on Usable Privacy and Security*, pages 437–454, 2021.
- [59] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Deciding on Personalized Ads: Nudging Developers About User Privacy. In *17th Symposium on Usable Privacy and Security*, pages 573–596, 2021.
- [60] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. Defining Privacy: How Users Interpret Technical Terms in Privacy Policies. *Proceedings on Privacy Enhancing Technologies*, 2021(3):70–94, July 2021.
- [61] Andrew J Thompson and Justin T Pickett. Are relational inferences from crowdsourced and opt-in samples generalizable? Comparing criminal justice attitudes in the GSS and five online samples. *Journal of Quantitative Criminology*, 36(4):907–932, 2020.

- [62] Jan Tolsdorf, Florian Dehling, Delphine Reinhardt, and Luigi Lo Iacono. Exploring mental models of the right to informational self-determination of office workers in Germany. *Proceedings on Privacy Enhancing Technologies*, 2021(3):5–27, July 2021.
- [63] United States Food and Drug Administration. Collection of race and ethnicity data in clinical trials: Guidance for industry and food and drug administration staff. <https://www.fda.gov/media/75453/download>, 2016.
- [64] Rick Wash, Norbert Nthala, and Emilee Rader. Knowledge and capabilities that non-expert users bring to phishing detection. In *17th Symposium on Usable Privacy and Security*, pages 377–396, 2021.
- [65] Shikun Zhang, Yuanyuan Feng, Lujo Bauer, Lorrie Faith Cranor, Anupam Das, and Norman Sadeh. “Did you know this camera tracks your mood?”: Understanding Privacy Expectations and Preferences in the Age of Video Analytics. *Proceedings on Privacy Enhancing Technologies*, 2021(2):282–304, April 2021.
- [66] Leah Zhang-Kennedy and Sonia Chiasson. “Whether it’s moral is a whole other story”: Consumer perspectives on privacy regulations and corporate data practices. In *17th Symposium on Usable Privacy and Security*, pages 197–216, 2021.

A Survey Questions

This appendix contains the list of questions asked during our user study. These questions are taken from the Pew American Trends Panel run between June 3-17, 2019. Each question could be skipped by the user.

1. Do you use any of the following social media sites?
The order of the first three of the following questions are randomized
 - (a) Facebook [*behav2*]
 - Yes, use this / No, do not use this
 - (b) Instagram [*behav3*]
 - Yes, use this / No, do not use this
 - (c) Twitter [*behav4*]
 - Yes, use this / No, do not use this
 - (d) Any other social media sites [*behav5*]
 - Yes, use this / No, do not use this
2. In your own words, what does “digital privacy” mean to you?
Participants are given a textbox to type in.
3. How often are you asked to agree to the terms and conditions of a company’s privacy policy? [*exp1*]

- Almost daily / About once a week / About once a month / Less frequently / Never
4. How confident are you, if at all, that companies will do the following things?
The order of the following questions are randomized
 - (a) Follow what their privacy policies say they will do with your personal information [*percep1*]
 - Very confident / Somewhat confident / Not too confident / Not confident at all
 - (b) Promptly notify you if your personal data has been misused or compromised [*percep2*]
 - Very confident / Somewhat confident / Not too confident / Not confident at all
 - (c) Publicly admit mistakes and take responsibility when they misuse or compromise their users’ personal data [*percep3*]
 - Very confident / Somewhat confident / Not too confident / Not confident at all
 - (d) Use your personal information in ways you will feel comfortable with [*percep4*]
 - Very confident / Somewhat confident / Not too confident / Not confident at all
 - (e) Be held accountable by the government if they misuse or compromise your data [*percep5*]
 - Very confident / Somewhat confident / Not too confident / Not confident at all
 5. How comfortable are you, if at all, with companies using your personal data in the following ways?
The order of the first, second, and last questions are randomized
 - (a) To help improve their fraud prevention systems [*percep6*]
 - Very comfortable / Somewhat comfortable / Not too comfortable / Not comfortable at all
 - (b) Sharing it with outside groups doing research that might help improve society [*percep7*]
 - Very comfortable / Somewhat comfortable / Not too comfortable / Not comfortable at all
 - (c) This question is not part of the survey and just helps us to detect bots and automated scripts. To confirm that you are a human, please choose ‘Strongly agree’ here.
 - Strongly disagree / Disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Agree / Strongly Agree

- (d) To help them develop new products *[percep8]*
- Very comfortable / Somewhat comfortable / Not too comfortable / Not comfortable at all
6. Have you recently decided NOT to use a product or service because you were worried about how much personal information would be collected about you? *[behav1]*
- Yes, have done this / No, have not done this
7. Here's a different kind of question. (If you don't know the answer, select "Not sure.") As far as you know... *The order of the following questions is randomized. For each question, the order of the first four options is randomized.*
- (a) If a website uses cookies, it means that the site... *[know1]*
- Can see the content of all the files on the device you are using
 - Is not a risk to infect your device with a computer virus
 - Will automatically prompt you to update your web browser software if it is out of date
 - Can track your visits and activity on the site *[correct]*
 - Not sure
- (b) Which of the following is the largest source of revenue for most major social media platforms? *[know2]*
- Exclusive licensing deals with internet service providers and cellphone manufacturers
 - Allowing companies to purchase advertisements on their platforms *[correct]*
 - Hosting conferences for social media influencers
 - Providing consulting services to corporate clients
 - Not sure
- (c) When a website has a privacy policy, it means that the site... *[know3]*
- Has created a contract between itself and its users about how it will use their data *[correct]*
 - Will not share its users' personal information with third parties
 - Adheres to federal guidelines about deceptive advertising practices
 - Does not retain any personally identifying information about its users
 - Not sure
- (d) What does it mean when a website has "https://" at the beginning of its URL, as opposed to "http://" without the "s"? *[know4]*
- Information entered into the site is encrypted *[correct]*
 - The content on the site is safe for children
 - The site is only accessible to people in certain countries
 - The site has been verified as trustworthy
 - Not sure
- (e) Where might someone encounter a phishing scam? *[know5]*
- In an email
 - On social media
 - In a text message
 - On a website
 - All of the above *[correct]*
 - None of the above
 - Not sure
- (f) Which two companies listed below are both owned by Facebook? *[know6]*
- Twitter and Instagram
 - Snapchat and WhatsApp
 - WhatsApp and Instagram *[correct]*
 - Twitter and Snapchat
 - Not sure
- (g) The term "net neutrality" describes the principle that... *[know7]*
- Internet service providers should treat all traffic on their networks equally *[correct]*
 - Social media platforms must give equal visibility to conservative and liberal points of view
 - Online advertisers cannot post ads for housing or jobs that are only visible to people of a certain race
 - The government cannot censor online speech
 - Not sure

- (h) Many web browsers offer a feature known as “private browsing” or “incognito mode.” If someone opens a webpage on their computer at work using incognito mode, which of the following groups will NOT be able to see their online activities? [know8]
- The group that runs their company’s internal computer network
 - Their company’s internet service provider
 - A coworker who uses the same computer [correct]
 - The websites they visit while in private browsing mode
 - Not sure
8. Do you think that ALL Americans should have the right to have the following information about themselves removed from public online search results?
The order of the following questions is randomized
- (a) Data collected by law enforcement, such as criminal records or mugshots [belief1]
- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches
- (b) Information about their employment history or work record [belief2]
- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches
- (c) Negative media coverage [belief3]
- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches
- (d) Potentially embarrassing photos or videos [belief4]
- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches
9. Today it is possible to take personal data about people from many different sources – such as their purchasing and credit histories, their online browsing or search behaviors, or their public voting records – and combine them together to create detailed profiles of people’s potential interests and characteristics. Companies and other organizations use these profiles to offer targeted advertisements or special deals, or to assess how risky people might be as customers. Prior to today, how much had you heard or read about this concept? [exp5]
- A lot / A little / Nothing at all
10. In the last 12 months, have you had someone. . .
The order of the following questions is randomized
- (a) Put fraudulent charges on your debit or credit card [exp2]
- Yes / No
- (b) Take over your social media or email account without your permission [exp3]
- Yes / No
- (c) Attempt to open a line of credit or apply for a loan using your name [exp4]
- Yes / No
11. What is your age?
- 18-29
 - 30-49
 - 50-64
 - 65+
12. What is your sex?
- Male
 - Female
13. Please indicate your highest level of education
- Less than high school
 - High school graduate
 - Some college, no degree
 - Associate’s degree
 - College graduate/some post grad
 - Postgraduate
14. Choose the race that you consider yourself to be
The first four options are presented in randomized order
- White
 - Black or African American
 - Asian or Asian American
 - Mixed Race
 - Some other race

B Survey Response Summaries

Table 4 and Table 5 summarize how our four datasets compare on each of the thirty individual questions. Responses are within $\pm 5\%$ Pew proportions are highlighted in green; responses are $\geq 10\%$ off from Pew proportions are highlighted in orange.

Q	Ans	Pew	Repr	Bal	MTurk
behav1	Yes	51.6	51.7	51.2	62.6
behav2	Yes	71.9	77.4	72	93.2
behav3	Yes	38.0	63.5	76.5	88.9
behav4	Yes	22.6	58.8	63.6	88.7
behav5	Yes	29.0	73.1	82.9	84.9
exp1	Daily	25.2	31.9	28	33.8
	Weekly	32.1	35.8	35.6	34.2
	Monthly	24.3	19.2	22.5	23.1
	Less	15.4	12.6	13.6	7.8
	Never	3.0	0.5	0.2	1.1
exp2	Yes	21.4	16.2	14.9	45.2
exp3	Yes	8.0	5.4	6.4	47.8
exp4	Yes	6.1	3.4	2.8	48.1
exp5	A lot	27.2	31.8	27.6	40.6
	A little	49.8	57	60.5	53.9
	Nothing	23.0	11.2	11.9	5.5
know1	Correct	62.6	89.7	86.6	48.6
	Incorrect	9.3	3.8	2.8	38
	Not sure	28.2	6.5	10.6	13.4
know2	Correct	58.9	80.5	80	48.9
	Incorrect	8.5	4.1	3.4	36.1
	Not sure	32.6	15.4	16.6	15
know3	Correct	47.8	68.8	71	44.4
	Incorrect	24.6	18.5	15.6	40.9
	Not sure	27.6	12.8	13.4	14.8
know4	Correct	30.3	56.2	54.2	37.8
	Incorrect	15.1	13	9.9	41
	Not sure	54.6	30.8	35.9	21.2
know5	Correct	67.1	75.5	76.2	31.6
	Incorrect	17.6	20.9	18.9	58
	Not sure	15.3	3.6	4.9	10.4
know6	Correct	28.7	67	73.1	64.7
	Incorrect	21.9	12.4	12.6	26
	Not sure	49.4	20.6	14.2	9.2
know7	Correct	44.6	59.9	58.3	43.2
	Incorrect	12.0	12.9	13.9	38.4
	Not sure	43.4	27.2	27.9	18.4
know8	Correct	24.4	37.1	53.4	26.6
	Incorrect	25.5	38.5	27.5	53
	Not sure	50.1	24.4	19.1	20.4

Table 4: Proportions of responses to each question for the full samples. **Green** responses are within $\pm 5\%$ Pew proportions, **orange** responses are $\geq 10\%$ of Pew proportions.

Q	Ans	Pew	Repr	Bal	MTurk
percep1	VC	4.8	5.5	5.9	26.9
	SC	37.1	41.2	42.8	44.4
	NTC	40.3	37.6	37.5	21
	NCAA	17.7	15.6	13.9	7.8
percep2	VC	5.1	4	3.5	26
	SC	29.6	34.5	32.6	38.4
	NTC	40.6	40.6	40.5	22.5
	NCAA	24.8	20.9	23.4	13.1
percep3	VC	2.9	2.8	2	22.8
	SC	17.8	21	17	39.1
	NTC	46.4	47	49	25.9
	NCAA	32.9	29.2	32	12.2
percep4	VC	3.6	3.4	3.1	26.8
	SC	27.2	28	27.6	41.6
	NTC	47.1	46	48.5	22.4
	NCAA	22.1	22.6	20.8	9.2
percep5	VC	3.6	4.4	2.9	22.8
	SC	27.2	18.9	18.2	40.4
	NTC	47.1	42.2	42.4	21.5
	NCAA	22.1	34.5	36.5	15.4
percep6	VC	10.4	9.9	8.4	28.9
	SC	47.0	49	49.8	44.5
	NTC	28.5	29.5	31.5	18.1
	NCAA	14.1	11.6	10.4	8.5
percep7	VC	5.7	6.2	4.5	23.5
	SC	30.2	31.4	34.1	40.8
	NTC	36.6	35.9	35.1	23.6
	NCAA	27.4	26.5	26.2	12.1
percep8	VC	8.1	6.6	5	29.8
	SC	42.1	38	39.5	42.8
	NTC	31.3	33.2	32.8	16.6
	NCAA	18.5	22.1	22.8	10.9
belief1	Yes	39.1	43.2	38.8	71
belief2	Yes	67.3	76.9	74.4	82.3
belief3	Yes	56.1	43.5	36.4	67.2
belief4	Yes	84.9	85.1	82	83

Table 5: Proportions of responses to each question for the full samples. **Green** responses are within $\pm 5\%$ Pew proportions, **orange** responses are $\geq 10\%$ of Pew proportions.

Aunties, Strangers, and the FBI: Online Privacy Concerns and Experiences of Muslim-American Women

Tanisha Afnan¹ Yixin Zou¹ Maryam Mustafa² Mustafa Naseem¹ Florian Schaub¹
¹University of Michigan School of Information ²Lahore University of Management Sciences

Abstract

Women who identify with Islam in the United States come from many different race, class, and cultural communities. They are also more likely to be first or second-generation immigrants. This combination of different marginal identities (religious affiliation, gender, immigration status, and race) exposes Muslim-American women to unique online privacy risks and consequences. We conducted 21 semi-structured interviews to understand how Muslim-American women perceive digital privacy risks related to three contexts: government surveillance, Islamophobia, and social surveillance. We find that privacy concerns held by Muslim-American women unfolded with respect to three dimensions of identity: as a result of their identity as Muslim-Americans broadly (e.g., Islamophobic online harassment), as Muslim-American women more specifically (e.g., reputational harms within one's cultural community for posting taboo content), and as a product of their own individual practices of Islam (e.g., constructing female-only spaces to share photos of oneself without a hijab). We discuss how these intersectional privacy concerns add to and expand on existing pro-privacy design principles, and lessons learned from our participants' privacy-protective strategies for improving the digital experiences of this community.

1 Introduction

Islam is the fastest growing religion in the United States [32]. Despite Islam's growing role and presence in U.S. history, Muslim communities in the U.S. have to contend with discrimination, prejudice, and mass surveillance [20, 47, 50, 68].

Muslim-American women are further subjected to a unique set of targeted attacks and stereotypes while also facing heightened vulnerability related to gender-specific veiling practices (such as the hijab), which act as visible identifiers of Islam [30, 99]. Western narratives paint Muslim women as meek, oppressed, and complicit in their own apparent subjugation [48, 64, 76]. These attributes can result in serious consequences in various contexts, such as hiring discrimination [4, 16]. Additionally, within their own religious and cultural communities, Muslim women might face restrictive gender norms and behavioral expectations, leaving them vulnerable to social consequences if transgressed. These stereotypes, coupled with implications related to other marginalized identities such as immigration status, race, and gender, mean that Muslim-American women may need more specific ways to control their information and own their narratives.

While privacy has been studied extensively [15, 27, 55, 71], the particular concerns and circumstances of Muslim women are relatively understudied. Prior work at the intersection of Muslim experiences and human-computer interaction (i.e., Islamic HCI), while offering rich insights into some of this community's experiences [1, 2, 80, 81, 100], often centers on Muslim women residing in Muslim-majority countries. Our research expands on existing Islamic HCI literature by exploring the additional challenges and perspectives of Muslim women living in countries where Muslims are a minority group, specifically in the United States. Prior research also reveals how an individual's level of religious adherence may influence their preferences and behaviors [58, 66, 67, 107]. We are interested in understanding to what extent individual religiosity (particularly how tenets of Islam, which often prescribe heightened values of modesty to women [2, 34]) may shape how Muslim-American women navigate their online privacy concerns.

To understand if and how Muslim-American women experience privacy concerns, we interviewed 21 Muslim-American women about their typical tech consumption, privacy-protective behaviors and strategies, and scenario-specific privacy concerns. Our findings show that privacy concerns held by Muslim-American women manifest in three distinct

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

dimensions. First, participants expressed privacy concerns as a result of identifying as Muslim-American broadly. Participants described deliberately choosing when and where to disclose this identity and how such disclosure could pose risks to them (e.g., feeling the need to constantly moderate their speech even in personal text messages, because a government agent may be monitoring them). Second, participants identified concerns about potential harms as a result of identifying more specifically as Muslim-American *women* (e.g., being held to higher scrutiny by their cultural community for sharing photos of themselves hanging out with individuals of the opposite gender). At the third and most personal level, participants' individual religiosity and relationship to Islam also shaped their privacy concerns and behaviors. Participants who described themselves as more deeply religious were more likely to have more private online presences (e.g., sharing fewer photos of themselves), but all of our participants' privacy preferences were shaped by their lived experiences as Muslim-Americans broadly and Muslim-American women specifically.

Our participants also shared key strategies they have adopted to mitigate their concerns (e.g., creating female-only spaces on social media to share more intimate content) and noted how existing technology does not meet their privacy needs. We discuss implications of our findings, including an intersectional lens in conceptualizing privacy and design recommendations for better addressing the privacy needs of Muslim-American women.

Researcher Positionality. Our research team consists of members with both insider and outsider perspectives, which contributed to our analysis approach and understanding of findings. Three authors identify as Muslim. Three authors identify as women, and two of them as Muslim women. The authors have diverse cultural backgrounds and religious attitudes, including Muslim women who wear the hijab and those who do not. The first author, who conducted all interviews, identifies as a cisgender Muslim-American woman.

2 Related Work

We examine existing research on Muslims in America, women and privacy in Islam, and the privacy risks Muslim women face.

2.1 Muslims in America

Muslims have been historically othered in America as a religious minority. Islamophobia, the specific prejudice against and hatred towards Muslims, surged after the 9/11 terrorist attacks [33, 50, 76, 94]. Since then, Muslims have often been portrayed by the media with “continuous reference to images of extremism, terrorism, and irrationality” [94]. Respective portrayals delineate the American ‘us’ and the alien ‘them,’ perpetuating a conflict for Muslim Americans who must reconcile these two seemingly disparate parts of

their identity. Hijab, a veil or headscarf worn publicly by some Muslim women, is a highly visible identifier of Islam. This makes hijab-wearing Muslim women particularly vulnerable targets of hate speech and crimes [76] while exposing them to gendered perceptions such as the stereotype of “oppressed Muslim woman” [87]. Through a Western lens, the image of a veiled female represents the subordination of women, falsely rendering Muslim women as either content in their disenfranchisement or in need of rescue [30].

The hypervisibility of Muslims in the U.S., due to amplified media depictions following 9/11, gave rise to growing “Muslim self-consciousness” [47, 91] and efforts to ‘repackage’ and ‘rebrand’ the Muslim identity to be more appetizing to Western values and norms. Muslim Americans may purposefully choose which aspects of themselves are publicly visible to distance themselves from the ‘Muslim’ label, e.g., by framing abstention from alcohol in social settings as a health-related concern rather than a religious conflict [91]. A more overt approach, often employed by community leaders, is to construct a ‘modern and moderate’ Muslim-American identity to be more compatible with American norms [78]. This approach ranges from smaller, self-policing behaviors (e.g., wearing ‘friendlier’ pink hijabs rather than more stigmatized black hijabs) to larger decisions such as moving to predominantly white neighborhoods [20, 91]. In our study, we explore how the Muslim-American identity conflict manifests in digital spaces, and how the mainstream stigmatization of Islam affects participants’ privacy concerns and experiences online.

2.2 Women and Privacy in Islam

Religiousness, or the degree to which an individual adheres to the tenets of their religion, may also influence one’s privacy needs, concerns, and behaviors. Prior research has studied religiousness in healthcare and consumer behavior [58, 66, 67, 107]. Higher levels of religious involvement have been shown to have positive correlation with psychological well-being [67], but can also be deterrents for seeking treatment for stigmatized diseases such as HIV [74]. Religious individuals are less likely to be impulsive shoppers [58], more likely to orient along traditional gender lines in purchases [107], and more likely to exhibit brand or quality consciousness [66]. In studies measuring religious involvement, women (compared to men) and individuals of racial/ethnic minority groups consistently have higher scores [57].

In our study, we explore how Islamic conceptualizations of privacy might influence privacy concerns and behaviors of Muslim-American women. Western conceptualizations of privacy tend to center individual freedoms [105]. By contrast, privacy in Islam is tied to ideals of modesty and family honor, often extending beyond the personal self [34]. Muslim women carry additional responsibilities to uphold their family’s reputation via their own individual actions and opinions. The concept of preserving family honor is

unevenly laid on Muslim women more than men [77, 80, 100], as reflected by Muslim women’s stricter privacy practices on social media platforms [1, 2]. Three notions of privacy are described in the Qu’ran [2]: the *awrah* represents the most intimate or private spaces that must be shielded from others (e.g., parts of a woman’s body), the *hurma* represents pure and sacred ‘spaces’ that must be protected to preserve their sanctity (e.g., the family home), and the *haq al-khososyah* is one’s right and responsibility to protect both their *awrah* and *hurma* through actions. Although Muslim women in the U.S. have roots in many different ethnic and cultural communities, acknowledging the interplay between gender, modesty and privacy in Islam is important for best understanding the values and attitudes of Muslim-American women. Privacy concerns as a result of gendered Islamophobia [30, 76] may further affect Muslim-American women’s online disclosures and behaviors.

2.3 Muslim Women’s Privacy Risks

For Muslim women, the main types of perceived threats discussed in media and prior work include social consequences within the Muslim community, government surveillance, and Islamophobia. As such, we base our interview protocol on these scenarios.

2.3.1 Social risk factors within community

Muslim women’s behaviors are often linked to their honor, and by extension, the honor of their families. When behaviors outside of cultural norms are discovered, erring individuals are subject to reputational harms within their communities. *Haram* behaviors, or behaviors not considered permissible by Islam, vary by community but typically include alcohol consumption, engaging in romantic relationships outside of marriage, privately communicating with individuals of the opposite gender, and getting tattoos [8]. Social media pose further privacy risks, requiring Muslim women to consider what information to make public and how their online content may be interpreted. In a study with Muslim-Kuwaiti youth, participants described “shame and loss of face” due to information exposure on social media and exhibited conservative usage as a result [34]. In another study with Muslim-Qatari women, participants viewed Facebook as a medium for simple correspondences rather than a space for deeper self-expression, and actively considered social repercussions of sharing content online that could be misunderstood as haram behavior [53].

For Muslim women living in Western societies such as Muslim-American women, online behaviors become further complicated as they must reconcile conflicting cultural values of ‘mainstream’ society with certain conventions of Islam. For example, Abokhodair and Vieweg document a scenario in which a Muslim woman grappled with the decision of accepting a male coworker’s Facebook friend request to be sociable or rejecting it out of obligation to family expectations [2].

Social media has become a challenging terrain to navigate for Muslim women who want to engage in different behaviors that correlate with different facets of their lives. This struggle aligns with prior research on context collapse, i.e., multiple social circles with varying norms become flattened into a singular audience on social media [56, 101]. Strategies for coping with context collapse are often burdensome, and individuals may opt to mute certain disclosures entirely [26, 31]. In our work, we sought to understand the role of context collapse and specific cultural or religious expectations on Muslim-American women’s online behaviors. Though participants assigned varying levels of significance to these factors, they influenced and constrained all of our participants’ digital activities.

2.3.2 Fear of government and military surveillance

Government actors are recognized as one of the largest threats to the Muslim-American community due to their history of targeted surveillance [20, 47, 50, 92]. Following 9/11, Muslim-Americans have been subjected to institutional surveillance on local and national levels. The PATRIOT Act, a counterterrorism act drafted in response to 9/11, ushered in a new era of surveillance programs by law enforcement targeting Muslims. For instance, the New York City Police Department’s Muslim Surveillance Program targeted Muslim-American communities in the city via undercover operations, secret informants, and other deceptive and invasive tactics [50]. The Pentagon’s Total Information Awareness System (TIA) was another predictive counterterrorism system aggregating data on individuals who may pose future terrorist threats, namely immigrants, Muslims, and other communities of interest. TIA data came from various sources, including financial and medical records, educational records, familial associations, and commercial data such as online shopping histories [68]. This expansion of government capabilities infringed on the civil liberties and rights of many Muslim Americans [33] while deepening mistrust between the American public and its Muslim communities.

In response to rising government surveillance, Muslim American communities exhibited drastic chilling effects in their online and offline behaviors [42, 44, 92]. Though many programs have been dismantled since, new surveillance efforts, claiming to no longer targeting Muslim and Arab communities, continue to make government tracking a relevant concern [9]. Emerging technologies allow for new avenues of data collection [104]. The US military, for example, is known to purchase location data of users from various smartphone apps; some of the data has been used to launch and plan drone attacks in Muslim-majority countries [86]. More recently, such trading of user data raised criticism among Muslim Americans when it was revealed that Muslim Pro, a mobile app for Islamic prayer times, was believed to have sold user data to the U.S. Special Operations Command through data broker intermediaries [17]. We explored the scenario of U.S. government and military surveillance in our study and found several tactics employed

by our participants to address related concerns.

2.3.3 Islamophobia online

Blatant Islamophobia, i.e., explicit hate crimes and speech targeted at Muslims, is prevalent online [12, 13]. Movement towards white nationalism following the 2016 U.S. elections has contributed to an increase in xenophobic behaviors towards Arabs and other Muslim-Americans [106]. Muslim women, particularly those wearing hijab, remain visible targets of these attacks online [48, 64, 76], leaving many vulnerable to assaults on their physical and psychological safety.

Latent Islamophobia, i.e., prejudice against Muslims enacted in implicit ways, can also thrive online [46]. Research on how social media data particularly affects job seeking Muslim-Americans suggests that screening practices have a discriminatory impact on their employability [16]. A hiring discrimination experiment in the U.S. found that Muslim job applicants, who were only identifiable as Muslim on their social media profiles, “received 38 percent fewer e-mails and 54 percent fewer phone calls” than replicated candidates with other religious affiliations [102]. Another study similarly revealed that applicants who had disclosed their Muslim-American identity on social media received 16% fewer callbacks than the identical Christian candidate in specific regions. This influence of online disclosure on U.S. firms’ hiring practices is an important reality to consider in studying Muslim-American women’s online behaviors.

3 Research Method

Prior work has primarily focused on Muslim women living in Muslim majority contexts [1, 2, 34, 97, 100]. In our study, we focused on the experience of Muslim women in the U.S., who additionally have to contend with being targets of mass surveillance, Islamophobia, and media stigmatization, among other concerns [33, 50, 76, 77, 92]. We explored how these factors affect Muslim-American women’s online privacy concerns and experiences.

3.1 Study Design

As Muslim-American women are a relatively understudied population, we opted for a qualitative approach. The first author conducted 21 semi-structured interviews between May and August 2021. Our study was approved by the University of Michigan’s Institutional Review Board (IRB).

Interested individuals were directed to complete a pre-study survey (see Appendix A), asking for demographic information, which we used to contextualize our sample. After completing the pre-survey, participants were invited to share their availability and given a written consent form to complete prior to their interview session. All interviews were conducted remotely via Zoom in English.

Each interview (see Appendix B for the interview script) began with questions to build rapport and gauge the participant’s daily tech use, followed by general questions about tech-related concerns, privacy and their faith. In the second part, we asked scenario-specific questions about four major categories of privacy risks—ad tracking, social surveillance, U.S. government surveillance, and Islamophobia. Our questions were informed by related work examining experiences of Muslim women (primarily in Muslim-majority countries), Muslim-Americans broadly, and women of color in the U.S. (see Section 2). Our goal was to bring these often separate conversations together. Participants were given the opportunity to discuss their personal concerns in Part 1 before being asked about these scenarios in Part 2; almost all participants mentioned at least one of the scenarios unprompted. At the end, we gave participants opportunity to share concerns not yet captured.

After the interview, participants completed an exit survey (see Appendix C) that consisted of the 5-item Islam-specific version of the Centrality of Religiosity scale [43] to measure their level of religious adherence, complementing what was shared during the interview. We slightly rephrased one question for better fit and added another, taking inspiration from the Pew Research Center’s work [65]. Upon completion of the exit survey, participants received a \$20 virtual gift card. Interviews lasted 67 minutes on average, ranging from 41 to 95 minutes.

3.2 Recruitment and Demographics

We sought adult participants who identified with the religion or culture of Islam, had a permanent home in the U.S., and were regular technology users. We also asked about immigration status but did not screen participants based on it. We advertised our study through social media in relevant online groups (e.g., Muslim Women’s Professional Network), by partnering with Islamic organizations (e.g., the Sister’s Committee at a local mosque), and snowball sampling. Leaders at the community organizations we collaborated with also served as pilot interviewees and provided valuable feedback on our interview protocol. While we did not record the exact channel each participant was recruited from, we did not observe concentration in any particular channel. Only two participants were recruited via snowball sampling. The first author kept recruiting participants and conducting interviews until reaching saturation [24].

Table 1 provides an overview of participant demographics. Our study captured the experiences of a specific subset (young and highly-educated professionals) of Muslim-American women. While this focus limits the generalizability of our findings, our study contributes new insights into the unique privacy experiences of this population. Participants were 22 to 39 years old (mean 28 years). All were college graduates, and 11 held graduate degrees. Participants exhibited similar levels of daily screen-time and tech use. Annual household income varied from less than \$25,000 to over \$150,000. Thirteen participants identified as South/Southeast Asian (Pakistani,

Table 1: Participant demographics

ID	Age	CRS	Education	Ethnicity
P01	39	3.8	Master's Degree	South Asian
P02	27	4.8	Master's Degree	MENA
P03	26	3.6	Master's Degree	South Asian
P04	22	2.8	Bachelor's Degree	South Asian
P05	34	4.8	Professional Degree	South Asian
P06	29	4.4	Master's Degree	South Asian
P07	25	4.4	Bachelor's Degree	MENA
P08	25	4.2	Master's Degree	South Asian
P09	25	4.8	Master's Degree	MENA
P10	35	3.8	Master's Degree	MENA
P11	26	4.6	Bachelor's Degree	Black or African
P12	29	4	Master's Degree	Central Asian
P13	29	4	Master's Degree	South Asian
P14	37	4.6	Master's Degree	South Asian
P15	25	4.8	Master's Degree	South Asian
P16	24	3.8	Bachelor's Degree	South Asian
P17	30	4	Bachelor's Degree	South Asian
P18	N/A	2.4	Doctorate Degree	South Asian
P19	23	4.8	Professional Degree	South Asian
P20	27	5	Master's Degree	Central Asian
P21	28	4.2	Bachelor's Degree	Black or African

MENA = Middle East and North Africa.

Indian, Bangladeshi, Indonesian), four as Middle Eastern or North African, two as Central Asian (Afghanistan), and two as Black or African. Participants' CRS scores ranged from 2.4 to 5 (scale range is 1 to 5), skewing toward the higher end. Scores were calculated using responses from items 1-5 on the exit survey. The mean score of 4.17 maps to 'highly religious' [43]. We discuss the validity of these scores later in our findings.

3.3 Data Analysis

Interview sessions were audio recorded with Zoom. One participant asked not to be recorded, and the interviewer took notes instead. Recordings were transcribed using a transcription service. The research team reviewed transcripts to ensure consistency with the recordings. Throughout the data collection process, the research team met regularly to discuss the collected data.

We used an inductive approach [84] to analyze our interview data so that findings would not be constrained by our research questions. We used thematic analysis [19] to organize and interpret interview transcripts and notes. The first author began with theoretical memoing and affinity diagramming to familiarize themselves with the data, while noting initial reactions and ideas. The first author then conducted open, inductive coding across the entire dataset to develop a codebook. The research team then reviewed themes and preliminary codes to check for their relevance to the entire

dataset. Themes were refined through further iterative rounds of coding. Final analysis focused on extracting illustrative examples for a cohesive narrative around our original research questions. Though the research team worked together to develop and evaluate codes throughout the analysis process, the first author coded the entire dataset themselves, therefore not requiring the calculation of inter-rater reliability [61].

3.4 Limitations

We chose an interview approach to gain insights into the privacy experiences of a relatively understudied group. This method also imposed certain constraints. Though our sample had diversity along some parameters such as income, we cannot claim that our sample is representative of the highly diverse population of Muslim-American women. Our sample primarily consists of young, highly educated Muslim-American women. The experiences highlighted in our study are only reflective of the lived experiences of those participants. This also differentiates our sample from Muslim woman populations studied in some prior research (i.e., women in the Global south with limited literacy [7, 10, 34, 80, 81]) and provides important insights about this subpopulation. Furthermore, the interviewer's identity as a Muslim-American woman may have made some participants more likely to disclose some details, but could also have introduced social desirability bias for others [54].

4 Findings

Our findings are organized based on three distinct dimensions of privacy concerns and the respective risks and harms experienced by our participants. First, participants shared privacy concerns tied to their identities as Muslims in the U.S., such as those related to targeted government surveillance. Second, participants described concerns associated with their identities more specifically as Muslim-American *women*, such as those related to gendered cultural norms. Lastly, individual religiosity and how participants practiced Islam (e.g., wearing a hijab) also shaped their online privacy concerns.

4.1 Privacy Concerns as Muslim-Americans

While participants held multiple intersecting minority identities, many related perceived privacy risks to their identity as Muslim-Americans. Participants viewed these risks as relevant to any Muslim-American regardless of gender, age, or other characteristics. Concerns centered on the U.S. government and military, strangers online, and companies.

4.1.1 Surveillance by the U.S. government and military

The most prominent concern, mentioned by almost all participants, was targeted surveillance by the U.S. government or military. While counterterrorism efforts targeting Muslims

emerged in the years immediately following 9/11 and many have been disbanded since, several participants described suspicion about the extent to which they were being monitored by the government. Participants recounted stories of invasive government practices they heard about from secondary sources (e.g., media outlets, podcasts) or from their own personal communities (e.g., a local mosque). Some participants described witnessing or experiencing negative actions by governmental entities (e.g., being disproportionately subjected to random TSA checks). P19 discussed how a suspected FBI agent had been monitoring and harassing community members at her mosque:

“Basically the FBI sent a fake convert...to [my] masjid ...This guy would go to people’s houses, befriend them, record their private conversations. He had a camera on one of the buttons of his shirt ...This guy would just bring up jihad [holy war in Islam] randomly and all the guys were like, ‘Okay...’ Eventually the masjid leadership ended up reporting this guy to the FBI and the FBI didn’t do anything about it because they were like, ‘Oh, it’s our guy.’ So the masjid got really suspicious.” (P19)

Governmental counterterrorism efforts have been intentionally hidden [93]. With little verifiable information, many participants speculated that the government simply had access to ‘everything,’ i.e., any data about them in existence. Participants thought that the government’s reach extended from public social media posts to private text messages. This concern of wide-reaching government access based on feelings of uncertainty has also been observed in other communities such as undocumented immigrants in the U.S. [39].

Additionally, participants often conflated what was accessible to private companies with what was accessible to the U.S. government or military. More than half of the participants expressed concerns about how their personal data may be exploited by private companies (e.g., companies profiting from targeted ads based on their personal data) as a generic privacy risk. Several participants further shared concerns about how private companies may share their information with the government. For instance, P10 highlighted the reported data flow from the Muslim Pro app to the U.S. military through data brokers:

“This is scary for me...Because I belong to a certain group like being a Muslim person, I have to be watched. This is kind of a burden...especially [when] anything that you can type or write on social media can be used against you...Maybe I’m overreacting, but since the Muslim Pro app thing, when we all knew that they were selling our data to the biggest bidder, I’ve questioned a lot what I’m doing.” (P10)

As a result of perceived targeted surveillance and concerns about how their data might be misused, many participants described experiencing chilling effects similar to those expressed by the Muslim community immediately after

9/11 [92]. This concern was exacerbated by the little autonomy participants felt they had against the entities in question. Most participants felt they had ‘some’ or ‘little control’ over information collected about them by private companies; 12 participants reported feeling ‘no control’ regarding information collected by the government.

Consequently, participants shared how they applied extra caution in day-to-day online and offline behaviors, such as avoiding posting about certain topics (e.g., political opinions critical of the U.S. government on Twitter). These chilling effects inhibited the degree to which participants felt they were able to freely express themselves online, meaningfully engage with others on social media, and consume media of interest. P14 shared why she adopted selective self-expression online:

“I, as a Muslim, would not say certain words over text or even online just because I know that those are not good words to use...That would trigger [someone] to monitor and look into my profile and what I’m doing, and potentially have people tracking me. There are certain things that we do online that would elicit a greater response from other people. I think those types of things are flagged...It would be taken to a whole other level versus a white person looking that up...” (P14)

Fear of government surveillance has been documented as a common privacy concern across the U.S. adult population [98, 108, 109], and our findings indicate a continuing salient level of anxiety among Muslim-American women.

4.1.2 Islamophobia and strangers online

Online hate speech and harassment was another dominant risk participants linked to their identity as Muslim-Americans. Unlike concerns related to government or corporate entities, participants felt more equipped to protect themselves against threats from strangers online. To avoid hostile or unwanted attention, 19 participants described setting their social media accounts to private so that their content was only viewable by approved friends or followers. On platforms designed for public engagement, such as YouTube or TikTok, many participants opted to be passive spectators rather than active content creators, a behavior also mirrored in other exposure-sensitive populations [39, 59].

To avoid inciting hate speech from their approved friends and followers, participants curated audiences with whom they shared Islamic content (e.g., only sharing photos of them celebrating Eid with a subset of friends). Participants noted how their strategies evolved over time. P02 provides an example:

“When I was a high schooler, I’d read maybe a Fox News post on Facebook, and then I would see people cussing out Muslims and I was so naive. I just thought I could convince them, so [I’d be] like, ‘No, Muslims are good’ ...So in those parts of [social media], I experienced very Islamophobic

rhetoric. First it was like the replies back, and then I learned to just block [them], and then I learned after that to just not interact. Because there's no point essentially." (P02)

While such strategies offered participants relief from becoming targets of Islamophobia, many still regularly encountered Islamophobic sentiments shared online. Though not directed at them individually, this constant exposure to harassment still caused distress in their everyday Internet use.

4.2 Concerns as Muslim-American Women

In addition to the concerns linked to being Muslims in the U.S., participants shared concerns and risks specifically tied to being Muslim-American *women*. Many of these risks were described to be equally motivated culturally and religiously, with some participants describing them as results of "outdated patriarchal values" (P18). Participants spoke at length about deep gendered divides in expectations between men and women within their communities. Almost all participants noted that expectations and consequences Muslim men were subject to were significantly different from those for Muslim women. P07 unpacked these uneven cultural gender norms:

"I think Muslim women probably have to be a lot more careful. Because we're definitely judged more harshly. I think men can get away with a lot more, and not get judged for it. The actions they take, [they] don't see him as like, 'oh, this is going to ruin your life' in the way that conversations happen with females in our community. That's how it feels. Like you've ruined your life with this thing. So I think the ways that our communities interact with us is very different." (P07)

While participants expressed being adept in dealing with Islamophobic strangers, they reacted differently when asked about navigating online spaces they shared more closely with their cultural and religious communities. Social surveillance [34, 53] was a phenomenon that almost all participants immediately recognized and felt subjected to. Feeling pressured to accept the friend requests of those in their extended communities out of social obligation, while dealing with the consequences of context collapse [26, 31, 56], greatly limited how participants shared content even on their private social media profiles.

4.2.1 Social taboos and inappropriate content

Definitions of appropriate content to share online varied depending on participants' specific circumstances. For example, a participant who grew up in an area with a large Muslim population and attended an Islamic high school, shared concerns about critiquing a popular Islamic scholar on her personal social media. By contrast, a different participant, who grew up as the only Muslim-American in town, worried

about untagging herself from photos in which she was holding a wine glass. The broad recurring categories of taboo content included photos with members of the opposite gender, photos that placed the participant in potentially inappropriate venues such as bars, photos of wearing clothes that could be considered immodest (e.g., ranging from wearing the hijab too loosely to wearing a bikini on the beach), content about romantic or intimate relationships, and sharing personal opinions on topics that participants felt Muslim women were not typically vocal about (e.g., mental illnesses).

While tensions between cultural and religious expectations of Muslim women and their online behaviors have been reported in prior work [1, 2, 80], our participants faced the added burden of navigating these cultural and religious expectations in a society with differing ideals. Trying to assimilate into western norms to subvert negative stereotypes [38, 94, 99] while upholding the cultural values of Islam left many participants distressed. For example, P01 described following behaviors similar to other American women while being cautious about her representation around family members:

"A lot of times you lead the double life. Not in a bad way, but I don't feel like I'm very different from most other American women because I pretty much do the same thing a lot of American women do. I dress the same as them, I eat the same kinds of foods. I'm single, so I date as well. But I have to hide certain parts of that when I'm around my family because it's inappropriate, and I always have to be aware of what's acceptable culturally, so I can never really share who I am." (P01)

Participants tied these amplified tensions to their intersecting identities as both Muslim and American women. Multiple participants shared feeling they led 'double lives' and being unable to find spaces in which they could share their full existences.

4.2.2 Protective strategies on social media

Participants noted that failing gendered expectations could have several negative consequences. Most concerning was the fear of reputational harm, which would affect participants personally as well as those around them. As P09 explained, "[It's an] obsession with their image. You're a Muslim woman. You can't do this. You're representing our whole community." Participants emphasized that the degree of potential harm depended on each individual family. Six participants had experienced *actual* social repercussions from sharing 'taboo content' on social media, while nine noted that they had not but were still deeply wary of the potential consequences. Participants mainly shared the fear of ostracism; other less commonly noted harms included explicit harassment and physical threats. Though risk does not always lead to tangible harms (e.g., in the form of financial loss), participants' perception of risks should not be dismissed as prior work has noted that perceived risk itself can

simultaneously create harm by affecting one's autonomy and psychological state (e.g., through chilling effects) [23].

To avoid these harms while still engaging in sincere self-expression, participants shared various strategies to create boundaries online. A common strategy was to use multiple social media profiles. All of our participants were active social media users and had accounts on at least three platforms. Having more than one platform meant that participants could add particularly judgemental community members on a selective set of social media accounts while hiding their profiles on others. Those most likely to pose threats typically included older extended family members and religious elders, who usually only used Facebook. As a result, some participants aligned their Facebook appearances more closely with the expectations of their communities while creating more authentic representations of themselves on other platforms like Twitter or Instagram. P12 provides an example:

"Facebook definitely gets the more conservative, modest, professional aspects of me, because not only is that my friends, but it's also family. I have some family that are really strict...Not much goes to Facebook, and if things do go into Facebook, they're still very modest, very conservative, very clean post in aspects of what I wanted to post." (P12)

Some participants also took steps to limit the content others could see on the accounts they shared with those in their community. Examples included using options for restricted audiences (e.g., the close friends feature on Instagram), configuring privacy settings (e.g., locking their profiles on Facebook), and carefully vetting what kinds of content they posted. P06 described having a 'no-list' of friends on Facebook who had limited visibility of the content she posted:

"I definitely had a list of people [on Facebook], I think it was just called my 'no-list' and it was just like family members that I felt like were a little...not trustworthy. I just felt like they would more like[ly] share things with older family members or other family members, and I just didn't really want to risk it... So if I posted a picture with me and all my friends at the beach, it was for everyone but my list of no people." (P06)

Despite best efforts, some participants shared experiences of data leakage, in which personal content they posted ended up reaching unintended viewers. P04 recounted how a photo in which she was tagged leaked to her family members and expressed her frustration with Facebook's privacy settings:

"There's a time that I was wearing shorts in August in Austin, Texas...It was just me standing there with my friends. They took a photo. I was like, 'Oh, that's fine. They took the photo. What are they going to do, send it to my family?' But then they posted it on Facebook. I think it auto-tagged me...Somehow my settings were configured so that my friends can see the photos that I'm tagged in from

other people. So, my family members had seen it because it was posted by someone else before I could notice and untag myself or delete it...I didn't know it was there until I logged in and I saw it was there. I would've preferred ... 'Hey, you're tagged in this photo. Do you want it to be on your timeline?' And it's up to me to say yes or no." (P04)

The desired feature P04 describes exists in Facebook but is not the default. Participants attributed many instances of unintended content sharing to the confusing choice architecture and privacy-unfriendly default settings on social media platforms, echoing existing privacy research on dark patterns around privacy controls [22, 41, 55]. Other participants attributed data leakage to individuals in their closer circles who might have exposed their content to others. Interface changes without sufficient notifications further pose barriers for participants to manage their content effectively, as P05 described:

"I think Facebook changes how you have to adjust your privacy settings, like every six months. And you are like, 'what is this new thing I have to do? I have to click how many buttons and do X, Y or Z?'" (P05)

Ultimately, most participants felt they had more control over the personal information they shared with others on their private social media compared to limiting what data was available to companies and the government. However, control did not necessarily match concern levels. Though participants may have felt less control over the information collected about them by private companies, most participants expressed heightened anxiety over social consequences than surveillance capitalism by private companies [109]. This finding stands in contrast to the reported privacy concerns of 'general' American Internet users, who typically identify private companies as the biggest threat to their information [11].

4.3 Religiosity's Influence on Privacy Concerns

In addition to concerns tied to being Muslim-Americans and Muslim-American women, our findings suggest that differences in personal beliefs (e.g., what constitutes prayer), religious practices (e.g., veiling practices such as wearing a hijab), and involvement with Muslim-American communities and causes (e.g., the frequency of visiting a local mosque) all played a role in participants' conceptualization of privacy concerns and harms.

To better understand how religion influenced participants' privacy concerns and behaviors, we asked about each participant's religious practices during the interview; we also asked participants to complete an Islam-specific version of the CRS-5 [43]. The majority of our participants scored a 4 or higher on CRS-5 (mean 4.17), suggesting that our sample is 'highly religious.' However, the interview data revealed much greater variation and nuance in religiosity than what the CRS-5 results indicate. Individual participants' relationships with

religion were deeply personal and were not accurately captured by CRS-5. To understand this disconnect between qualitative and quantitative responses, consider participants P07 and P14. Both had similar CRS-5 scores (4.4 and 4.6) but described their religious practices quite differently. P07, while regarding herself as deeply spiritual, shared her deliberation of engaging in only a subset of practices that she felt comfortable with:

“I think I’m a pretty deeply spiritual person and I’ve had a lot of back and forth in terms of how I like to practice with congregations...I’ve stepped away a lot from more organized practice...When I was in a bigger city where there was a lot more community, it just didn’t always feel like the most comfortable. And when I was in very Muslim spaces, it didn’t always feel like a great fit either, so I think I’ve moved away from things that are more established.” (P07)

P14, on the other hand, shared her adherence to more traditional practices, and how visiting and engaging with her local mosque has always been important to her:

“I do the simple [things] like greetings, [celebrating] the holidays, things like that...But I also grew up going to the mosque too, very regularly...And then, I moved around and I continued to always constantly go to the mosque, and even here now, where I live now, I do as well. That was a big part of my religion too, going to the mosque. That cultural aspect, that socialization, is a heavy part for me...Being part of a community, knowing that I’m part of a community too.” (P14)

Based on this insight, we decided to focus our analysis on how participants described practicing Islam in the interviews and how they integrated religious practices into their daily lives. We found that more frequent intentional religious practices coincided with participants who defined privacy, in all regards, as an extremely important personal value. For example, participants who reported praying all five requisite prayers daily showed equal amounts of concern with regards to government surveillance, social surveillance, and surveillance capitalism. In contrast, participants who identified as Muslim more culturally (e.g., only praying on religious holidays) were more likely to show heightened concern for social surveillance, but exhibited signs of resignation or apathy [29] toward data collection practices of corporate entities, viewing them as a trade-off between privacy and convenience [11, 85]. For instance, P06 shared that she preferred having sufficient control over information shared on social media, but was willing to be tracked in other contexts such as shopping:

“On social media, I like being able to exercise a certain modicum of control, just because different people can see different things...like different family members. I don’t necessarily want everything out there all the time. I’d like opportunities to regulate that. And then in terms of other kinds of data, it would depend based on what

it is. There’s some data that I think is important for me to give...that makes things a whole lot easier, like for shopping...Tracking sometimes make[s] things easier and is more targeted. I just would like to exercise a little bit more control in that way, but to a certain degree. I think I’d be okay with giving up some autonomy too.” (P06)

4.3.1 The impact of hijab

Veiling practices, as in whether or not a participant chooses to wear a head or face covering, substantially impacted participants’ privacy concerns. All participants who wore hijab emphasized their autonomy and agency in wearing the hijab as a personal decision. Some wore the hijab as an act of visibility to present themselves as Muslim in all spaces, while others felt it aligned with their conceptions of Islamic privacy and their duty to protect their awrah [2]. Twelve participants mentioned potential consequences of wearing the hijab as a particular religious practice. Some of them felt that they were subjected to more scrutiny by other Muslims. As an example, P20 shared her frustration of having to contend with shaming around *how* one wears the hijab:

“I think for a lot of Muslim women, there is a lot of constant conversation about hijab, what is hijab, how to wear hijab, how should you not wear it...blah, blah, blah. It’s just ongoing. Often times I feel [it’s a] very unhealthy conversation that really doesn’t benefit anyone. And those conversations are driven by people who are not women...I think that’s something a lot of Muslim women can relate to, having to deal with that from outside the community and within the community, being constantly critiqued.” (P20)

Other participants noted that wearing the hijab might disadvantage them in interactions with non-Muslims. For example, hiring managers looking at an applicant’s social media profiles would be able to conclude immediately that they were Muslim based on the hijab, and act in discriminatory ways [5, 48, 73].

While all participants who wore the hijab were proud of their choice and excited to represent themselves in digital spaces, they described how this decision also comes with costs. Our hijab-wearing participants shared unique strategies they adopted to navigate the nuances of appearing visibly Muslim online. Similar to some practices discussed earlier to keep judgemental community members at bay, participants leveraged multiple social media platforms. By dedicating different accounts for different purposes, participants were able to uphold certain outward images while still cultivating safe zones for more authentic expression. Snapchat was particularly popular for its ephemerality of posts, with a few participants sharing how they created women-only spaces with their closest friends on Snapchat to share photos of themselves without hijab.

In addition to managing multiple accounts with different content, our hijab-wearing participants shared other strategies to preserve their privacy when needed. Examples included

using images of inanimate objects or scenery as profile pictures, utilizing internal networks to crowdsource information for their needs (e.g., relying on Muslim Women’s Professional Network instead of LinkedIn to look for jobs), and avoiding certain platforms that could be hostile spaces for Muslim women like themselves. P09 described her practice of selective disclosure and self-representation (showing the hijab or not) based on connections on the platform:

“I already feel like I have a lot working against me being brown, being a hijabi...so I’m twice as cautious about what information I post or how I express my views, which is unfortunate because I am very outspoken and opinionated and still feel that fear. I have a Finsta with the girls and the gays that will see my hair. But I do not trust men. And so especially [on] Snapchat, where I have basically no men, I am more candid with what I will post there.” (P09)

Although our participants varied in their veiling practices and respective motivations, participants shared consistently that wearing a hijab exposed them specific risks and vulnerabilities that were not experienced by Muslim women who chose not to physically veil and non-Muslim women.

4.3.2 Closeness to community and activism

Participants who engaged in public Muslim activism or relevant leadership expressed a particular subset of privacy considerations. These participants publicly advocated for specific social causes affecting Muslim communities online (e.g., on a public Twitter account) or offline (e.g. attending a protest), or have taken on public leadership roles in Islam-affiliated organizations (e.g., being the president of a Muslim students’ association).

Supporting certain social causes, particularly those highlighting the plight of different Muslim communities, often placed participants on the side of issues that could be perceived as ‘un-American’ (e.g., critiquing the U.S. military in the war on terror). As a result, several participants shared how they had personally experienced privacy harms due to their activist work, ranging from targeted online harassment to more intense threats like doxing [96]. Such experience was particularly common when it came to controversial issues such as advocating for Palestinian liberation in discussions of the Israel-Palestine conflict. For instance, P02 shared her concern of being listed on Canary Mission, which keeps a blocklist of pro-Palestine activists, and how that might impact her job prospect:

“Canary Mission is a website that [documents] anyone working in anything related to boycotting or divesting Israel, or is Pro-Palestine...They basically dox people on that website and employers look through that website, so then those people can’t get jobs. That’s something I am very careful [about] around my privacy or my identity anywhere. I do have separate accounts for different

things...but if my face and name is on there, it opens you up to a lot of harassment.” (P02)

Participants felt helpless with regards to these concerns and struggled to develop meaningful strategies to mitigate privacy risks associated with public Muslim activism other than opting for more low-effort and anonymous ‘slacktivism’ [82]. However, as P19 unpacked, hiding traces of engagement with Muslim activism is hard, and any slip-up could lead to severe reputational damage:

“If you go to a protest, your name will be on there. You [might] just share a picture of you at a protest, right? Cool. You’re supporting a really worthwhile cause. Meanwhile someone...could be like, ‘Oh my God.’ And then post you on their website and your job prospects gone, your social image tainted, people are calling you anti-Semitic, [or] they’re calling you all these hurtful things that aren’t true.” (P19)

Ultimately, this left participants feeling as though they were at an impasse. Participants had to either curb their activist work or risk facing serious repercussions if they continued, a dilemma also echoed in the continued chilling effects of fears of government surveillance.

5 Discussion

Privacy needs are shaped by environmental, contextual, and individual factors [3, 55, 71, 75]. However, the privacy choices available in mainstream technology are often oriented along profit margins and the larger goals of private-interest companies. Privacy dark patterns are common among online service providers [18, 70], deceiving users into surrendering their personal information to maximize profit [109]. Value-sensitive design suggests that technological artifacts are not value-neutral and instead reflects the creators and communities they are borne from [36]. Even in cases where users’ privacy needs are prioritized, technology developed and designed for a ‘typical’ user in the U.S. will deviate from the preferences of marginalized individuals and users across the globe [27, 108]. Prior Islamic HCI work, primarily situated in Muslim-majority countries, has recognized the role of Islam in users’ interactions with digital technologies. Most notably, Islamic sociocultural norms, widely adhered to by Muslim families and individuals, can significantly impact how privacy is understood and put into practice (e.g., women consider their *awra* when posting photos of themselves) [2, 45, 69]. Our study shows how boundaries between Islamic norms and Western-influenced technology get blurred in the experiences of Muslim-American women — members of both mainstream American society and of their particular religious and cultural communities. Next, we discuss the crossroads of intersectionality and privacy, and outline design opportunities to support the needs of Muslim-American women.

5.1 Privacy Through an Intersectional Lens

Our findings align with similar concerns expressed by women in previous Islamic HCI research (e.g., upholding expectations of modest dressing by community elders [2, 45, 77]), but at heightened degrees because of our participants' intersecting identities as Muslim and American women. Our participants further contended with unique considerations due to their identity as Muslim women in the U.S. (e.g., being part of a stigmatized minority religion, being members of minority ethnic communities), and these tensions manifested in different ways involving a variety of actors. For some, the fear of government surveillance inhibited how they shared their political opinions on specific topics online. Some struggled with crafting an online presence that upheld the 'rules' enforced by their elders while reflecting their more 'American' sensibilities (e.g., debating whether to post a photo at the beach in swimwear). Others worried about Islamophobic threats, some pertaining to their physical safety, when interacting with strangers online. These situational anxieties as a result of being Muslim-American, coupled with concerns of other Muslim women documented in prior work (e.g., debating whether to share photos of oneself without hijab [53]), left our participants feeling vulnerable.

In addition to the unique context of being a Muslim woman in the U.S., we must also recognize the diversity within the Muslim-American women population compared to populations of women in Muslim-majority countries [6]. Women in our sample, and across the Muslim-American women population, hail from various ethnic, racial, and socioeconomic backgrounds. These different visible and invisible social identities interact and intersect in many ways, exposing individuals to varying experiences of discrimination, privilege, and acceptance. This broad range of social identities results in very different lived experiences, even among our small sample, which further differs from the more unified set of challenges experienced by those living in more homogeneous Muslim-majority countries.

Examining the experiences of people who live with multiple marginalized identities, like Muslim-American women, enables a deeper understanding of how privacy concerns are rooted in the intersection of identities; such insights may not as readily appear when focusing on a single or few minority characteristics. Crenshaw developed the concept of intersectionality [25], drawing on the work of many before her, as a framework for better understanding the intersections of race and gender. Work since then has discussed the application of intersectionality in HCI research [52, 72, 79, 90]. Women of color in the U.S. are subjected to the ramifications of male superiority and white supremacy among other hegemonic structures. Muslim-American women, more specifically, are regularly exposed to sexism, racism, and religious discrimination [20, 64]. The intersections of oppression mean that Muslim-American women often face prejudice for each of their individual identity characteristics, but also in compounded ways that cannot be un-

tangled. This insight was revealed in conversations with many of our participants, including one who was unsure if the hostile looks she received from strangers was due to her hijab or her visible Blackness, making her further protective of both identities.

Our findings add nuance to existing understanding of Islamic norms in the digital world. While Muslim women in Muslim-majority countries face similar religious and cultural expectations within their communities, our participants, as Muslim women in the U.S., described the extra burdens of having to dispel stereotypes to those outside their community, including the 'violent extremist,' the 'oppressed Muslim woman,' and other stereotypes associated with their race, gender, and class identities. The minoritized experience of Muslim-American women helps conceptualize the privacy needs of marginalized Internet users and how they relate to and differ from those of more dominant groups [60].

5.2 Designing for Muslim-American Women

Our findings on the privacy concerns and experiences of Muslim-American women reveal perspectives of individuals living with multiple marginalized identities in relation to privacy, usability, and design. While design improvements alone cannot address deep-rooted structural and cultural issues, we provide some key design insights and opportunities. Our recommendations are closely based on insights provided by our participants and further support prior frameworks for designing usable and useful privacy interfaces [35, 88, 89], social justice-oriented design [28], trauma-informed computing [21], feminist HCI [14], and more. This alignment with prior work indicates the broader benefits of considering—and centering—marginalized users in the design process: as more diverse perspectives are included to better represent the wide spectrum of individuals' privacy needs, users from all backgrounds also stand to benefit from more robust applications of inclusive privacy design.

Considering identity-specific needs. Privacy settings are often difficult to find and use [22, 40]. Our participants echoed this sentiment, and several found privacy settings hard to configure for their goals. While usability issues of privacy settings affect all users, our participants expressed greater insecurity and anxiety due to their identity-specific concerns about consequences of Islamophobia, social surveillance, and more. Participants were particularly frustrated when different platforms had drastically different privacy settings, which posed challenges to their impression and identity management.

Existing guidelines for designing privacy controls often focus on general usability, modality, and legal requirements [35, 88, 103]. Following these principles, making privacy controls easier to find and requiring consistency across platforms might help resolve some of our participants' tensions and provide a stronger sense of safety. As an important next step, usable privacy design needs to shift from solely focusing

on the affordances of privacy controls to also considering how identity and contextual aspects, such as digital literacy skills [35], may affect users' needs. For example, though some design ideologies advocate for less notifications to alleviate burdens on users' cognitive load [83], some of our participants felt extremely anxious about unanticipated system updates and changes to privacy settings due to social surveillance concerns. These participants would benefit from timely and trauma-informed notifications about such changes [21]. Though our participants held many of the same general privacy concerns as other Internet users, the unique contextual factors that affect Muslim-American women must be treated with care and should be reflected in system and interface design.

Enabling identity-based audience controls. Many participants engaged in privacy-protective strategies that were directly tied to their identities as Muslim-American women. For example, some participants were part of closed groups on Facebook or had created private alternatives spaces (e.g., secret accounts under pseudonyms) to share specific content with subgroups of peers. These behaviors allowed our participants to draw clear boundaries and differentiate audiences to cope with context collapse [56], similar to the practices of other marginalized populations such as LGBTQ+ communities [31], sex workers [59], and undocumented immigrants [39].

We suggest that platforms should explore more direct opportunities for users' audience stratification to help users find better channels for peer support and grants users more autonomy. For example, many hijab-wearing participants mentioned a need for women-only digital spaces. Instead of having to go through the tedious process of adding individual users to custom audiences, platforms could offer automatic differentiation options such as 'XYZ trait only' in dropdown lists based on other users' disclosed traits, similar to existing choices like 'friends of friends only' [62]. This type of functionality, however, also presents its own set of challenges. Allowing users to filter others by identity traits could reinforce echo chambers [49] and online segregation [37]. Spaces catering to those who share similar experiences and identities could be abused by predatory individuals for targeted harassment. The feature's design, if not done carefully, could lead to users revealing sensitive characteristics about themselves unintentionally due to the groups they are added to; a potential idea to mitigate this risk is enabling users to only allow particular other users to exercise these filters about them. To avoid misuse and abuse, identity-specific design approaches require further research. Respective guidelines must be crafted carefully in collaboration with community leaders, members, and organizations.

Supporting cross-platform data management. Aside from lists, groups, and audience settings on a particular platform, part of our participants' strategies depended on the ability to curate content and segregate audiences across multiple social media platforms. All participants reported using at least three

different platforms, each for distinct purposes. This strategy comes under fire as private companies move towards merging different services and developing integrated ecosystems. For example, Facebook and Instagram, both owned by Meta, are tightly intertwined: Instagram may suggest 'People you may know' based on connections on Facebook, and vice versa [63]. This context collapse creates harms—not just for our participants but also for other marginalized populations [59]—by violating the boundaries users intentionally set to avoid unwanted exposure. Companies should assuage the concerns of these populations by being transparent about how these suggestions are made, and create features that allow them to control if they are suggested to other users, and if yes, to whom.

Providing stronger privacy defaults. Our participants faced repercussions as a result of unexpected default settings on certain platforms. For instance, one participant dealt with reputational damage when family members saw a photo that was unintentionally shared as a result of Facebook's auto-tagging feature. Following this incident, the participant was forced to become more familiar with Facebook's privacy settings and configure them to suit her needs. Prior work suggests that more granular privacy choices can sometimes deter users [51, 95], suggesting the efficiency of improving default options. The instances described by our participants could be avoided by requiring companies to practice privacy by default and set initial privacy settings to be most restrictive (e.g., photo tags requiring user approval). The platform could then ask the user if they want to enable certain features such as auto-tagging, and in doing so, explain both the benefits and potential risks of the feature [89]. This suggestion can come into conflict with the business goals of private-interest companies, and therefore may be better enforced through stronger legislation and regulation.

6 Conclusion

Our findings corroborate with prior Islamic HCI research and show how cultural and religious expectations can be unevenly imposed upon Muslim women [77, 100], and how these expectations shape their practices of navigating online and offline spaces. By focusing on Muslim women living in the U.S., our study contributes new insights into this population's concerns and experiences as they live in societies oriented around Western norms and attitudes. Our participants expressed privacy concerns as a result of being Muslim broadly, as Muslim-American women, and on their individual practice of Islam. Participants adopted countermeasures to make technology work for them, such as developing women-only spaces for self-expression and using Muslim-friendly workplaces to find job postings. Our findings contribute to an intersectional understanding of privacy. We further presented design recommendations for technologies to better cater to the privacy needs of Muslim-American women.

7 Acknowledgements

We thank our community partners and participants for their valuable time and insights. We also thank the anonymous reviewers for their constructive feedback. This research has been partially supported by the Defense Advanced Research Projects Agency (DARPA) under grant No. HR00112010010. The content of the information does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred. Approved for public release; distribution is unlimited.

References

- [1] Norah Abokhodair, Adam Hodges, and Sarah Vieweg. Photo sharing in the Arab Gulf: Expressing the collective and autonomous selves. In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 696–711, 2017.
- [2] Norah Abokhodair and Sarah Vieweg. Privacy & social media in the context of the Arab Gulf. In *ACM Conference on Designing Interactive Systems*, pages 672–683, 2016.
- [3] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.
- [4] Alessandro Acquisti and Christina Fong. An experiment in hiring discrimination via online social networks. *Management Science*, 66(3):1005–1024, 2020.
- [5] Tanisha Afnan, Hawra Rabaan, Kyle ML Jones, and Lynn Dombrowski. Asymmetries in Online Job-Seeking: A Case Study of Muslim-American Women. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):404:1–404:29, 2021.
- [6] Sam Afridi. Muslims in America: Identity, Diversity and the Challenge of Understanding, 2001. <https://files.eric.ed.gov/fulltext/ED465008.pdf>.
- [7] Syed Ishtiaque Ahmed, Md Romael Haque, Jay Chen, and Nicola Dell. Digital privacy challenges with shared mobile phone use in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):17:1–17:20, 2017.
- [8] Nader Al Jallad. The concepts of al-halal and al-haram in the Arab-Muslim culture: a translational and lexicographical study. *Language Design: Journal of Theoretical and Experimental Linguistics*, 10(1):77–86, 2008.
- [9] Arshad Imtiaz Ali. The impossibility of Muslim citizenship. *Diaspora, Indigenes, and Minority Education*, 11(3):110–116, 2017.
- [10] Sajeda Amin. The poverty–purdah trap in rural Bangladesh: implications for women’s roles in the family. *Development and Change*, 28(2):213–233, 1997.
- [11] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. Technical report, Pew Research Center, 2019.
- [12] Imran Awan. Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet*, 6(2):133–150, 2014.
- [13] Imran Awan. *Islamophobia in cyberspace: Hate crimes go viral*. Routledge, 2016.
- [14] Shaowen Bardzell. Feminist HCI: taking stock and outlining an agenda for design. In *ACM Conference on Human Factors in Computing Systems*, pages 1301–1310, 2010.
- [15] Louise Barkhuus. The mismeasurement of privacy: using contextual integrity to reconsider privacy in HCI. In *ACM Conference on Human Factors in Computing Systems*, pages 367–376, 2012.
- [16] Timothy Bartkoski, Ellen Lynch, Chelsea Witt, and Cort Rudolph. A meta-analysis of hiring discrimination against Muslims and Arabs. *Personnel Assessment and Decisions*, 4(2):1:1–1:16, 2018.
- [17] Johana Bhuiyan. Muslims reel over a prayer app that sold user data: ‘a betrayal from within our own community’, 2020. <https://www.latimes.com/business/technology/story/2020-11-23/muslim-pro-data-location-sales-military-contractors>.
- [18] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies*, 2016(4):237–254, 2016.
- [19] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
- [20] Louis A Calkins. *Homeland insecurity: the Arab American and Muslim American experience after 9/11*. Russell Sage Foundation, 2009.
- [21] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. Trauma-informed computing: Towards safer technology experiences for all. In *ACM Conference on Human Factors in Computing Systems*, pages 544:1–544:20, 2022.
- [22] Yi Chen, Mingming Zha, Nan Zhang, Dandan Xu, Qianqian Zhao, Xuan Feng, Kan Yuan, Fnu Suya, Yuan Tian, Kai Chen, et al. Demystifying hidden privacy settings in mobile apps. In *IEEE Symposium on Security and Privacy*, pages 570–586, 2019.
- [23] Danielle Keats Citron and Daniel J Solove. Privacy harms. *SSRN*, 2021. <http://dx.doi.org/10.2139/ssrn.3782222>.
- [24] Juliet Corbin and Anselm Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.
- [25] Kimberle Crenshaw. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299, 1990.
- [26] Vanessa P Dennen and Kerry J Burner. Identity, context collapse, and Facebook use in higher education: Putting presence and privacy at odds. *Distance Education*, 38(2):173–192, 2017.

- [27] Tamara Dinev, Massimo Bellotto, Paul Hart, Vincenzo Russo, and Ilaria Serra. Internet users' privacy concerns and beliefs about government surveillance: An exploratory study of differences between Italy and the United States. *Journal of Global Information Management*, 14(4):57–93, 2006.
- [28] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. Social justice-oriented interaction design: Outlining key design strategies and commitments. In *ACM Conference on Designing Interactive Systems*, pages 656–671, 2016.
- [29] Nora A Draper and Joseph Turow. The corporate cultivation of digital resignation. *New Media & Society*, 21(8):1824–1839, 2019.
- [30] Rachel Anderson Droogsmas. Redefining Hijab: American Muslim women's standpoints on veiling. *Journal of Applied Communication Research*, 35(3):294–319, 2007.
- [31] Stefanie Duguay. "He has a way gayer Facebook than I do": Investigating sexual identity disclosure and context collapse on a social networking site. *New Media & Society*, 18(6):891–907, 2016.
- [32] John L Esposito. *The future of Islam*. Oxford University Press, 2010.
- [33] Jennifer C Evans. Hijacking civil liberties: The USA PATRIOT Act of 2001. *Loyola University of Chicago Law Journal*, 33(4):933–990, 2001.
- [34] Maha Faisal and Asmaa Alsumait. Social network privacy and trust concerns. In *ACM International Conference on Information Integration and Web-based Applications and Services*, pages 416–419, 2011.
- [35] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. A design space for privacy choices: Towards meaningful privacy control in the internet of things. In *ACM Conference on Human Factors in Computing Systems*, pages 64:1–64:16, 2021.
- [36] Batya Friedman, Peter Kahn, and Alan Borning. Value sensitive design: Theory and methods. Technical report, University of Washington, 2002.
- [37] Matthew Gentzkow and Jesse M Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.
- [38] Peter Gottschalk and Gabriel Greenberg. From Muhammad to Obama: Caricatures, cartoons, and stereotypes of Muslims. *Islamophobia: The challenge of pluralism in the 21st century*, pages 191–209, 2011.
- [39] Tamy Guberek, Allison McDonald, Sylvia Simioni, Abraham H Mhaidli, Kentaro Toyama, and Florian Schaub. Keeping a low profile? Technology, risk and privacy among undocumented immigrants. In *ACM Conference on Human Factors in Computing Systems*, pages 114:1–114:15, 2018.
- [40] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. "It's a Scavenger Hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *ACM Conference on Human Factors in Computing Systems*, pages 384:1–384:12, New York, NY, USA, 2020.
- [41] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites. In *Symposium on Usable Privacy and Security*, pages 387–406, 2019.
- [42] William Hobbs and Nazita Lajevardi. Effects of divisive political campaigns on the day-to-day segregation of Arab and Muslim Americans. *American Political Science Review*, 113(1):270–276, 2019.
- [43] Stefan Huber and Odilo W Huber. The centrality of religiosity scale (CRS). *Religions*, 3(3):710–724, 2012.
- [44] Sunny Skye Hughes. US domestic surveillance after 9/11: An analysis of the chilling effect on first amendment rights in cases filed against the Terrorist Surveillance Program. *Canadian Journal of Law and Society*, 27(3):399–425, 2012.
- [45] Samia Ibtasam. For God's sake! Considering Religious Beliefs in HCI Research: A Case of Islamic HCI. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, pages 9:1–9:8, 2021.
- [46] Namira Islam. Soft Islamophobia. *Religions*, 9(9):280:1–280:16, 2018.
- [47] Amaney Jamal and Nadine Naber. *Race and Arab Americans before and after 9/11: From invisible citizens to visible subjects*. Syracuse University Press, 2008.
- [48] Amaney A Jamal. Trump (ing) on Muslim women: The gendered side of Islamophobia. *Journal of Middle East Women's Studies*, 13(3):472–475, 2017.
- [49] Kathleen Hall Jamieson and Joseph N Cappella. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press, 2008.
- [50] Sara Kamali. Informants, Provocateurs, and Entrapment: Examining the Histories of the FBI's PATCON and the NYPD's Muslim Surveillance Program. *Surveillance & Society*, 15(1):68–78, 2017.
- [51] Stefan Korff and Rainer Böhme. Too much choice: End-user privacy decisions in the context of choice proliferation. In *Symposium On Usable Privacy and Security*, pages 69–87, 2014.
- [52] Neha Kumar and Naveena Karusala. Intersectional computing. *Interactions*, 26(2):50–54, 2019.
- [53] Rodda Leage and Ivana Chalmers. Degrees of caution: Arab girls unveil on facebook. *Girl wide web*, 2:27–44, 2010.
- [54] Douglas Macbeth. On "reflexivity" in qualitative research: Two readings, and a third. *Qualitative Inquiry*, 7(1):35–68, 2001.
- [55] Kirsten Martin and Katie Shilton. Why experience matters to privacy: How context-based experience moderates consumer privacy expectations for mobile applications. *Journal of the Association for Information Science and Technology*, 67(8):1871–1882, 2016.
- [56] Alice E Marwick and Danah Boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133, 2011.

- [57] Joanna Maselko and Laura D Kubzansky. Gender differences in religious practices, spiritual experiences and health: Results from the US General Social Survey. *Social Science & Medicine*, 62(11):2848–2860, 2006.
- [58] Michael E McCullough and Brian LB Willoughby. Religion, self-regulation, and self-control: Associations, explanations, and implications. *Psychological Bulletin*, 135(1):69–93, 2009.
- [59] Allison McDonald, Catherine Barwulor, Michelle L Mazurek, Florian Schaub, and Elissa M Redmiles. “it’s stressful having all these phones”: Investigating sex workers’ safety goals, risks, and practices online. In *USENIX Security Symposium*, 2021.
- [60] Nora McDonald and Andrea Forte. The Politics of Privacy Theories: Moving from Norms to Vulnerabilities. In *ACM Conference on Human Factors in Computing Systems*, pages 40:1–40:14, 2020.
- [61] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):72:1–72:23, 2019.
- [62] Meta. How do I change who can add me as a friend on Facebook?, 2022. <https://www.facebook.com/help/217125868312360>.
- [63] Anna Middleton. How does Instagram know my friends and who to suggest?, 2021. <https://www.alphr.com/how-does-instagram-know-friends/>.
- [64] Heidi Safia Mirza. Embodying the veil: Muslim women and gendered islamophobia in ‘new times’. In *Gender, Religion and Education in a Chaotic Postmodern World*, pages 303–316. Springer, 2013.
- [65] Travis Mitchell. How Does Pew Research Center Measure the Religious Composition of the US? Answers to Frequently Asked Questions, 2018. <https://www.pewresearch.org/religion/2018/07/05/how-does-pew-research-center-measure-the-religious-composition-of-the-u-s-answers-to-frequently-asked-questions/>.
- [66] Safiek Mokhlis. The effect of religiosity on shopping orientation: an exploratory study in Malaysia. *Journal of American Academy of Business*, 9(1):64–74, 2006.
- [67] Alexander Moreira-Almeida, Francisco Lotufo Neto, and Harold G Koenig. Religiousness and mental health: a review. *Brazilian Journal of Psychiatry*, 28(3):242–250, 2006.
- [68] Nancy Murray. Profiling in the age of total information awareness. *Race & Class*, 52(2):3–24, 2010.
- [69] Maryam Mustafa, Shaimaa Lazem, Ebtisam Alabdulqader, Kentaro Toyama, Sharifa Sultana, Samia Ibtasam, Richard Anderson, and Syed Ishtiaque Ahmed. IslamicHCI: Designing with and within Muslim Populations. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, pages 20:1–20:8, 2020.
- [70] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. Dark patterns: Past, present, and future: The evolution of tricky user interfaces. *Queue*, 18(2):67–92, 2020.
- [71] Helen Nissenbaum. *Privacy in context*. Stanford University Press, 2009.
- [72] Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. Critical race theory for HCI. In *ACM Conference on Human Factors in Computing Systems*, pages 265:1–265:16, 2020.
- [73] Teresa Valerio Parrot and Stacia Tipton. Using social media “smartly” in the admissions process. *College and University*, 86(1):51–53, 2010.
- [74] Sharon K Parsons, Peter L Cruise, Walisa M Davenport, and Vanessa Jones. Religious beliefs, practices and treatment adherence among individuals with HIV in the southern United States. *AIDS Patient Care & STDs*, 20(2):97–111, 2006.
- [75] Paul A Pavlou. State of the information privacy literature: Where are we now and where should we go? *MIS Quarterly*, pages 977–988, 2011.
- [76] Barbara Perry. Gendered Islamophobia: hate crime against Muslim women. *Social Identities*, 20(1):74–89, 2014.
- [77] Hawra Rabaan, Alyson L Young, and Lynn Dombrowski. Daughters of men: Saudi women’s sociotechnical agency practices in addressing domestic abuse. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):224:1–224:31, 2021.
- [78] Angel Rabasa, Cheryl Benard, Lowell H Schwartz, and Peter Sickle. *Building moderate Muslim networks*. Rand Corporation, 2007.
- [79] Yolanda A Rankin, Jakita O Thomas, and Nicole M Joseph. Intersectionality in HCI: Lost in translation. *Interactions*, 27(5):68–71, 2020.
- [80] Mohammad Rashidujjaman Rifat, Mahiratul Jannat, Mahdi Nasrullah Al-Ameen, SM Taiabul Haque, Muhammad Ashad Kabir, and Syed Ishtiaque Ahmed. Purdah, Amanah, and Gheebat: Understanding Privacy in Bangladeshi “pious” Muslim Communities. In *ACM Conference on Computing and Sustainable Societies*, pages 199–214, 2021.
- [81] Mohammad Rashidujjaman Rifat, Toha Toriq, and Syed Ishtiaque Ahmed. Religion and Sustainability: Lessons of Sustainable Computing from Islamic Religious Communities. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):128:1–128:32, 2020.
- [82] Dana Rotman, Sarah Vieweg, Sarita Yardi, Ed Chi, Jenny Preece, Ben Shneiderman, Peter Pirolli, and Tom Glaisyer. From slacktivism to activism: participatory culture in the age of social media. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, pages 819–822, 2011.
- [83] Manuel Rudolph, Denis Feth, and Svenja Polst. Why users ignore privacy policies—a survey and intention model for explaining user privacy behavior. In *International Conference on Human-Computer Interaction*, pages 587–598. Springer, 2018.
- [84] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. Sage, 2021.
- [85] M Angela Sasse, Matthew Smith, Cormac Herley, Heather Lipford, and Kami Vaniea. Debunking security-usability tradeoff myths. *IEEE Security & Privacy*, 14(5):33–39, 2016.

- [86] Jeremy Scahill and Glenn Greenwald. The NSA’s secret role in the US assassination program, 2014. <https://theintercept.com/2014/02/10/the-nsas-secret-role/>.
- [87] Christina Scharff. Disarticulating feminism: Individualization, neoliberalism and the othering of ‘Muslim women’. *European Journal of Women’s Studies*, 18(2):119–134, 2011.
- [88] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Symposium on Usable Privacy and Security*, pages 1–17, 2015.
- [89] Florian Schaub and Lorrie Faith Cranor. Usable and useful privacy interfaces. In Travis D Breaux, editor, *An Introduction to Privacy for Technology Professionals*, pages 176–238. International Association of Privacy Professionals, 2020.
- [90] Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. Intersectional HCI: Engaging identity through gender, race, and class. In *ACM Conference on Human Factors in Computing Systems*, pages 5412–5427, 2017.
- [91] Tahseen Shams. Visibility as resistance by Muslim Americans in a surveillance and security atmosphere. *Sociological Forum*, 33(1):73–94, 2018.
- [92] Dawinder S Sidhu. The Chilling Effect of Government Surveillance Programs on the Use of the Internet by Muslim-Americans. *University of Maryland Law Journal of Race, Religion, Gender and Class*, 7(2):375–394, 2007.
- [93] Andrew Silke. *Routledge handbook of terrorism and counterterrorism*. Routledge, 2019.
- [94] Derek MD Silva. The othering of Muslims: Discourses of radicalization in the New York Times, 1969–2014. *Sociological Forum*, 32(1):138–161, 2017.
- [95] Daniel Smullen, Yuanyuan Feng, Shikun Aerin Zhang, and Norman Sadeh. The best of both worlds: Mitigating trade-offs between accuracy and user burden in capturing mobile app privacy preferences. *Proceedings on Privacy Enhancing Technologies*, 2020(1):195–215, 2020.
- [96] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *ACM Internet Measurement Conference*, pages 432–444, 2017.
- [97] Sharifa Sultana, François Guimbretière, Phoebe Sengers, and Nicola Dell. Design Within a Patriarchal Society: Opportunities and Challenges in Designing for Rural Women in Bangladesh. In *ACM Conference on Human Factors in Computing Systems*, pages 536:1–536:13, 2018.
- [98] Joseph Turow and Michael Hennessy. Internet privacy and institutional trust: insights from a national survey. *New Media & Society*, 9(2):300–318, 2007.
- [99] Margaretha A Van Es. Muslim women as ‘ambassadors’ of Islam: Breaking stereotypes in everyday life. *Identities*, 26(4):375–392, 2019.
- [100] Sarah Vieweg and Adam Hodges. Surveillance & modesty on social media: How Qataris navigate modernity and maintain tradition. In *ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 527–538, 2016.
- [101] Jessica Vitak. The impact of context collapse and privacy on social network site disclosures. *Journal of Broadcasting & Electronic Media*, 56(4):451–470, 2012.
- [102] Michael Wallace, Bradley RE Wright, and Allen Hyde. Religious affiliation and hiring discrimination in the american south: A field experiment. *Social Currents*, 1(2):189–207, 2014.
- [103] Na Wang, Heng Xu, and Jens Grossklags. Third-party apps on Facebook: privacy and the illusion of control. In *ACM Symposium on Computer Human Interaction for Management of Information Technology*, pages 4:1–4:10, 2011.
- [104] Nina Wang, Allison McDonald, Daniel Bateyko, and Emily Tucker. American dragnet: Data-driven deportation in the 21st century. Technical report, Center on Privacy & Technology at Georgetown Law, 2022.
- [105] Alan F Westin. *Privacy and Freedom*. Scribner, 1967.
- [106] Andrew L Whitehead, Samuel L Perry, and Joseph O Baker. Make America Christian again: Christian nationalism and voting for Donald Trump in the 2016 presidential election. *Sociology of Religion*, 79(2):147–171, 2018.
- [107] Robert E Wilkes, John J Burnett, and Roy D Howell. On the meaning and measurement of religiosity in consumer research. *Journal of the Academy of Marketing Science*, 14(1):47–56, 1986.
- [108] Yue “Jeff” Zhang, Jim Q Chen, and Kuang-Wei Wen. Characteristics of Internet users and their privacy concerns: A comparative study between China and the United States. *Journal of Internet Commerce*, 1(2):1–16, 2002.
- [109] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile books, 2019.

A Pre-Study Survey

1. In which year were you born? Please enter your birth year in 4 digits.
2. I identify my gender as Women Men Non-binary Prefer to self-describe: ____ Prefer not to disclose
3. What is the highest level of education you have completed? Less than high school High school graduate or equivalent Some college Trade, technical or vocational training Associate’s degree Bachelor’s degree Master’s degree Professional degree (JD, MD, etc.) Doctoral degree Other: ____ Prefer not to disclose
4. I identify myself as (please select all that apply): American Indian or Alaska Native Middle Eastern or North African Asian (including South Asian) Hispanic, Latinx, or of Spanish origin Native Hawaiian or Pacific Islander Caucasian Black or African American Other: ____ Prefer not to disclose

5. What is your current employment status? Employed
 A student A homemaker Military Retired
 Out of work and looking for work Out of work but not looking for work Other: ____ Prefer not to disclose
6. If you selected “employed” in the previous question, please describe your primary occupation: ____
7. What is your immigration status in the United States?
 Citizen (born or naturalized) Permanent resident
 Non-immigrant (student visa, K-1 visa, etc.)
 Refugee/asylum seeker Other: ____ Prefer not to disclose
8. What is your present religion, if any? Christian (including Protestant, Catholic, etc.) Jewish Muslim (including “Islam, Islamic, Nation of Islam, etc.”) Hindu Buddhist No religion, not a believer (including atheist, agnostic) Other: ____ Prefer not to disclose
9. What was your total household income before taxes during the past 12 months? Less than \$25,000 \$25,000 to \$49,999 \$50,000 to \$74,999 \$75,000 to \$99,999 \$100,000 to \$124,999 \$125,000 to \$149,999 \$150,000 or more Prefer not to disclose
10. Do you or anyone in your household own any of the following devices? Please check all that apply. Personal computer Smartphone (can access the Internet, etc.) iPad or other tablet devices E-reader (e.g., Kindle, Nook, etc.) Music Playing Device (e.g., iPod) Console-based gaming system (e.g., Xbox, Nintendo, or Playstation) Voice-activated smart speaker (e.g., Alexa/Echo device, Google Home) Smart TV that connects to the internet Digital media player and microconsole (e.g., Apple TV, Amazon Fire TV) Other: ____ None of the above Prefer not to disclose
11. (For each device selected in the previous question) Thinking about a typical day, how much time do you spend per day using your [Device X]? Never 0-1 hour 1-2 hours 2-3 hours 3-4 hours 4-5 hours 5+ hours I don’t know Prefer not to disclose
12. How would you like to be contacted for more information regarding this study? Email (please enter your full email address): ____ Phone (please enter your preferred phone number): ____

B Interview Protocol

Hello, thanks so much for your time and participation today!

I am a PhD student at [anonymized institution], and I’m really interested in understanding the everyday technology practices of Muslim American women, and how technology can be further innovated to best support your needs and leave

you feeling empowered. I am also a part of a larger research project at [anonymized institution] who is conducting similar research with other cross cultural populations.

In this interview, we hope to learn how you use technology in your day to day for gathering information and communicating with others, what some of your most pressing questions and concerns are, and how you might feel better supported. We hope to eventually use this research to develop tools to support you and other members of your community in your daily tech practices.

You can expect our conversation to take between an hour and an hour and a half today.

A couple of things before we start:

- We will compensate you \$20 for your super valuable time.
- I would like to record this interview to help me remember your responses and later analyze your responses. If you are not comfortable with this conversation being audio or video recorded, please let me know right now.
- To the extent possible, we will ensure that your identity remains completely confidential. This means that we will aggregate comments from all interviews so that your comments are not easily traced to an individual. If we quote you in our final report, we will do so without identifying your name or specific role. If there’s anything you really don’t want on the record, even if it’s anonymous, please let me know that, too.
- This interview is entirely voluntary–if you want to stop the interview at any point during this session, please let me know. We can end the interview at any point and you will still be fully compensated for your time.

Do you have any questions for me? Alright, then let’s get started! I’m going to begin the recording and want to confirm that you are consenting to participate in the study.

Part 1: Opening Questions

- On a typical day, what kinds of devices, websites, apps, online services do you usually use? [Probe: Do you own these devices or share them?]
- Are there aspects that concern you when using technology? [Probe: One topic we hear a lot about lately is privacy – to what extent does privacy matter to you if at all?]
- What does ‘privacy’ mean to you? [Probe: Are there different types of privacy? Does your definition of privacy change when you are online vs. offline?]
- Are you motivated to protect [reiterate what participants said when defining privacy]? Why or why not?

- What ‘stuff’ do you think about when it comes to privacy risks? What specific things would you want to protect? [Probe: Information about yourself? Certain kinds of information? Information about others in your community or network?]
- Who or what do you need to protect these things from? Who or what poses a risk to your information?
- Are there groups of people who have to worry about protecting their information more than others? [Probe: Yourself? Other members in your community? Muslim-American women in general? Why does this group/person have to worry about it more than others?]

Faith-Related Questions

- What do you think it means to be a Muslim-American woman today? [Probe: How would you describe yourself? Your identity? What’s part of that?]
- Are there any experiences unique to being Muslim-American woman today? [Probe: Are there any experiences you would identify as collective experiences for all Muslim-American women?]
- People practice their religion in many different ways. How often do you do something related to practicing your religion? What kinds of things? [Probe: How long have you practiced this way? Have you always practiced this way? Are there times you present as Muslim and other times you do not? How about in online spaces? Do you attend or visit any mosques or religious community centers?]

Part 2: Scenario-Specific Privacy Concerns

Perfect! Thank you for those answers, we’re going to be moving on to the next section of our interview now. These next questions are going to be less general and more specific to a couple of different contexts.

Scenario: Ad Tracking

Today it is possible to take personal data about people from many different sources – such as their purchasing and credit histories, their online browsing or search behaviors, or their public records – and combine them together to create detailed profiles of people’s potential interests and characteristics. Companies and other organizations use these profiles to offer targeted advertisements or special deals, or to assess how risky people might be as customers.

- Is this something you’ve already heard about? [Probe: How many companies do you think use profiles like this for their own goals? Would private companies or organizations use this information for any other reasons [than the ones mentioned in blurb]?]

- When you are online, do you ever see advertisements that look like they might be based on a profile of you that uses your personal data?
- What information do you think is used to create these profiles? [Probe: Personal information (e.g., social identities)? Posts on social media? Search terms? Purchases online? Private conversations via text? Can location data from your personal phone’s location services be used for these profiles? Is this a good or bad thing?]
- Is there any information about you that might be used for these profiles that you wouldn’t want to be used? (E.g., health data, religion, sexual orientation)? Why?
- How accurately do these advertisements actually reflect your interests and personal characteristics?
- How might private companies use a data profile of you in ways that you find acceptable? [Probe: Share your info w/ outside groups doing research that might help improve society? Develop new products? Optimize functionality of the service? Tailor product recommendations?]
- How might private companies use a data profile of you in ways that you find unacceptable? [Probe: What are some concerns you might have about the data private companies are collecting about you?]
- How much control do you feel you have with regards to the information private companies collect about you?

Scenario: US Government/Military Threats

- Based on what you know, do you think what you do (including on your cell phone or offline) is being monitored by the US government or military? How much? Why? [Probe: Does your understanding of what information might be collected about you by the US government or military change the things you do or how you act online?]
- What information do you think the US government or military is particularly interested in collecting about individuals? Why? [Probe: Are they interested in collecting information about some individuals/communities more than others? Why?]
- Do you believe the government collects data about all Americans to assess who might be a potential terrorist threat? [Probe: Is this an acceptable or unacceptable practice? Why or why not? Are some individuals more likely to be monitored closely than others? Why or why not?]
- Do you have any concerns about what information is being collected about you by the US government or military? [Probe: Are any of these concerns related to your identity as a Muslim-American?]

- Have you heard of any instances in which information about people from your community (at large) was collected or used by the US government or military in a way that was harmful?
- Do you think it's possible to go about your daily life without having any government or military entity collect data about you?
- How much control do you feel you have with regards to the information the US government or military collects about you?

Scenario: Online Islamophobia

- Do you think information you share online can be used against you by people you don't know? How?
- Do you think information you share online can be used against you in discriminatory ways? How? [Probe: What kind of information can be used to harm you? What kind of people might want to use information about you to harm you? How might they access that information about you? Do you do anything to protect your information from people you don't know?]
- What platforms or spaces do you feel are people most likely to engage with you in harmful ways? [Probe: Why do you feel this way?]
- Have you ever witnessed or seen an instance of Islamophobia online? [Probe: Would you mind describing that experience?]
- Have you ever personally experienced an instance of Islamophobia online? [Probe: If yes, would you mind describing that experience? If not, have any of your family or friends ever experienced an instance of Islamophobia online?]
- Have you experienced a situation in which what you did online affected your life outside of that space? [Probe: Can your online presence or behavior give rise to discrimination in other environments?]
- How much control do you feel you have over the information you share publicly online with everyone?

Scenario: Social Surveillance & Social Media Use

- What social media platforms or social networking sites do you typically use?
- What kind of information do you share [on mentioned platforms]? Can you give me an example?
- Have you ever hesitated to share something online, even if you weren't posting it publicly? Why?

- What kinds of considerations do you have when posting or sharing something on your personal social media? [Probe: Why do you have these considerations? Does the type of content matter (e.g., political opinions, photos of you, sharing personal thoughts and reflections)? Why or why not? Does the particular social media platform matter? Why or why not?]
- Do you share everything you post online with all of your connections on a given platform? [Probe: Do you have specific audiences that you share specific content with? Do you have specific platforms you share specific content on?]
- Recall a time when you posted something on [particular platform] that you only shared with some of your connections. Can you walk me through the thought process you had as you went through with posting it? [Probe: What would happen if the people you didn't want to share that post with happened to see it?]
- Do you think there can be social consequences to posting certain kinds of content online with your online connections? What are they? [Probe: Where do those consequences come from? Are these consequences related to your identity as a Muslim woman? How? Are these consequences different for Muslim women than they are for Muslim men?]
- We talked about what might be problematic to post/share online. In your practice of your faith, how would you define 'haram' behaviors? [Probe: Is this definition different from how others in your community might describe it? In what ways?]
- Are there similar considerations you have with regards to any other online behaviors (e.g., who you follow, who you are friends with, what you 'like')?
- How much control do you feel you have over the information you share privately online with your connections?

Are there any other important concerns or considerations you have when using the Internet that we have not discussed yet today? Are any of these concerns related to your identity as a Muslim-American woman?

Part 3: Privacy Mitigation Behavior

Now we're going to move away from those context specific questions and think more broadly about all the different concerns we've discussed today.

- I was wondering if you have changed the way you use technology in response to any of those concerns? (E.g., changed settings, used a browser extension/software/other protective tool, abstained from

certain tech usage, etc.) [Probe: Can you tell me about a recent time when you avoided using a specific technology or platform, if that happened? Are there any topics you deliberately choose not to discuss or share via tech (messaging apps, social networks, devices etc.)?]

- Have there been any events in your own life that made you change your technology practices? [Probe: Can you tell me about any specific instances? Is this an active change?]
- Have there been any events related to your identity as a Muslim women broadly that made you change your technology practices? [Probe: Can you tell me about any specific instances? Are you still doing it now?]
- Have you ever experienced a privacy violation, breach, or other negative experience (related to privacy) online? [Probe: For instance, someone gained unwanted access to your personal information?]
- In general, what specific steps, actions or strategies have you taken to protect your personal information and privacy online? Could you give me any specific examples? [Probe: Where or from whom did you learn that strategy? Are these strategies easy or difficult for you to use?]
- What sources do you trust when seeking privacy advice?
- When it comes to protecting your information [or privacy] how helpful or hurtful are the [features/options/settings] on the different apps and platforms you use? [Probe: How could they be better for your needs?]
- How much do you feel you understand the laws and regulations that are currently in place to protect your data privacy?

Part 4: Closing Questions

In your opinion, what are some ways Muslim American women like yourself could better protect themselves online? What seems to be missing for you? (E.g. better tools to allow people to control their personal information, stronger laws regulating what companies can and cannot do with people's

personal information, privacy laws and policies that are easier for people to understand and engage with, better/free educational opportunities that teach individuals about online defense tools and strategies)?

Would you be interested in being contacted for future studies? What would be the best way to reach you?

Any questions about our study or any of the topics we discussed today? If you have any questions later you can always contact me at [anonymized email address].

Thank you so much for participating! As we wrap up and I still have you on the line, I'm going to go ahead and send you the virtual gift card and make sure you received it. While I'm doing that, I'm just going to send you a link to this last post-interview survey that's super brief and you can go ahead and leave whenever you're done. [Link]

C Post-Study Survey

1. How often do you think about religious issues? Never Rarely Occasionally Often Very often
2. To what extent do you believe that Allah or something divine exists? Not at all Not very much Neutral Somewhat Very much
3. How often do you take part in religious services? Never Rarely Occasionally Often Very often
4. How often do you experience situations in which you have the feeling that Allah or something divine allows for an intervention in your life? Never Rarely Occasionally Often Very often
5. People practice their religion in different ways. How often, if at all, do you pray? Hardly ever, only during religious holidays Only on Fridays Only on Fridays and religious holidays More than once a week Every day at least once Every day five times
6. How important is religion in your life? Not at all important Not too important Somewhat important Very important

An open door may tempt a saint: Examining situational and individual determinants of privacy-invading behavior

Markus Langer[§], Rudolf Siegel^{*}, Michael Schilling^{*}, Tim Hunsicker[§], Cornelius J. König[§]

[§] Saarland University, Industrial and Organizational Psychology, Saarbrücken, Germany

^{*} CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

The first two authors contributed equally to the article (shared first authorship)

Abstract

Digital life enables situations where people invade other's privacy – sometimes with harmful intentions but often also without such. Given negative effects on victims of privacy invasions, research has examined technical options to prevent privacy-invading behavior (PIB). However, little is known about the sociotechnical environment where PIB occurs. Therefore, our study ($N = 95$) examined possible situational (effort necessary to invade privacy) and individual determinants (e.g., personality) of PIB in a three-phase experiment. 1) Laboratory phase: participants were immersed into the scenario; 2) privacy-invasion-phase at home: automatically and covertly capturing participants' PIB; 3) debriefing-phase at home: capturing whether participants admit PIB. Our results contribute to understanding the sociotechnical environment in which PIB occurs showing that most participants engaged in PIB, that the likelihood of PIB increased when it required less effort, that participants less likely admitted PIB for more sensitive information, and that individual characteristics affected whether participants admitted PIB. We discuss implications for privacy research and design.

1 Introduction

Everyday, people provide private information in digital spaces. This way, we reveal information that attracts the attention of companies, governments, but also of people in our every day's life [1]. For example, we might observe somebody else's smartphone conversations while sitting in the bus (i.e., "shoulder surfing"; [2]). Sometimes it may be tempting to look into someone's browser history when they have left their device unattended [3] or to read an email that has accidentally been sent to the wrong recipient [4]. These examples describe behavior where people access private information of others – not necessarily to do any harm but due to curiosity [2]. We subsume this behavior under the term privacy-invading behavior (PIB).

Although PIB may not be intended to harm anyone, it seems to be socially unacceptable and can lead to negative conse-

quences [2, 5]. Specifically, PIB evokes negative feelings for people whose privacy is being invaded (e.g., feeling observed, harassed; [5]). Furthermore, experiences with PIB can crucially affect future social interactions or handling of sensitive information [6, 7] (e.g., stop using social media; remain overcautious in digital communication). Although for people engaging in PIB it can come with positive feelings such as amusement, they can also end up feeling uneasy or guilty [2].

PIB happens frequently [2] but research knows little about *when* such behavior becomes more likely and about *who* will be more likely to engage in such behavior. Moreover, to the best of our knowledge what we know stems from correlative research which is understandable given the possible ethical issues associated with experimental intervention studies. To overcome this limitation and to shed further light on situational and individual characteristics that determine PIB in digital spaces, we conducted an experimental study including three phases in a highly controlled but nevertheless realistic setting enabling to control ethical issues associated with this kind of research. In phase 1, participants were asked to provide private information about themselves in a laboratory study. In this phase, participants also responded to questionnaires capturing individual characteristics (e.g., personality). In phase 2, participants received an email including private information captured in phase 1 from a supposed other participant (a text and a video file). This way, participants were given the opportunity to show PIB by accessing information of the other person. We manipulated the "necessary effort" to access these information: Whereas one half of participants was able to access the information directly via a link, the other half needed to insert a password that was easy to guess. Access of the files was tracked. In phase 3, participants received another email telling them that there has been an error during the sending of emails in phase 2. Then, they were asked to indicate whether they accessed the files and were asked to justify their possible PIB. With these three phases, it was possible to show a) that most participants engaged in PIB (66% accessed the text, 57% the video file), b) that situational characteristics (i.e., necessary effort) influenced the likelihood that people

invade others' privacy, c) that individual characteristics captured in phase 1 negligibly affected whether people invade other's privacy in phase 2, and d) that the type of information as well as individual characteristics influenced whether people admitted that they accessed other's private information. Our study contributes to our understanding of PIB by shedding light on sociotechnical environments that may promote or prevent PIB.

2 Related Work

2.1 Privacy-invading behavior

Research has investigated PIB predominantly from the perspective of deliberate attacks and with the goal to prevent privacy invasions. For instance, Bošnjak and Brumen [8] reviewed research on ways to prevent harmful privacy invasion through shoulder surfing. Shoulder surfing describes behavior where people covertly observe somebody else's screen of those people's electronic devices [2]. This way, observers can access sensitive data and the victim can be harmed (e.g., by finding out someone's passwords; [9]). As another example of PIB, in an infamous incident, employees at a company selling security cameras had access to customers' cameras for administrative or maintenance reasons but also occasionally accessed camera recordings for other reasons thus invading those people's privacy [10]. Even though some of the aforementioned behavior could be aimed towards harming someone (e.g., blackmailing; voyeurism; [11]), PIB also happens without any intention to harm. Results by Eiband et al. [2] as well as research on social curiosity [12–14] support that PIB is often not aimed towards harming anyone. When people engage in shoulder surfing, they often do so with harmless intentions, and due to curiosity and boredom [2]. Accordingly research sometimes distinguishes PIB according to whether it was intentional (vs. unintentional) and according to the nature of the consequences for the victim (harmful vs. non-harmful) [15]. Following this classification, our study addresses unintended PIB with non-harmful consequences. In our opinion, however, the classification of PIB in such a scheme lacks objectivity and may therefore fall short of the mark: Even though PIB might be without any harmful intent, people might later start to use the collected information in a harmful way. For example, pupils might make fun of another pupil because of watching a video they deem to be "uncool". Collecting this information by shoulder surfing might have happened without a certain intention but results in negative consequences. In addition, even if a certain information is not used in a harmful way, the person whose privacy is invaded may still feel attacked or angry because someone accessed information without consent. Thus, even benign PIB without intent may result in negative consequences for the victim.

2.2 Determinants of PIB

Even without harmful intentions, PIB can involve negative consequences for people whose privacy is being invaded. To prevent such negative outcomes, it is crucial to understand such behavior and to prevent privacy invasions. Whereas technical means to hinder PIB are one important factor [8], examining why people engage in such behavior helps to better understand sociotechnical environments that affect PIB [2].

Initial work has investigated the determinants of PIB. In this regard, Eiband et al. [2] was the main inspiration for the current study. They used a survey to understand the situations where shoulder surfing happens and to investigate motivations that may determine shoulder surfing. From their results, we identified two possible categories of determinants of PIB: 1) situational determinants, 2) individual characteristics. However, the results by Eiband et al. [2] stemmed from a survey where participants reflected situations where they engaged in PIB in the past, only allowing for post-hoc interpretations of possible determinants of PIB. We aimed to go beyond these results and examine the hypothesis that situational and individual determinants affect PIB by experimentally investigating their possible effects.

Regarding situational determinants, arguably there is a large variety of situational determinants that may affect PIB. In line with the findings by Eiband et al. [2] we focus on *effort to access private information*. Their participants reported that other people's electronic devices were in their line of sight which made them inadvertently watch other people's interaction on their devices. Consequently, in this situation PIB may happen because people do not need to take efforts to engage in behavior that allows them to access private information. In digital environments, passwords are one example to make it more effortful to access someone's private information. Uploading files into a shared cloud and using a password for these files increases the effort necessary to access this information. This is also true if the password is easily available (e.g., because it is written down in another file) or guessable (because the password is the other persons' date of birth, which is still often the case; [16]).

In our study, we experimentally manipulated whether more effort was necessary to access someone's private information. Specifically, participants in our study received an email that was clearly addressed to a different person. In one condition, private information was password-encrypted (with the possibility to derive the password from the information available in the email). Participants thus needed to find out the password, type in the password, and only then had access to private information of the other person – they needed to show effort before they could access private information. In contrast, in the no-password condition, participants had to click on a link to readily access the private information. In line with this argumentation, we propose the following hypothesis:

Hypothesis 1: Participants show more PIB when accessing

the private information requires less (i.e., clicking on a link) compared to more effort (entering a password that participants need to derive from an email).

Beyond situational determinants, individual characteristics may influence the likelihood that people engage in PIB. Prior research suggested that a large range of individual characteristics could affect PIB [2, 14]. We thus focus on a range of individual characteristics that can roughly be put into three groups: personality, privacy concerns, and individual characteristics that align with possible motivations behind PIB.

Under personality we subsume the Big Five personality facets openness for experience, conscientiousness, extraversion, agreeableness, and neuroticism [17], honesty-humility as proposed to be the sixth general personality facet [18], as well as the three dark personality facets Machiavellianism, narcissism, and psychopathy [19]. People with a high openness for experience enjoy new activities [20], thus they may also be more interested in acquiring private information of others. Highly conscientious people take obligations seriously [21] and see possibly harmful behavior as more problematic [22]. Thus, they may be less likely to invade other's privacy. Extroverted people are sociable and enjoy getting to know others [21]. Research has shown that introverted people value privacy more strongly which also seems to apply to digital privacy (e.g., they share less information online; [23]). Consequently, it might be less likely that introverted people will invade others privacy. Agreeable people tend to shun away from confrontation and want to get along with others [18]. For them, PIB may be less likely because privacy invasions may offend others. People high on neuroticism worry about many things and are more easily upset [21]. They have been found to be especially concerned about sharing sensitive information [22] thus they might also be less likely to take a look at sensitive information of others. Honesty-humility describes a personality facet of people who are honest and who tend to follow rules [18]. People with high levels of honesty-humility may be less likely to engage in PIB because it constitutes behavior that contradicts the rules of good citizens' behavior.

Dark personality facets subsume Machiavellianism, narcissism, and psychopathy [19]. All of these traits may increase the likelihood that people will engage in PIB. Machiavellianism describes a trait of people who use cunning methods to manipulate others to their own benefit and who have little emotional involvement in interpersonal relations. People high on Machiavellianism have been found to disrespect others' privacy [24, 25] and may thus also be more likely to show PIB. Narcissism reflects a trait of people who believe they are more important than others and who want to be admired. People with strong narcissistic personality tend to disclose more private information on social media [23] and might be more likely to invade other's privacy because they want to compare themselves to others [19]. People with high levels of psychopathy lack empathy and remorse and are less concerned with morality of their actions [19]. They may thus be

more likely to invade other's privacy because they may not realize that this is uncomfortable for the person whose privacy is being invaded.

Privacy concerns might additionally be an important determinant of individuals' PIB. People with strong privacy concerns are sensitive about their private data and more generally about the topic of privacy [7, 26]. Consequently, such people may be less likely to engage in PIB as they are more aware of the sensitivity of private information. Furthermore, they may less likely access others private information because they themselves would not want their privacy to be invaded.

Finally, we hypothesize that people with tendencies for online exhibitionism, thrill-seeking, and social curiosity (i.e., individual characteristics that align with possible motivations behind PIB) make PIB more likely. Online exhibitionism describes a characteristic of people who enjoy presenting themselves on the internet [27]. As people who engage in online exhibitionism reveal much private information, they might also be more interested in other's private information. Thrill-seekers are people who like to engage in dangerous and semi-legal activities for the sake of experiencing novelty [13]. Thrill-seekers may enjoy that they are engaging in unaccepted behavior thus being more likely to invade others privacy. Finally, social curiosity describes a characteristic of people who are interested in the lives of other people, but in extreme versions can also be a characteristic of people who enjoy observing others without their knowledge [14]. This indicates that social curious people may be more likely to invade other's privacy.

In sum, for all these individual characteristics, there is reason to believe that they can influence PIB. We thus propose the following research question (RQ):

RQ1: Do individual characteristics affect the likelihood of PIB?

2.3 Admitting PIB

Since PIB is socially unacceptable and associated with negative consequences for the parties involved [2, 5], admitting such behavior might not be easy and may depend on situational and individual characteristics. Regarding situational determinants, if it required more effort to access someone's private information, wrongdoing may be more salient and it might be less plausible to deny it. In our case, if participants only need to click on a link, they can admit that they did so because they can plausibly say it happened because they wanted to check what kind of information are provided behind these links. However, if people have to find out a password, type in this password and thus take effort to access the provided information, it is less plausible to deny that they accessed private information "by accident". Thus, these people may be less likely to admit that they accessed private information due to being aware of their wrongdoing.

Hypothesis 2: Participants will less likely admit PIB when

accessing the private information requires less compared to more effort.

Regarding individual characteristics, some of the aforementioned characteristics may increase the likelihood that people admit that they engaged in PIB, whereas others may make it less likely. Specifically, conscientiousness, agreeableness, and honesty-humility may increase the likelihood that people admit wrongdoing. In contrast, the dark personality facets may decrease this likelihood. For the other individual characteristics we capture in our study, it is less straightforward to propose associations with admitting PIB. However, to explore the relation of individual determinants and admitting PIB we propose,

RQ2: Do individual characteristics affect the rate with which people admit PIB?

3 Method

Following suggestions to prevent typical fallacies of non-reproducible science [28], we preregistered our hypotheses and research questions, dependent variables that we wanted to capture, experimental manipulations, data analysis plan, and planned number of participants before data collection started. While conducting the experiment and the analyses, we followed our preregistration available at <https://aspredicted.org/e4m32.pdf>. The materials, data, and analysis to reproduce the current findings are available at <https://osf.io/zcq2e>.

3.1 Ethics statement

The first authors' university's ethical review board evaluated and approved this research project. Since this study included deception, a complex design, and a sensitive context, a requirement for ethics approval was a controlled environment that enabled managing participant flow, ensuring that participants do not reveal the purpose of the study to others, and ensuring that participants are contactable for debriefing. This was only possible with a student sample. To enable informed consent, we debriefed participants about the study objective, informed that participation was voluntary and highlighted repeatedly that participants could withdraw from the study. Further, we did not collect any personally identifiable information, and none of the authors were instructors of the student participants to ensure voluntariness. The evaluation materials that were sent to the participants were created together with a professional actress. At the end of phase 3, we conducted a formal debriefing for participants, informing them of the true purpose of this study and assuring them that no personal information was stored from them.

3.2 Procedure

The study consisted of three phases. In phase 1, participants took part in job interview study in a laboratory setting which served as the cover story for our study. In phase 2 (one day after phase 1), participants received an email with the experimental manipulation. In phase 3 (two days after phase 2), there was a post-experimental survey. Figure 2 shows a flow chart summarizing the study procedure.

3.2.1 Phase 1: Laboratory examination

Participants were invited to our laboratory and were told that they will conduct a study on the automatic evaluation of job interviews. After giving their informed consent, participants were asked to complete a questionnaire capturing demographic data and individual characteristics (see section 3.3). Then, they participated in a mock job interview with a trained interviewer. During the interview, a video camera filmed participants, and they were able to see that a video recording software recorded them on a nearby computer monitor. This was done to ensure that participants believed that they were recorded during the interview and later automatically assessed by a computer program (however, we never actually recorded the interviews). The interview started with common questions (e.g., tell us about your strengths) and continued with increasingly personal questions (for instance asking about participants' family planning, marital status; we chose questions that are illegal in selection interviews but where research has shown that they are common in practice [29]). To increase plausibility of our cover-story, after the interview, participants responded to questions assessing their perception of the interview process (e.g., regarding fairness of the process; [30]). However, the only purpose of the interview was to draw attention to the interview setting and to make participants believe that they, and other participants who would take part in the study, had to share private information with the interviewer and that this information was recorded. In the end of phase 1, participants were told that they would receive an evaluation of their interview via email.

3.2.2 Phase 2: At home, privacy-invasion phase including email with experimental manipulation

One day after the interview, participants received an email containing links to a text evaluation and a video of the interview (see Appendix Table 3 for the email). However, the email was addressed to another supposed participant of the same study named "Luca". This way, participants were tricked to believe that they received an email actually intended for another person. In the email, Luca was thanked for their participation and was informed that they will receive the evaluation of their interview via two links contained in this email. The email further informed the recipient that the first link will direct them to a text file that presents an evaluation of the

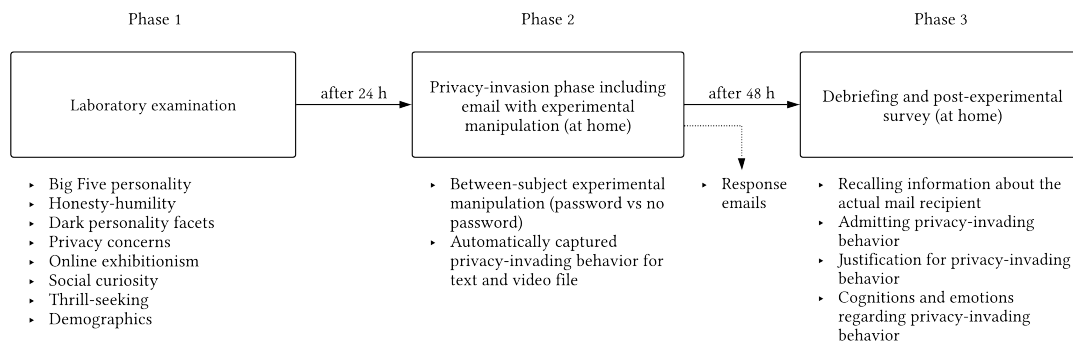


Figure 1: Flow chart of the study procedure. Items below the boxes indicate the measured variables. See Section 3.2 for further details.

interview and that the second link will direct them to a video file of their interview that has been conducted in phase 1.

At this point, we experimentally manipulated the effort necessary to access the other person's private information. In one group, the email contained information that access to the text and video files is password-protected using the email recipient's first name (thus, the correct password was "Luca"). The links for this group led to a page where participants were asked for the password and only after correctly inputting the password the files could be accessed. In the other group, no password protection was mentioned and the links led directly to a page with the respective files. The web pages containing the text and the video file logged the duration of accessing the information.

At this point, some people wrote us that they received a mail containing evaluation materials from another person. We responded that we will clarify what happened but did not disclose our experiment. Disclosing our experiment would have destroyed our manipulation and participants who wrote us an email might have reacted differently to our final survey (see Phase 3).

3.2.3 Phase 3: At home, debriefing-phase and post-experimental survey

Two days after the first email, participants received an apology and information that due to a technical error several emails had been sent to wrong recipients (see Appendix Table 4 for the content of this email). To uphold the cover story, the email asked participants to respond to another questionnaire that was supposedly aimed towards estimating the extent of this supposed error (details on the questionnaire in phase 3 see section 3.3 and Appendix Table 6). First, participants were asked multiple-choice questions to assess whether they had received an email to make sure that phase 2 worked as intended. Then, participants were asked open-response questions that captured whether participants remembered the name of the person who was the actual recipient of the email, whether and how they realized that they had received a wrong email, and whether they remembered details about the other person.

Afterwards, participants responded to a manipulation check for the password condition (i.e., "Were you asked to insert a password at any point?"). Subsequently, participants were asked whether they accessed the other person's private information, thus recording whether participants admitted their potential PIB. Participants were then asked to justify their possible PIB via an open-ended question.

Then, participants were debriefed about the actual objective of the study. Specifically, they were informed that the objective was to examine whether people would access other's private information when given the opportunity to do so and whether they would admit accessing this private information. Furthermore, we told participants that we covertly captured whether they accessed the other person's text and/or video file thus informing them that we knew whether they accessed the files. We emphasized that we never recorded or stored their interview answers and there was no leakage of their private information. Afterwards, participants were asked about the credibility of the cover story of the study. In addition, we asked participants about their cognitions and emotions regarding the private information that they accessed and they were once more given the opportunity to justify their potential PIB the same way as before the debriefing.

3.2.4 Rationale for the procedure

To provide detail on the rationale behind the procedure of the current study, we now present prerequisites for our study and consequences that have manifested in our procedure.

1. Participants needed to be immersed into a realistic situation. If it would have been too obvious that the study was about PIB, people might adapt their behavior in a socially-desirable way. *Consequence:* We used the cover of a "job interview study". These studies are common at the main authors' research institution. We also embedded the questionnaires that were central to our research questions with diversion questionnaires that maintained the "job interview" facade.
2. We needed the option to experimentally manipulate the

effort necessary to access another person's private information. Thus, we needed a situation where we could make it harder to access private information and at the same time not too hard for laypeople. *Consequence:* During phase 2, we manipulated the necessary effort using an easy-to-derive password. Having a password makes it implausible that participants just clicked on a link "by accident". Moreover, since we decided to have the password be the name of the actual recipient of the email, typing in the correct password makes it less plausible that participants believed that the information in the email was actually intended to be for them.

3. We needed to covertly capture participants' actual PIB. Since PIB is socially unaccepted, it is not possible to ask people overtly whether they performed such behavior because we can expect that even under the assurance of anonymity, people may not honestly report their behavior [2]. *Consequence:* During phase 2, we covertly captured participants' behavior with the text and video files.
4. To enable us to examine whether participants would admit PIB, we needed a situation where they were asked to report whether they engaged in PIB. *Consequence:* In phase 3, we gave participants the opportunity to admit their behavior before they were debriefed about the study objectives.
5. In studies including deception, ethics requirements demand to debrief participants about the study objectives, and about where they can get more information about the study. *Consequence:* In phase 3, we debriefed participants, informed them about the study objectives, and provided them with the contact details of the principal investigators. Also, we checked whether people accessed the debriefing in phase 3 and contacted participants who did not complete phase 3 to also debrief them.

3.3 Measurements

Participants responded to all items on a scale from 1 to 5 ("strongly disagree" to "strongly agree"). For all scales we report Cronbach's α as a measure of reliability. Reliability of all scales was acceptable or good, except for psychopathy which was not used for further analyses. If items were not available in the study language, we applied a team approach (following [31,32]). That is, two researchers independently translated the English items into German, discussed possible disparities and resolved them. See <https://osf.io/zcq2e> for further information on the used items and materials.

Accessing private information. We automatically captured whether people accessed the text and video file. For both files, we also captured how long participants interacted with these files.

Admitting PIB. In phase 3, participants were asked whether they accessed the other person's private information, thus recording whether participants admitted their potential privacy-violating behavior. The respective questions were introduced by the prompt "If you received an email with a personal evaluation that was not intended for you ..." and then captured the possibilities for privacy-violating behavior: "...did you look at the text evaluation file?", and "... did you watch (parts of) the video recording?"

Big Five personality. The Big Five personality dimensions were assessed with the Big Five Inventory (BFI) by John and Srivastava [33] with 44 items in a German version [34]. This inventory captures the personality dimensions openness to experience (e.g., "I see myself as someone who is original, comes up with new ideas"; Cronbach's $\alpha = .85$), conscientiousness (e.g., "I see myself as someone who does a thorough job"; Cronbach's $\alpha = .82$), extraversion (e.g., "I see myself as someone who is talkative"; Cronbach's $\alpha = .88$), agreeableness (e.g., "I see myself as someone who is considerate and kind to almost everyone"; Cronbach's $\alpha = .76$) and neuroticism (e.g., "I see myself as someone who gets nervous easily"; Cronbach's $\alpha = .85$).

Honesty-humility. Honesty-humility was captured using 10 items from the HEXACO Personality Inventory-Revised by Ashton and Lee [35]. A sample item was "I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed" (Cronbach's $\alpha = .76$).

Dark personality facets. Dark personality facets were measured with the German version of the Dirty Dozen by Jonason and Webster [19] by Kűfner et al. [36]. This scale captures the Dark Triad with its three dimensions Machiavellianism (e.g., "I tend to manipulate others to get my way"; Cronbach's $\alpha = .80$), Narcissism (e.g., "I tend to want others to admire me"; Cronbach's $\alpha = .70$), and Psychopathy (e.g., "I tend to lack remorse"; Cronbach's $\alpha = .39$), using 12 items (4 per subscale).

Privacy concerns. Privacy concerns were measured with 6 items by Dinev and Hart [37]. A sample item was "I am concerned that the information I submit on the Internet could be misused" (Cronbach's $\alpha = .84$).

Online exhibitionism. Online exhibitionism was assessed with the Social Exhibitionism on the Internet scale by Vetter et al. [27]. We used the short version of this scale, which measures online social exhibitionism with 8 items. A sample item was "I like to post details of my private life on the internet" (Cronbach's $\alpha = .81$). Due to the too explicit reference to sexuality, the item "I like to use communication platforms on the Internet to share my sexual fantasies with people I do not know" was removed.

Social curiosity. Social curiosity was measured with the 10-item version of the Social Curiosity Scale by Renner [14]. A sample item was "When other people are having a conversation, I like to find out what it's about."; Cronbach's $\alpha = .74$).

Thrill-seeking. Thrill-seeking was captured with the five items of the corresponding dimension from the Five-Dimensional Curiosity Scale by Kashdan et al. [13]. A sample item was "The anxiety of doing something new makes me feel excited and alive" (Cronbach's $\alpha = .79$).

3.4 Qualitative measures

Recalling information. In phase 3 we assessed via open-ended questions whether participants recalled information about the actual mail recipient. Those started with the prompt "If you received an email with a personal evaluation that was not intended for you..." and were followed by three questions to capture a) the name of the supposed other person ("... do you still know to whom this mail was addressed?"), b) whether and how participants realized the alleged error ("... how did you realize that the email was not intended for you?"), and c) details about the supposed other person ("... what information about the other person did you see?").

Justification. Before and after the debriefing in phase 3, participants were asked to justify their possible privacy violations via the open question "If you received an email with a personal evaluation that was not intended for you and you clicked on the link to a text or video file or watched it, why did you do this?".

Cognitions and emotions. Participants were asked to report on their cognitions and emotions regarding the private information that they accessed with the question "What did you think or feel when you were dealing with another person's private information (the text or video file)?"

Response emails. During the study, participants responded to the emails that we had sent them. Since these emails could contain interesting information that informs about participant reactions and behavior, we qualitatively analyzed these emails.

4 Results

4.1 Sample characteristics

Overall, 95 participants took part in our study. All participants gave their informed consent and were compensated for their participation either with course credit or with 5€. Of these participants, 72 (75.8%) were female, 22 (23.2%) male, and one person stated another gender. Participants were $M = 22.96$ ($SD = 5.99$) years old. All participants were undergraduates at a German university and the majority was enrolled in a psychology course (86, 90.5%).

4.2 (Sub-)Samples used for hypothesis testing and control questions

Our entire sample completed the first two phases of the experiment, resulting in data from 95 participants usable for the

analysis of determinants of PIB. In phase 3, 81 participants completed the survey. Accordingly, the subsample for the analyses of the determinants of admitting PIB and our qualitative analyses consisted of data from 81 participants. After the debriefing in phase 3, we asked participants whether they believed the scenario of the study. Overall, most participants agreed that they found the emails and the scenario believable ("agree" was coded with 4, range 1 to 5; $M = 3.81$, $SD = 1.13$).

4.3 Investigating hypotheses and RQs

4.3.1 Data structure and analysis plan

To test our hypotheses and RQs, we collected data on two different measures of PIB for each participant: The access/admission of access of the text file and the access/admission of access of the video file. To account for the data structure with two different measurements of PIB per subject (i.e., nested data) and to test our hypotheses/RQs parsimoniously, we decided to analyze our data using a multi-level logistic regression analysis. For our analyses, we used R 4.4.1 [38] and the lme4 package [39] and followed recommendations from the multi-level analysis literature [40, 41]. More specifically, we opted for a step-by-step approach, starting from the simplest model with no predictors and gradually adding predictors to explain our data. Following best practice [40] we tested each model against the respective previous model and rejected more complex models if they could not explain the data significantly better than more parsimonious models to prevent overparameterization.

For, both, the analysis of determinants of PIB and for the question regarding the admittance of this behavior, we specified three models:

1. The Null-Model, containing only a random intercept for the participants. This models that two data points of the dependent variable belong to one participant, but does not consider any predictors.
2. Model 1 extends the Null-Model by adding a predictor for the experimental condition (no-password vs. password). The corresponding fixed effect of this predictor enabled testing of our two hypotheses regarding PIB and admitting this behavior. In addition, a predictor for the type of private information (text vs. video) was added in this model step because we imagined that this may also affect participant behavior.
3. Model 2 extends Model 1 by adding predictors for the individual characteristics in focus of our RQs.

Model comparisons were carried out using AIC values as measures of model fit (lower values indicate better fit) as well as χ^2 -difference tests that compare performance of the models to explain the empirical data.

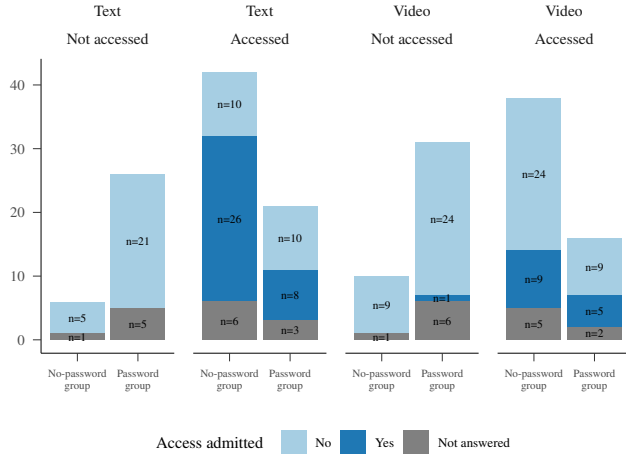


Figure 2: Frequencies of participants per experimental condition who accessed the text and/or video file, of participants who admitted respective behavior, and of participants who did not respond to the respective questions.

4.3.2 Determinants of PIB

Hypothesis 1 stated that participants will show more PIB when accessing the private information requires less effort. Figure 2 provides descriptive information on the number of people who accessed the text and video file, as well as about how many people admitted doing so. In the no-password group, 42 (87.5%) of 48 participants accessed the text file. In the group with password protection, only 21 (44.7%) of 47, accessed the text file. Regarding the video file, in the no-password group, 38 (79.1%) of 48 participants accessed the video file. In the group with password protection, only 16 (34.0%) of 47 participants accessed the video file.

The statistical testing of Hypothesis 1 and the examination of RQ1 followed the multi-level logistic regression approach described in section 4.3.1. Table 1 shows the results of the model comparisons. Model 1 predicted participants' PIB significantly better than the Null-Model ($\chi^2(2) = 35.73$, $p < .001$). Model 2 did not demonstrate a better fit to the data than Model 1 ($\chi^2(12) = 10.06$, $p = .611$). Accordingly, Model 1 was selected as the final model for the dependent variable PIB. Since Model 2 did not explain the empirical data better than Model 1, we found no evidence for any effect of participants' individual characteristics on PIB (see RQ1).

The left side of Table 2 (column: Access) shows the regression coefficients and corresponding significance tests for the final model of PIB. Our results suggest that participants in the password group invaded the other person's privacy significantly less frequently than participants in no-password group (Odds Ratio = 0.03, $p < .001$). This supports Hypothesis 1. We found no effects for the type of information (Odds Ratio = 0.45, $p = .074$).

We further explored our data by examining for how long

Model	AIC	ICC	R^2	LogLik	χ^2 (df)
Dependent variable: Access ($N = 95$, $Obs. = 190$)					
Null-Model	240.21	.58		-118.11	
Model 1 [†]	208.48	.53	.30	-100.24	35.73* (2)
Model 2	222.43	.47	.40	-95.21	10.06 (12)
Dependent variable: Admittance ($N = 60$, $Obs. = 101$) ^b					
Null-Model	143.77	- ^a		-69.88	
Model 1	134.12	.22	.18	-63.06	13.65* (2)
Model 2 [†]	126.50	.02	.54	-47.25	31.62* (12)

Table 1: Model comparisons for the dependent variables accessing private information (Access) and admitting PIB (Admittance).

Note. χ^2 -scores and df reflect the comparison between the models in the current row vs. the previous row (* denotes a significant improvement). R^2 denotes the marginal R^2 and reflects the proportion of total empirical variance explained by fixed effects only (see [42]) and is thus omitted for the Null-Models.

^aThe Null-Model has a singular fit, meaning that one of the variance components in the model has been estimated as zero. In this case, the variance of the random intercepts for the subjects has been estimated to zero, therefore no ICC can be calculated.

^b21 participants were excluded because they did not access neither text nor video file and 14 participants were excluded because they did not answer the survey in phase 3.

[†] indicates the best model according to model comparison.

participants accessed the respective files. The text file was accessed for a mean of $M = 93.13$ ($SD = 137.09$, Median = 35.00) seconds, where 75% of participants accessed the text for longer than 20.5 seconds. Whereas the no-password group accessed the text for a mean of $M = 105.67$ ($SD = 148.84$, Median = 36.50) seconds, the password group accessed the text for a mean of $M = 68.05$ ($SD = 108.87$, Median = 25.00) seconds. The video file was accessed for a mean of $M = 191.82$ ($SD = 244.91$, Median = 60.00) seconds and 75% of the participants who accessed the video did so for longer than 14.50 seconds. Whereas the no-password group accessed the video for a mean of $M = 202.92$ ($SD = 257.69$, Median = 64.00) seconds, the password group accessed the video for a mean of $M = 165.44$ ($SD = 216.99$, Median = 60.00) seconds.

4.3.3 Determinants of admitting PIB

Hypothesis 2 proposed that participants will less likely admit PIB when accessing the private information requires more effort. Regarding the text file, 32 participants did not access this file and 9 participants did not respond to the question whether they accessed the text file. Because it was not possible to determine admittance for those participants, they were excluded from the analysis. In the no-password group, 36 participants accessed the text file and 26 of them (72.7%) admitted doing so. In the password group, 18 participants accessed the text file and 8 (44.4%) admitted doing so. Regarding the video file, 42 participants did not access this file and 7 participants did

	Final Model Access				Final Model Admittance			
	Odds Ratio	β	95% CI	<i>p</i>	Odds Ratio	β	95% CI	<i>p</i>
Null-Model								
Intercept	20.72	3.03	1.49 – 5.57	<.001	0.33	-1.10	-10.67 – 8.47	.821
Model 1								
Type of Information ^a	0.45	-0.79	-1.66 – 0.08	.074	0.10	-2.27	-3.75 – -0.80	.003
Experimental Condition ^b	0.03	-3.39	-5.19 – -1.59	<.001	0.88	-0.13	-1.42 – 1.17	.849
Model 2								
Openness					2.17	0.77	-0.22 – 1.77	.126
Conscientiousness					0.88	-0.13	-1.11 – 0.86	.804
Extraversion					3.26	1.18	0.10 – 2.26	.032
Agreeableness					0.23	-1.46	-3.06 – 0.14	.073
Neuroticism					0.89	-0.12	-1.03 – 0.79	.800
Honesty-humility					1.39	0.33	-0.91 – 1.56	.603
Machiavellianism					0.15	-1.91	-3.27 – -0.55	.006
Narcissism					3.75	1.32	-0.04 – 2.69	.058
Privacy concerns					1.01	0.01	-0.84 – 0.87	.976
Thrill-seeking					0.52	-0.65	-0.15 – 0.21	.139
Social curiosity					0.89	-0.12	-1.50 – 1.26	.864
Online exhibitionism					4.69	1.55	0.12 – 2.12	.034

Table 2: Regression coefficients of the final models for the dependent variables PIB (Access) and admitting PIB (Admittance). Note. Significant *p*-values ($\alpha < .05$) are printed bold. ^aReference category = text. ^bReference category = no password.

not respond to the question whether they accessed the video file. Because it was not possible to determine admittance for those participants, they were excluded from the analysis. In the no-password group, 33 participants accessed the video and 9 (27.3%) participants admitted doing so. In the password group, 14 participants accessed the video and 5 (35.7%) admitted doing so (see also Figure 2).

Testing of Hypothesis 2 and examining RQ2 followed the multi-level logistic regression approach described in section 4.3.1. Table 1 shows the results of the model comparisons. Model 1 predicted significantly better whether participants admitted their PIB than the Null-Model ($\chi^2(2) = 13.65, p = .001$). Model 2 showed an even better fit to the data than Model 1 ($\chi^2(12) = 31.62, p = .002$). Accordingly, Model 2 was selected as the final model for the further analysis.

The right side of Table 2 (column: Admittance) shows the regression coefficients and corresponding significance tests for the final model of admitting PIB. Participants in the password group were equally likely to admit their PIB like those in the no-password group (Odds Ratio = 0.88, $p = .849$). Accordingly, there was no support for Hypothesis 2. Unexpectedly, we found a significant effects for the type of information (Odds Ratio = 0.10, $p = .003$) in this model. Participants less likely admitted that they had accessed the video (14 out of 47, 29.8%) compared to the text file (34 out of 54, 63.0%). One possible explanation for this effect is that accessing the video file was associated with a stronger feeling of wrongdoing and was thus overall a behavior that people

would less likely admit compared to accessing a text file.

RQ2 asked whether individual characteristics affect whether people admit PIB. Based on our model comparisons between Model 1 and Model 2, we conclude that certain individual characteristics affected whether participants admitted PIB. Specifically, higher scores on online exhibitionism (Odds Ratio = 4.69, $p = .034$) and extraversion (Odds Ratio = 3.26, $p = .032$) made it more likely that participants admitted their behavior, whereas higher scores on Machiavellianism (Odds Ratio = 0.15, $p = .006$) made it less likely (see right side of Table 2).

4.3.4 Qualitative results

For further insights regarding why participants behaved and reacted the way they did, we used qualitative analyses to obtain further information from our phase 3 questionnaire and from the emails that participants had sent us. We followed suggestions for reflexive thematic analyses by Braun and Clarke as qualitative analyses [43, 44]. Specifically, in a first step, we coded the text passages in response to the qualitative questions. Second, one of the authors and one research assistant derived superordinate topics from these codings. Third, two independent raters coded all text passages again meaning that they independently assigned text passages to the aforementioned topics. This allowed us to determine reliability of the topics we found in the qualitative analyses. In our case, we calculated interrater reliability (i.e., the agreement between

the two raters in assigning the text passages to the superordinate categories). Precisely, we used Cohen's κ which is calculated using the percentage of matches out of the total number of codings, plus adjusting for the probability of random matches. All of our qualitative analyses showed good to excellent reliability with Cohen's κ values between .65 and .91. The procedure was the same for all qualitative questions from the phase 3 questionnaire as well as for the email texts. Of the 95 participants, 14 participants did not respond to the post-experimental survey so it was not possible to include them in the qualitative analysis, accordingly, the following analyses are based on a sample of 81 participants.

Recalling information about the actual mail recipient.

The majority of participants recalled the name of the actual mail recipient (70, 86.4%). Participants were also asked whether they recalled information that made them aware of the fact that the email was not meant for them. Many participants noticed through the salutation in the email that it was not their name (73, 90.1%). Participants also reported that they noticed that the email was not meant for them due to the text file (16, 19.8%), the video file (19, 23.5%), or the password (8, 9.9%; i.e. they typed in their own first name to access the files instead of the name "Luca" which was the name of the correct recipient). In summary, these results indicated that the majority of participants realized that the information they received was meant for another person. Importantly, 90.1% of participants already realized through the salutation in the email that this email was not meant for them.

Justification for PIB. Many participants justified accessing the text and/or video by stating that they wanted to check whether the email salutation was just addressed to the wrong person but the files were the correct ones (46, 56.8%). Others remarked that they were interested in the analysis (13, 16.0%) or that they did not notice the wrong salutation (8, 9.8%). A minority stated that they believed that we send a wrong information on purpose (2, 2.5%). These findings imply that many participants justified their behavior with "checking behavior" – they supposedly wanted to check whether the provided information was really not meant for their eyes. Fewer participants justified their behavior by reporting that they were interested in the other person's information. After the debriefing, participants' justifications did not change compared to pre-briefing meaning that they did not change their justification after being informed that we had covertly captured their PIB in phase 2.

Cognitions and emotions regarding PIB. Our results revealed that 30 (37.0%) participants reported negative feelings such as guilt, shame, or concern, 17 (21.0%) participants reported no emotional involvement, and 17 (21.0%) reported that they were concerned about their own private information and who might have access to it. Furthermore, 25 (30.9%) participants expressed empathy with the actual recipient and that they felt bad for them that other people can access their information. Other participants were angry about the mistake (7, 8.6%) and some were suspicious whether this was part

of the study or not (6, 7.4%). In sum, these results indicate that participants were (mostly negative) emotionally involved when they realized that they had access to another person's private information. Additionally, participants felt empathy for the other person and at the same time were worried about their own private information as access to another person's private information has made concerns about participants' own privacy salient.

Response emails. Of the 95 participants, 76 (80.0%) responded to our email in phase 2. Of those, 75 (99.0%) wrote that there must have been a mistake during the sending of the mail, that it should be sent to another person or that they are not the person who was addressed in the original email. Furthermore, 14 (18.0%) participants reported concern or anger in their email. They insisted that private data should be handled more conscientiously, that authorities should be informed, and that they wanted immediate clarification of the issue. There were 2 (3.0%) participants who personally came to the laboratory where phase 1 was conducted to contact the research assistants who conducted the study. Also, 14 (18.0%) participants expressed privacy concerns (e.g., concerns about what has happened to their own data). Additionally, 13 (17.0%) participants wanted to make us aware of their privacy-respecting behavior in that they reported that they did not look at anything, that they stopped immediately after realizing that it was not their information, or that they deleted the material instantly. Also, 7 (9.0%) participants explicitly mentioned the full name of the supposed other person. Since the full name of this person was only visible after having clicked on at least one of the links, this indicates that they have accessed the private information. Finally, only 2 (3.0%) participants saw through our mock scenario and wrote that they believe that our email was sent as part of the experimental procedure.

5 Discussion

The goal of this paper was to enhance our understanding of sociotechnical environments that may promote or prevent digital PIB by experimentally investigating situational and individual determinants of PIB that have been proposed in prior research [2]. The main findings of our study are that a) a majority of participants showed PIB, b) many participants had negative feelings about access to another person's private information, c) PIB was less likely when it required more effort to access private information, d) participants were less likely to admit PIB if they accessed the video file compared to the text file, and e) individual characteristics (e.g., personality) only had a minor influence on PIB but influenced whether people admitted such behavior. In sum, our findings indicate that it is less a matter of individual characteristics that drove our participants to engage in PIB but they may have been tempted by effortless access to private information. Furthermore, admitting this kind of behavior seems to differ with respect to the kind of information that was invaded and also

depends on people's individual characteristics.

5.1 Privacy-invading behavior

Our study supports research that implied that people are tempted to behave in a privacy-invading manner if given the opportunity [2, 11, 45]. In fact, a majority of our participants accessed private information that was clearly addressed to another person. They mostly reported that they had accessed the files to check whether it was really not information that was meant for them. Clearly, there is a high probability that participants clicked on the links to check whether the information is really not supposed to be for their eyes. However, since a large proportion of participants accessed the text and video files for more than just a few seconds, we argue that participants' behavior did not just reflect "checking behavior". We propose that some participants used checking behavior also as a socially desirable response when asked for the reasons for socially unacceptable behavior [46]. In the case of the password group, checking behavior seems even less plausible because they had to enter the original recipient's first name as a password. Still, in the password group, nearly half of participants accessed the text and almost a third of participants the video file. We thus argue that our participants were aware that accessing others' private information is socially unacceptable which is supported by the number of participants who a) responded to our mail to clarify that they had received a mail that was intended for another person, b) reported negative feelings about accessing another person's private information, and c) described concerns about what might have happened to their own data. Yet, many participants still accessed the other person's private information, watched it for more than just a few seconds, and thus showed PIB.

Furthermore, our study supports that situational characteristics can affect PIB: more effort necessary to access others' private information made it less likely that participants showed PIB. Since deriving the password from the information in the email was easy, our study additionally showed that already low required effort decreases the likelihood that people invade others' privacy. This finding is in line with research on technical design to prevent PIB that has shown that even small changes and small increases of possible required effort to observe private information (e.g., not using graphical passwords; increasing the length of passwords; [8, 47]) can prevent privacy invasions.

Beyond the effort necessary to access private information, future research could explore other situational characteristics that may influence PIB. In hindsight, it is possible that having a password did not only affect the necessary effort to access private information but also whether there is an active action necessary, and/or whether available information is considered to be private. First, a password makes it a more active action to access private information compared to just clicking on a link since this may happen nearly automatically [48]. Second,

a password might make it salient that information secured with the password is private. In our study, participants in the no-password condition may have been less aware that information accessible through the links is private information worth protecting. In contrast, "protection" and maybe also "privacy" are salient attributes when something is password-encrypted. With our study, we cannot be sure which of these factors were the most influential to reduce PIB but future research can use our analysis as a starting point to investigate the importance of further situational determinants of PIB.

In contrast to situational characteristics, individual characteristics negligibly affected whether participants engaged in PIB. This could mean that there is not much difference between people regarding PIB. Especially in the case of characteristics such as social curiosity where an association with PIB seems straightforward [11, 14] this finding is surprising. One explanation for this finding and a limitation to our study is that participants were predominantly female and mostly students. Although there was some variance regarding individual differences even in our homogeneous sample, there may be less compared to a more representative sample thus reducing the potential to reveal possible effects of individual differences. Therefore, we do not dismiss that there are individual characteristics that influence PIB but still conclude that situational may be more influential than individual characteristics. In other words, it is less a question of who shows PIB but when and under what circumstances.

5.2 Admitting PIB

We hypothesized that situational characteristics that make PIB less likely would also make it less likely to admit PIB but found no support for this hypothesis. Instead, we unexpectedly found that admitting PIB was less likely for the video than for the text file. Possibly, participants perceived the video as containing more sensitive information than the text file. We do not want to imply that videos are always considered to include more sensitive information than text files; there clearly is textual information that will be considered very sensitive (e.g., bank account information). Nevertheless, our findings support research indicating that people ascribe different value or sensitivity to different information [26] and goes beyond that by showing that people may be less likely to admit that they accessed private information for which they assign particular value. The problem that arises from this is that if people are already less likely to admit wrongdoing in our study, where they did not have to fear any punishment, it is possible that in real-life people will not take responsibility for PIB when accessing sensitive information. On the one hand, this can diminish trust in relationships where admitting wrongdoing could facilitate rebuilding trust [49]. On the other hand, if unintended access to sensitive information makes admittance less likely, this may also decrease the likelihood with which data leaks or data security issues within organizations will

be realized [50]. In other words, if people access sensitive, private information they are not supposed to see, this could reveal security issues. However, if people who access this information are not willing to report doing so, such issues will remain undetected. This is a tentative hypothesis that may be worth investigating in future studies. Furthermore, this finding may be useful for information security training in organizations, where it may be necessary to highlight that access to private information can be a sign for security issues and where interventions may need to be implemented to motivate people to report such possible issues.

Whereas individual characteristics did not affect whether participants engaged in PIB, they did influence whether participants admitted their behavior. Admitting PIB was captured by asking participants whether they accessed private information which may have confronted them with their possible wrongdoing. This confrontation may be the point where people with certain individual characteristics may be more (or less) likely to admit PIB. Our findings implied that more extroverted participants more likely admitted PIB. It is possible that for more introverted participants admitting that they have invaded another person's privacy may be unpleasant because they value privacy more than extroverted ones [23, 51]. Consequently, admitting PIB may have been easier for extroverted compared to introverted participants. Furthermore, participants with stronger tendencies regarding online exhibitionism were more likely to admit PIB. Possibly, they were less likely to believe that accessing private information constitutes problematic behavior because they enjoy presenting private information of themselves online [27]. In line with this, admitting PIB may become more likely if people are less aware that such behavior is inadequate. Finally, participants with higher levels of Machiavellianism were less likely to admit PIB. This finding supports previous work finding that people with high levels of Machiavellianism were more likely to misreport their actual behavior [52]. In summary, our findings imply that engaging in PIB is different from admitting to engage in such behavior. Whereas the former was more strongly influenced by the effort necessary to access private information, the latter was influenced by the type of accessed private information and by individual characteristics.

5.3 Design implications

Although our study was aimed towards a better understanding of sociotechnical environments that affect PIB and not towards deriving design implications, we still want to emphasize design implications of our findings. It may sound obvious but to prevent PIB it makes sense to increase the effort necessary to access private information. On the one hand, our study shows that people are tempted to invade others' privacy when there is no effort required. On the other hand, it shows that even flawed security measures (like easy guessable passwords) can reduce PIB. This behavior affects people whose

privacy is being invaded and those who invaded other's privacy since a significant proportion of our participants reported negative feelings after showing PIB (see also [2]). To decrease the temptation to engage in behavior that has mostly negative consequences, designers and individuals need to consider ways to increase the effort necessary for possible PIB. For designers, research on shoulder surfing provides examples regarding how to design technology to make accessing private information require more effort (for a review see [8]). For individuals, our recommendation is to consider ways, even seemingly simple ones, to at least increase the effort for others to engage in PIB. The same way that curtains prevent privacy invasions, sending out password-encrypted files reduces the temptation for people who may accidentally have access to these files to engage in PIB – even if the password is easily available.

5.4 Limitations and Future Work

There are at least two limitations that we need to address. First, our sample consisted mostly of female psychology students which decreases the generalizability of our findings. This group of participants may differ in age, gender, and personality variables from a broader population. For instance, employees will have different characteristic and also different motivations when given the opportunity to access sensitive information. Second, although we designed our study to reflect a realistic situation (i.e., accessing private information because an email was mistakenly sent to the wrong recipient; [4]), it is still a specific situation. PIB in other contexts may differ from the situation in our study. For instance, shoulder surfing has the possibility that the observed person will realize the privacy invasion and will show a reaction that immediately affects the observer. Nevertheless, we believe that our conclusions hold for other PIB and that it is important to consider ways to increase the effort necessary to engage in PIB.

6 Conclusion

Research seems to only be beginning to understand the sociotechnical environment in which digital PIB occurs [2, 8]. Our study supports that PIB may be an everyday behavior that many people would show if given the opportunity, and whose likelihood is affected by situational characteristics. Furthermore, admitting this behavior will never be easy but also seems to depend on the kind of information that was accessed and on individual characteristics of the person who has engaged in PIB. Consequently, technical design but also individuals' privacy-securing behavior need to be guided towards preventing situations where people may be tempted to access information that is not meant for them.

A Appendix

Subject: Job interview study evaluation

Hey Luca,
Thank you again for your participation in our job interview study.

As already announced, we will send a corresponding evaluation to the study participants for whom our analysis tool strongly deflected at one point of the interview. A part of your behavior seems to have been so remarkable for the algorithm that it gave you extreme values.

Under the following links we have provided the automatically generated evaluation as well as the recording of your interview. In the evaluation, there are also time stamps for the remarkable parts, which the algorithm threw out.

Evaluation: [Link](#)

Video recording: [Link](#)

If you have any questions or want personal feedback just contact me again.

Best regards,
Laura

Table 3: Email from phase 2

Note: Text of the email sent to participants 24 hours after the job interview in the laboratory. The original email was in German and contained links to the research institute and an automatic signature with the contact details of the Lab.

Subject: Error evaluation job interview study

Dear participant,

Unfortunately, due to a technical error, there was some confusion in the sending of the personal evaluations.

In order to estimate the extent of this error, we kindly ask you to fill out the following questionnaire:

[Link](#)

Answering the questionnaire will take about 5-10 minutes and you will receive 0.25 subject hours as compensation for your efforts.

Best regards,
Laura

Table 4: Email from phase 3

Note: Text of the email sent to participants 48 hours after the email from phase 2. The original email was in German and contained links to the research institute and an automatic signature with the contact details of the Lab.

Response emails from participants to the emails from phase 2

information / notification about mistake
information / notification about wrong video
own name explicitly mentioned (not just in closing formula)
own name completely absent (not in text nor in closing formula)
privacy concerns
warning / threatening behavior
integrity / privacy-respecting behavior
full name "Luca Schmidt" mentioned
failed cover story

Table 5: Codebook from the reflexive thematic analysis of the response emails

Note: Cohen's $\kappa = .91$

Scale	Item text	Response format
Big-Five-Inventory [34]	I see myself as someone who . .	1 (strongly disagree) to 5 (strongly agree)
Agreeableness	<ul style="list-style-type: none"> ... tends to find fault with others. (r) ... is helpful and unselfish with others. ... starts quarrels with others. (r) ... has a forgiving nature. ... is generally trusting. ... can be cold and aloof. (r) ... is considerate and kind to almost everyone. ... is sometimes rude to others. (r) ... likes to cooperate with others. 	
Conscientiousness	<ul style="list-style-type: none"> ... does a thorough job. ... can be somewhat careless. (r) ... is a reliable worker. ... tends to be disorganised. (r) ... tends to be lazy. (r) ... perseveres until the task is finished. ... does things efficiently. ... makes plans and follows through with them. ... is easily distracted. (r) 	
Extraversion	<ul style="list-style-type: none"> ... is talkative. ... is reserved. (r) ... is full of energy. ... generates a lot of enthusiasm. ... tends to be quiet. (r) ... has an assertive personality. ... is sometimes shy, inhibited. (r) ... is outgoing, sociable. 	
Neuroticism	<ul style="list-style-type: none"> ... is depressed, blue. ... is relaxed, handles stress well. (r) ... can be tense. ... worries a lot. ... is emotionally stable, not easily upset. (r) ... can be moody. ... remains calm in tense situations. (r) ... gets nervous easily. 	
Openness	<ul style="list-style-type: none"> ... is original, comes up with new ideas. ... is sophisticated in art, music, or literature. ... is curious about many different things. ... is ingenious, a deep thinker. ... has an active imagination. ... is inventive. ... values artistic, aesthetic experiences. ... prefers work that is routine. (r) ... likes to reflect, play with ideas. ... has few artistic interests. (r) 	

Table 6: Items for phase 1 and phase 3

Note: The items for the Social Exhibitionism on the Internet scale (online exhibitionism) are originally in German and were translated to English for the Appendix. (r) = reverse-coded item.

Scale	Item text	Response format
Social curiosity [14]	<p>I'm interested in people.</p> <p>When other people are having a conversation, I like to find out what it's about.</p> <p>I like finding out how others "work."</p> <p>When on the train, I like listening to other people's conversations.</p> <p>I find it fascinating to get to know new people.</p> <p>When people quarrel, I like to know what's going on.</p> <p>When I meet a new person, I am interested in learning more about him/her.</p> <p>Every so often I like to stand at the window and watch what my neighbors are doing.</p> <p>I like to learn about the habits of others.</p> <p>I like to look into other people's lit windows.</p>	1 (strongly disagree) to 5 (strongly agree)
Online exhibitionism [27]	<p>The idea that theoretically millions of people could look at my site on the Internet is appealing to me.</p> <p>I like to post details of my private life on the internet.</p> <p>I enjoy putting intimate details of my private life on the Internet.</p> <p>I don't like the idea that unknown people on the Internet get information about my leisure activities from me. (r)</p> <p>I like to post photos showing me on the internet for everyone to see.</p> <p>I enjoy posting private videos of myself on the web for everyone to see.</p> <p>I struggle with not knowing who is reading the information I provide online. (r)</p>	1 (strongly disagree) to 5 (strongly agree)
Dark personality facets [19, 36]	<p>I tend to manipulate others to get my way.</p> <p>I have used deceit or lied to get my way.</p> <p>I have use flattery to get my way.</p> <p>I tend to exploit others towards my own end.</p> <p>I tend to lack remorse.</p> <p>I tend to not be too concerned with morality or the morality of my actions.</p> <p>I tend to be callous or insensitive.</p> <p>I tend to be cynical.</p> <p>I tend to want others to admire me.</p> <p>I tend to want others to pay attention to me.</p> <p>I tend to seek prestige or status.</p> <p>I tend to expect special favors from others.</p>	1 (strongly disagree) to 5 (strongly agree)
Privacy concerns [37]	<p>I am concerned that the information I submit on the Internet could be misused.</p> <p>When I shop online, I am concerned that the credit card information can be stolen while being transferred on the Internet.</p> <p>I am concerned about submitting information on the Internet, because of what others might do with it.</p> <p>I am concerned about submitting information on the Internet, because it could be used in a way I did not foresee.</p> <p>When I am online, I have the feeling of being watched.</p> <p>When I am online, I have the feeling that all my clicks and actions are being tracked and monitored.</p>	1 (strongly disagree) to 5 (strongly agree)

Table 6: Items for phase 1 and phase 3

Note: The items for the Social Exhibitionism on the Internet scale (online exhibitionism) are originally in German and were translated to English for the Appendix. (r) = reverse-coded item.

Scale	Item text	Response format
Thrill-seeking [13]	The anxiety of doing something new makes me feel excited and alive. Risk-taking is exciting to me. When I have free time, I want to do things that are a little scary. Creating an adventure as I go is much more appealing than a planned adventure. I prefer friends who are excitingly unpredictable.	1 (strongly disagree) to 5 (strongly agree)
Honesty-humility [35]	I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed. If I knew that I could never get caught, I would be willing to steal a million dollars. (r) Having a lot of money is not especially important to me. I think that I am entitled to more respect than the average person is. (r) If I want something from someone, I will laugh at that person's worst jokes. (r) I would never accept a bribe, even if it were very large. I would get a lot of pleasure from owning expensive luxury goods. (r) I want people to know that I am an important person of high status. (r) I wouldn't pretend to like someone just to get that person to do favors for me.	1 (strongly disagree) to 5 (strongly agree)
Controll questions	Did you receive an email from us with an evaluation? Please check your spam folder. Did you find an email from us with an evaluation? If you received an evaluation, was it your own?	1 (yes), 2 (no)
Recalling information about actual recipient	If you received an email with a personal evaluation that was not intended for you do you still know to whom this mail was addressed? ... how did you realize that the email was not intended for you? ... what information about the other person did you see?	open
Admitting privacy-invasive behavior	If you received an email with a personal evaluation that was not intended for you did you read the content of the mail? ... did you click on the link to the evaluation document? ... did you look at the evaluation document? ... did you click on the link to the video recording? ... did you watch (parts of) the video recording? ... was there a password prompt at any point?	1 (yes), 2 (no)
Justification	If you received an email with a personal evaluation that was not intended for you and you clicked on the link to a text or video file or watched it, why did you do this?	open
Credibility	I found it credible that a mistake was made when sending out the evaluations.	1 (strongly disagree) to 5 (strongly agree)
Cognitions and emotions	What did you think or feel when you were dealing with another person's private information? (text or video file)	open
Previous experience	Have you experienced a case of privacy violation before? If you have already experienced a case of privacy violation, briefly describe it.	1 (yes), 2 (no) open

Table 6: Items for phase 1 and phase 3

Note: The items for the Social Exhibitionism on the Internet scale (online exhibitionism) are originally in German and were translated to English for the Appendix. (r) = reverse-coded item.

Recalling information about the actual mail recipient		
Information from person	Recognition other recipient	Name recall
name	name	Luca
evaluation sheet	text file	
video	video	
image	image	
conspicuity from the algorithm	password	
unclear, which information they saw	evaluation (unspecific)	
gender		
age		
Justification, cognitions and emotions regarding privacy-invading behavior		
Cognitions and emotions	Justification	
strange	to check whether the files were the correct ones	
relief that other person reacted similarly	curiosity	
embarrassing	not noticed the wrong salutation	
impressed by the computer evaluation	believe wrong information was sent on purpose	
curiosity	believe attachment is a dummy evaluation	
indifference		
negative feelings		
<i>guilt, shame, concern, shocked, unpleasant, bad feeling, uncomfortable, queasy feeling, bad conscience</i>		
empathy with Luca		
<i>process is unfair, empathy with Luca, feeling sorry for Luca, protect the privacy of the other person</i>		
concern about own data		
<i>threatening, concern about own data</i>		
distrust in the study		
<i>thought that it was intentional, thought that it is only a cover story</i>		
disappointment / anger		
<i>disappointment, furious, anger</i>		

Table 7: Codebook from the reflexive thematic analysis of the qualitative questions from the phase 3 survey

Note: Cohen's $\kappa = .65$

References

- [1] M. Andrejevic, “The work of watching one another: Lateral surveillance, risk, and governance,” *Surveillance & Society*, vol. 2, no. 4, pp. 479–497, 2004. [Online]. Available: <https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/3359>
- [2] M. Eiband, M. Khamis, E. von Zezschwitz, H. Hussmann, and F. Alt, “Understanding shoulder surfing in the wild: Stories from users and observers,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Denver, Colorado, USA: ACM Press, May 2017, pp. 4254–4265. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=3025453.3025636>
- [3] D. Marques, T. Guerreiro, L. Carriço, I. Beschastnikh, and K. Beznosov, “Vulnerability & blame: Making sense of unauthorized access to smartphones,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–13. [Online]. Available: <https://doi.org/10.1145/3290605.3300819>
- [4] E. Lieberman and R. C. Miller, “Facemail: Showing faces of recipients to prevent misdirected email,” in *Proceedings of the 3rd symposium on Usable privacy and security*, ser. SOUPS ’07. New York, NY, USA: Association for Computing Machinery, Jul. 2007, pp. 122–131. [Online]. Available: <https://doi.org/10.1145/1280680.1280696>
- [5] S. Petronio, “Communication privacy management theory,” in *The international encyclopedia of interpersonal communication*, C. R. Berger, M. E. Roloff, S. R. Wilson, J. P. Dillard, J. Caughlin, and D. Solomon, Eds. New York, US: Wiley, 2015, pp. 1–9. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118540190.wbeic132>
- [6] H. J. Smith, S. J. Milberg, and S. J. Burke, “Information privacy: Measuring individuals’ concerns about organizational practices,” *MIS Quarterly*, vol. 20, no. 2, pp. 167–196, Jun. 1996. [Online]. Available: <https://www.jstor.org/stable/249477?origin=crossref>
- [7] N. K. Malhotra, S. S. Kim, and J. Agarwal, “Internet Users’ Information Privacy Concerns (IUIPC): The construct, the scale, and a causal model,” *Information Systems Research*, vol. 15, no. 4, pp. 336–355, Dec. 2004. [Online]. Available: <https://pubsonline.informs.org/doi/10.1287/isre.1040.0032>
- [8] L. Bošnjak and B. Brumen, “Shoulder surfing experiments: A systematic literature review,” *Computers & Security*, vol. 99, pp. 1–34, Dec. 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404820302960>
- [9] W. Goucher, “Look behind you: the dangers of shoulder surfing,” *Computer Fraud & Security*, vol. 2011, no. 11, pp. 17–20, Nov. 2011. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1361372311701166>
- [10] J. Hruska, “Amazon’s ring security camera let employees spy on customers,” Jan. 2019. [Online]. Available: <https://www.extremetech.com/internet/283665-amazons-ring-security-camera-let-employees-spy-on-customers>
- [11] J. A. Litman and M. V. Pezzo, “Dimensionality of interpersonal curiosity,” *Personality and Individual Differences*, vol. 43, pp. 1448–1459, Oct. 2007.
- [12] F.-M. Hartung and B. Renner, “Social curiosity and interpersonal perception: A judge \times trait interaction,” *Personality and Social Psychology Bulletin*, vol. 37, no. 6, pp. 796–814, 2011.
- [13] T. B. Kashdan, M. C. Stikma, D. J. Disabato, P. E. McKnight, J. Bekier, J. Kaji, and R. Lazarus, “The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people,” *Journal of Research in Personality*, vol. 73, pp. 130–149, Apr. 2018. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0092656617301149>
- [14] B. Renner, “Curiosity about people: The development of a social curiosity measure in adults,” *Journal of Personality Assessment*, vol. 87, no. 3, pp. 305–316, Oct. 2006. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327752jpa8703_11
- [15] R. Parks, H. Xu, C.-H. Chu, and P. B. Lowry, “Examining the intended and unintended consequences of organisational privacy safeguards,” *European Journal of Information Systems*, vol. 26, no. 1, pp. 37–65, 2017. [Online]. Available: <https://doi.org/10.1057/s41303-016-0001-6>
- [16] R. Wash, E. Rader, R. Berman, and Z. Wellmer, “Understanding password choices: How frequently entered passwords are re-used across websites,” in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. Denver, CO: USENIX Association, Jun. 2016, pp. 175–188. [Online]. Available: <https://www.usenix.org/conference/soups2016/technical-sessions/presentation/wash>

- [17] R. R. McCrae and P. T. Costa Jr., “The five-factor theory of personality,” in *Handbook of personality: Theory and research*, 3rd ed. New York, NY, US: The Guilford Press, 2008, pp. 159–181.
- [18] K. Lee and M. C. Ashton, “Psychometric properties of the HEXACO personality inventory,” *Multivariate Behavioral Research*, vol. 39, no. 2, pp. 329–358, Apr. 2004. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr3902_8
- [19] P. K. Jonason and G. D. Webster, “The dirty dozen: A concise measure of the dark triad,” *Psychological Assessment*, vol. 22, no. 2, pp. 420–432, Jun. 2010. [Online]. Available: <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0019265>
- [20] T. Halevi, J. Lewis, and N. Memon, “A pilot study of cyber security and privacy related behavior and personality traits,” in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW ’13 Companion. New York, NY, USA: Association for Computing Machinery, May 2013, pp. 737–744. [Online]. Available: <https://doi.org/10.1145/2487788.2488034>
- [21] S. Soldz and G. E. Vaillant, “The Big Five personality traits and the life course: A 45-year longitudinal study,” *Journal of Research in Personality*, vol. 33, no. 2, pp. 208–232, Jun. 1999. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0092656699922432>
- [22] G. Bansal, D. Gefen *et al.*, “The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online,” *Decision Support Systems*, vol. 49, no. 2, pp. 138–150, 2010.
- [23] C. Liu, R. P. Ang, and M. O. Lwin, “Cognitive, personality, and social factors associated with adolescents’ online personal information disclosure,” *Journal of adolescence*, vol. 36, no. 4, pp. 629–638, 2013.
- [24] D. P. Bhave, L. H. Teo, and R. S. Dalal, “Privacy at work: A review and a research agenda for a contested terrain,” *Journal of Management*, vol. 46, no. 1, pp. 127–164, 2020.
- [25] S. J. Winter, A. C. Stylianou, and R. A. Giacalone, “Individual differences in the acceptability of unethical information technology practices: The case of machiavellianism and ethical ideology,” *Journal of Business Ethics*, vol. 54, no. 3, pp. 273–301, 2004.
- [26] H. J. Smith, T. Dinev, and H. Xu, “Information privacy research: an interdisciplinary review,” *MIS Quarterly*, vol. 35, no. 4, pp. 989–1015, 2011. [Online]. Available: <https://www.jstor.org/stable/41409970>
- [27] M. Vetter, C. Eib, S. Hill-Kloss, P. Wollscheid, and D. Hagemann, “Development and validation of a scale for Social Exhibitionism on the Internet (SEXI),” *Diagnostica (Göttingen)*, vol. 60, no. 3, pp. 153–165, 2014. [Online]. Available: <http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-107208>
- [28] J. P. Simmons, L. D. Nelson, and U. Simonsohn, “False-positive psychology,” *Psychological Science*, vol. 22, no. 11, pp. 1359–1366, 2011, pMID: 22006061.
- [29] H. G. Hern Jr, H. J. Alter, C. P. Wills, E. R. Snoey, and B. C. Simon, “How prevalent are potentially illegal questions during residency interviews?” *Academic Medicine*, vol. 88, no. 8, pp. 1116–1121, 2013.
- [30] J. A. Colquitt, “On the dimensionality of organizational justice: A construct validation of a measure,” *Journal of Applied Psychology*, vol. 86, no. 3, pp. 386–400, Jun. 2001. [Online]. Available: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.86.3.386>
- [31] J. A. Harkness, B. Edwards, S. E. Hansen, D. R. Miller, and A. Villar, “Designing questionnaires for multipopulation research,” in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. John Wiley & Sons, Ltd, 2010, pp. 31–57. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470609927.ch3>
- [32] J. A. Harkness, A. Villar, and B. Edwards, “Translation, adaptation, and design,” in *Survey methods in multinational, multiregional, and multicultural contexts*, J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell, and T. W. Smith, Eds. Hoboken, NJ: John Wiley & Sons, Inc., May 2010, pp. 115–140. [Online]. Available: <http://doi.wiley.com/10.1002/9780470609927.ch7>
- [33] O. P. John, S. Srivastava *et al.*, *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*. University of California Berkeley, 1999, vol. 2.
- [34] C. B. Fell and C. J. König, “Cross-cultural differences in applicant faking on personality tests: A 43-nation study,” *Applied Psychology*, vol. 65, no. 4, pp. 671–717, 2016. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/apps.12078>
- [35] M. C. Ashton and K. Lee, “The HEXACO-60: A short measure of the major dimensions of personality,” *Journal of Personality Assessment*, vol. 91, no. 4, pp. 340–345, Jul. 2009. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.1080/00223890902935878>

- [36] A. C. P. Küfner, M. Dufner, and M. D. Back, “Das Dreckige Dutzend und die Niederträchtigen Neun [The dirty dozen and the infamous nine],” *Diagnostica*, vol. 61, no. 2, pp. 76–91, Jan. 2015. [Online]. Available: <https://econtent.hogrefe.com/doi/abs/10.1026/0012-1924/a000124>
- [37] T. Dinev and P. Hart, “Internet privacy concerns and their antecedents - measurement validity and a regression model,” *Behaviour & Information Technology*, vol. 23, no. 6, pp. 413–422, Nov. 2004. [Online]. Available: <https://doi.org/10.1080/01449290410001715723>
- [38] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: <https://www.R-project.org/>
- [39] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *arXiv preprint arXiv:1406.5823*, 2014.
- [40] J. J. Hox, M. Moerbeek, and R. Van de Schoot, *Multi-level analysis: Techniques and applications*. Routledge, 2017.
- [41] F. Steele, “Multilevel models for longitudinal data,” *Journal of the Royal Statistical Society: series A (statistics in society)*, vol. 171, no. 1, pp. 5–19, 2008.
- [42] S. Nakagawa, P. C. D. Johnson, and H. Schielzeth, “The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded,” *Journal of the Royal Society Interface*, vol. 14, no. 134, p. 20170213, 2017.
- [43] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa>
- [44] —, “Reflecting on reflexive thematic analysis,” *Qualitative Research in Sport, Exercise and Health*, vol. 11, no. 4, pp. 589–597, Aug. 2019. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/2159676X.2019.1628806>
- [45] J. R. Frampton and J. Fox, “Monitoring, creeping, or surveillance? A synthesis of online social information seeking concepts,” *Review of Communication Research*, vol. 9, pp. 1–42, 2021. [Online]. Available: <https://www.rcommunicationr.org/index.php/rcr/article/view/75>
- [46] A. J. Nederhof, “Methods of coping with social desirability bias: A review,” *European Journal of Social Psychology*, vol. 15, no. 3, pp. 263–280, 1985. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420150303>
- [47] L. Bošnjak and B. Brumen, “Shoulder surfing: From an experimental study to a comparative framework,” *International Journal of Human-Computer Studies*, vol. 130, pp. 1–20, Oct. 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1071581918305366>
- [48] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson, “Going spear phishing: Exploring embedded training and awareness,” *IEEE Security & Privacy*, vol. 12, no. 1, pp. 28–38, Jan. 2014. [Online]. Available: <http://ieeexplore.ieee.org/document/6585241/>
- [49] N. Gillespie and G. Dietz, “Trust repair after an organization-level failure,” *Academy of Management Review*, vol. 34, no. 1, pp. 127–145, Jan. 2009. [Online]. Available: <https://journals.aom.org/doi/10.5465/amr.2009.35713319>
- [50] P. Mayer, Y. Zou, F. Schaub, and A. J. Aviv, ““Now I’m a bit angry:” Individuals’ awareness, perception, and responses to data breaches that affected them,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 393–410. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity21/presentation/mayer>
- [51] D. L. Stone, “Relationship between introversion/extraversion, values regarding control over information, and perceptions of invasion of privacy,” *Perceptual and Motor Skills*, vol. 62, no. 2, pp. 371–376, 1986.
- [52] P. R. Murphy, “Attitude, machiavellianism and the rationalization of misreporting,” *Accounting, Organizations and Society*, vol. 37, no. 4, pp. 242–259, 2012.

Investigating How University Students in the United States Encounter and Deal With Misinformation in Private WhatsApp Chats During COVID-19

K. J. Kevin Feng
Princeton University

Kevin Song
University of Chicago

Kejing Li
University of Chicago

Oishee Chakrabarti
University of Chicago

Marshini Chetty
University of Chicago

Abstract

Misinformation can spread easily in end-to-end encrypted messaging platforms such as WhatsApp where many groups of people are communicating with each other. Approaches to combat misinformation may also differ amongst younger and older adults. In this paper, we investigate how young adults encountered and dealt with misinformation on WhatsApp in private group chats during the first year of the COVID-19 pandemic. To do so, we conducted a qualitative interview study with 16 WhatsApp users who were university students based in the United States. We uncovered three main findings. First, all participants encountered misinformation multiple times a week in group chats, often attributing the source of misinformation to be well-intentioned family members. Second, although participants were able to identify misinformation and fact-check using diverse methods, they often remained passive to avoid negatively impacting family relations. Third, participants agreed that WhatsApp bears a responsibility to curb misinformation on the platform but expressed concerns about its ability to do so given the platform's steadfast commitment to content privacy. Our findings suggest that conventional content moderation techniques used by open platforms such as Twitter and Facebook are unfit to tackle misinformation on WhatsApp. We offer alternative design suggestions that take into consideration the social nuances and privacy commitments of end-to-end encrypted group chats. Our paper also contributes to discussions between platform designers, researchers, and end users on misinformation in privacy-preserving environments more broadly.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

1 Introduction

WhatsApp is a widely used end-to-end encrypted messaging platform worldwide, with an estimated 74 million users in the United States (U.S.) alone as of 2021 [4]. The platform's widespread usage rose sharply with the global spread of COVID-19. By late March 2020, WhatsApp grew by 40% compared to pre-pandemic months [55]; this growth was likely fueled by its connective capabilities during the pandemic, such as for organizing mutual aid groups [16] and, in the case of millions of immigrants, connecting with family members abroad [42]. WhatsApp's end-to-end encryption [80] means that the platform is unable to easily detect or flag misleading messages, i.e., misinformation¹, which is problematic given its global user base [71]. It has therefore been identified as an effective misinformation pipeline by academics, journalists, and fact-checking organizations [31, 56, 74]. Consequences of this rapid dissemination of misinformation on the platform include the spread of misleading health claims and associated health risks [27, 39], tampering of elections abroad [5], and deaths [10, 34].

Many researchers have studied characteristics of online misinformation including prevalence [1, 22, 38], speed of spread [37], user perceptions [26, 32], and strategic participatory campaigns [67]. However, research on misinformation in WhatsApp specifically has been limited and mainly focuses on users outside of the U.S. [6, 41, 49]. These studies observe user behavior through theoretical frameworks and collect message content from large public WhatsApp groups [31, 41, 46, 49] rather than using empirical user studies of private chats² [25, 45, 46, 57, 58]. Private chats yield valuable insights into users' daily communication practices

¹In this paper, we use the definition of misinformation on social media presented by Wu et al. [85]: an umbrella term that includes all false or inaccurate information that is spread.

²A WhatsApp *private* chat can only be joined with an invitation link that is not typically shared publicly or when a group admin adds members to a group chat. A WhatsApp *public* chat can be joined by anyone on the Internet via an invitation link that is usually posted on a public website, making it easier for researchers to study.

since WhatsApp users mainly communicate in small, pre-selected groups of people [64], notably families. Although misinformation within smaller private group chats may not be broadcasted to large audiences at once, they can still reach high numbers of users through group chats' popularity and frequent forwarding activity between chats [46].

To properly combat misinformation on WhatsApp, we need a better understanding of how WhatsApp users deal with misleading messages, particularly in private chats. Since there is a generally an unreciprocated concern directed towards older family members about health misinformation due to them being perceived as a vulnerable population on the Internet [69], we also need to balance this out with an investigation of the perspectives of younger adults around misinformation on WhatsApp. To address this research gap, we conducted interviews with 16 young adults who were university students in the U.S.—a country with the third most WhatsApp users globally [70]—to better understand their experiences with COVID-19-related misinformation in close-knit private chats. Our study was driven by the following research questions:

- **RQ1:** How do U.S.-based university students currently perceive and encounter misinformation in WhatsApp private chats?
- **RQ2:** How do U.S.-based university students identify misinformation on the platform and respond to it?
- **RQ3:** How aware are U.S.-based university students of current WhatsApp features to combat misinformation and what would improve how the platform handles misinformation?

We uncovered three main findings. First, all participants encountered misinformation multiple times a week in group chats, often attributing the source of misinformation to be well-intentioned family members. Most participants also claimed not to forward information without fact-checking first. Second, although participants were able to identify misinformation using similar indicators seen in previous studies on other social media platforms [26, 32, 47], they often did not confront misinformation senders to avoid negatively impacting family relations. Third, participants were not aware of most existing features to combat misinformation on WhatsApp and agreed that WhatsApp bears a responsibility to curb misinformation on the platform. However, participants expressed concerns about its ability to do so given the platform's commitment to content privacy. Based on our findings, we suggest, assuming users can be made more aware of new features, that empowering users on the platform to better fact-check or flag misinformation for themselves may combat the effects of misleading content. We also suggest that designs that allow users to subtly provide resources for misleading messages within a group could offset the power dynamics in chats that prevent users from confronting misinformation

senders. Future work should investigate older adults' role in misinformation on WhatsApp and how to educate users about misinformation leveraging the fact that misinformation is often spread out of care and not malicious intent.

To summarize, our primary contributions are:

- **Findings from a U.S.-based WhatsApp user study:** we contribute novel insights about how U.S.-based WhatsApp university students in our study perceived and reacted to misinformation in private WhatsApp chats. For instance, we found that our participants felt that misinformation was often sent to them from well-intentioned family members out of care for others and that family dynamics make it harder for younger adults to confront older misinformation senders. This contributes to a growing set of studies of public WhatsApp chat data [25, 45, 46, 57, 58].
- We corroborate findings from misinformation studies on other social media platforms such as Facebook and news [26, 32, 47] about the indicators people use to identify misleading content; adding a novel finding about how WhatsApp users weigh the relationship with a misinformation sender to determine if content can be trusted.
- Finally, our paper adds to the literature on how to tackle misinformation in end-to-end encrypted platforms that conventional content moderation techniques used by open platforms such as Twitter and Facebook cannot address, owing to the tradeoff between user-privacy and having to access data for labeling content [43].

Next, we describe related work, our methods, findings, and discussion points before concluding the paper.

2 Background and Related Work

2.1 Misinformation on Social Media

COVID-19 has swept the world, and so has the misinformation associated with it [6, 9, 36, 61, 72]. Kouzy et al. [36] estimates 25% of tweets include misinformation about the pandemic, while 17% include unverifiable information. To date, researchers have studied misinformation and its dissemination through social media extensively [3, 6, 15, 26, 32, 39, 50, 67]. Studies have also shown that misinformation's impact is global, from increasing tensions between neighboring countries [28], to suppressing government-critical voices within borders [52], to interfering with democratic elections [3, 14, 51]. Yet, the scale of social media and the Internet's replacement of expert advice make combating misinformation challenging [3, 39, 67].

To combat misinformation, some studies have explored users' motives for spreading news and misinformation on social media specifically and found that while most participants shared news to inform others, a third share for others'

entertainment, with 19% doing so just to upset others [15]. Sharing misinformation can be influenced by culture as shown by Madrid-Morales et al. [50] who found that sharing habits differed by country and age in six sub-Saharan African countries. For example, some users in Kenya only shared tweets by verified Twitter accounts while students in South Africa shared news that was entertaining. Sometimes sharing misinformation depends on the content format. For instance, Singh et al. found that participants were more likely to share questionable claims on Twitter containing Uniform Resource Locators (URLs) with their friends than the same claims without URLs [66]. Often, once misinformation is shared, it is not corrected. For instance, prior works in the United Kingdom suggested that less than 20% of news sharers on social media are informed by others when they have shared dubious information [15] and on Facebook and Twitter, studies show that sometimes users ignore posts they consider misleading with no further action [26].

Other research has focused on the design of combative measures against misinformation. For instance, there have been qualitative experiments and surveys exposing users to ‘fake news’ on Facebook to see if and how they identified misleading content [22, 26]. Some studies found that lightweight interventions and frictions, such as nudging users to assess information accuracy or even preventing them from accessing known disinformation, helps users identify and avoid disinformation [32, 33]. Companies have also been employing warning labels and other strategies to combat misinformation. For example, Twitter encourages users to add their own commentary to a retweet [24], and Facebook displays a pop-up asking users if they want to share an article they have not yet opened [17]. Our study contributes to this body of knowledge by extending the study of users’ encounters and responses to misinformation to WhatsApp private chats.

2.1.1 Generational Challenges With Misinformation

There has been debate in the academic community on whether web-based misinformation can amplify inter-generational gaps. For instance, concerns have been raised around older adults’ susceptibility to misinformation due to their lack of experience with technology [48] and higher likelihood of deteriorating memory [60]. Researchers have investigated this phenomenon. Loos and Nihenhuis [40] tracked audience reach with deceptive Facebook ads linking to made-up news articles and found that the ads had higher reach amongst older age groups. Similarly, Madrid-Morales et al. [50] revealed that students and other younger users of social media in sub-Saharan Africa mostly blamed older generations for circulating fake news. Adding to this sentiment, Guess et al. [30] found older Americans more likely to share misinformation during the 2016 presidential election and Tandoc Jr. and Lee [69] found that young Singaporean adults in their 20s were more concerned for parents and older family members

about uncertainty around COVID-19 information.

Yet studies about whether age plays a part in misinformation online are mixed [54]. For example, Trninc et al. [75] concluded that both younger and older populations lack media literacy upon measuring both groups’ abilities to recognize, verify, and relate to misinformed content. Additionally, Brosius et al. [13] used survey data across 10 European countries and did not find differing levels of trust in media between generations. On the other hand, Wineburg and McGrew [84] suggest that younger generations of “digital natives” are especially at high risk of being duped by misinformation due to the amount of time spent on social media and the speed at which they consume online media. Some work even investigates younger population’s perceptions of misinformation, from feeling frustrated [11], to being under peer pressure to consume certain media [23]. Yet despite previous work, we still lack a detailed empirical understanding of how younger users interact with misinformation-related topics in intergenerational environments such as WhatsApp family chats, particularly during times of crisis such as COVID-19. Our work serves to bridge this gap.

2.2 Misinformation on WhatsApp

The study of misinformation on WhatsApp is not new. Quantitative studies have explored misinformation dissemination on WhatsApp [25, 35, 41, 45, 46, 49, 53, 57, 58]. Using publicly available data from public WhatsApp group chats, researchers have studied the effects of limiting message forwarding on misinformation’s spread on the platform [46]³, characteristics of misleading messages [57, 58], and percentages of false information in chats [35]. Studies have shown, for instance, that political and election-based misinformation is prevalent in WhatsApp group chats in Brazil [41], Indonesia [46], India [49], and Nigeria [31], among others. Researchers have typically focused on public WhatsApp group chats in their studies because these chats can be rampant misinformation spreaders and since anyone with an invitation link can join them, it makes data access for research easier. We focus on private WhatsApp chats since existing research lacks insight into misinformation encounters in private, direct messages or group chats with close friends and family. These chats can still be effective conduits for misinformation owing to forwarding on the platform [46].

In other studies of misinformation on WhatsApp, researchers have created tools for detecting misinformation and alerting users to these misleading messages. For instance, some qualitative studies examined public WhatsApp group

³WhatsApp introduced new forwarding limits in April 2020 [82]. Messages that are identified as “highly forwarded”—sent through a chain of five or more people—are marked with a double arrow icon and can only be forwarded to a single chat instead of 5. Prior to this change, in 2019, each message could be forwarded to a max of 20 chats [29], regardless of forwarding status.

chat messages [35, 41, 58] for detectable misinformation indicators such as excessively capitalized text and flashy images. In another study by Palomo and Sedano in Spain [53], they created a fact-checking tip line tool so that users could use WhatsApp as to verify claims in local news. Unlike our work, these researchers interviewed a chief editor of a local news publication rather than WhatsApp users themselves to inform design of the tool. Other researchers have developed automated misinformation detection approaches with limited success [25]. In Brazil, researchers also created WhatsApp Monitor, a tool intended to limit the spread of misinformation on WhatsApp in Brazil in public group chats [45]. However, due to WhatsApp's privacy policies and end-to-end encryption, the tool functioned as a window into the prevalence of various content categories (images, videos, audio, text) of misleading content in public WhatsApp chats for researchers rather than a direct intervention on misinformation for users. Finally, some work has looked at the efficacy of family chats in disseminating misinformation in Brazil [58] and Kenya [76].

There are a few studies of COVID-19 misinformation with WhatsApp users but not in the U.S.. Bowles et al. [12] showed from surveying WhatsApp users in Zimbabwe that information sent from trusted authorities have significant impacts on individuals' knowledge and ultimately crowd behavior. In another study of Indian WhatsApp users, Bapaye and Bapaye [8] conducted a web questionnaire survey to better understand the impact of COVID-related misinformation on WhatsApp users in India. They found that users aged over 65 years and those involved in common labor (e.g., street vendors, housekeepers) were found to be the most vulnerable to false information. The study also found that the presence of an attached link can add significant false credibility to a piece of misinformation. Finally, some work has looked at the efficacy of family chats in disseminating misinformation in Brazil [58] and Kenya [76].

While existing research has been focused on analyzing collected messages to *infer* the effect of misinformation dissemination on WhatsApp users, there have been fewer qualitative studies with WhatsApp users to understand their experiences with misinformation and no studies of misinformation encounters in private WhatsApp chats. Finally, prior studies did not investigate U.S.-based experiences with misinformation on the platform; the third most populous user base of WhatsApp users in the world [70]. Since country context affects misinformation encounters, our work serves to fill these gaps.

3 Methods

3.1 Data Collection Process

To answer our research questions, we conducted semi-structured interviews with 16 WhatsApp users who were university students in the U.S. to better understand their experiences with COVID-19 related misinformation on the platform, particularly in their private chats. Interviews were conducted

between October and November 2020 and we stopped recruiting upon reaching data saturation i.e., when we encountered repeating themes without detecting new ones from freshly enrolled participants [63]. Our study was approved by the Institutional Review Boards (IRB) of our two institutions. We designed a demographic survey and interview questions based on prior literature discussed in Section 2. For instance, since prior works had investigated the spread of misinformation in different media formats, we asked about text, image-based, and URLs as sources of misinformation. We also investigated how users perceive current measures for combating misinformation online.

Demographic Survey: Participants were asked to provide their demographic information in a Qualtrics survey prior to participating in their interview. We collected their age range, gender, highest level of education completed, estimated annual income, frequency of WhatsApp usage, and the number of years they had been using WhatsApp. Additionally, this survey was used to collect their consent to audio and video recording during the interview.

Interview Guide: We had three main categories of inquiry for our interviews to answer our research questions:

General usage: We asked questions about frequency and duration of WhatsApp usage to confirm participants' answers on the demographic survey, why they used WhatsApp over other messaging platforms, and what relationships they had with their contacts (friends, family, co-workers, etc.).

Misinformation encounters: We asked participants what concerns if any, they had about false, inaccurate, or misleading information on WhatsApp. We also asked how often they encountered this type of content and what factors they considered when deciding to trust information sent to them via WhatsApp. Specifically, we also asked if this content was text-based, an image, or a URL.

Fact checking strategies and technologies: Finally, we asked participants how they fact-checked information they received in WhatsApp. Additionally, we asked participants about current anti-misinformation tools, shown in Figure 1, such as WhatsApp's limitation on message forwarding, their magnifying glass (search) icon (WhatsApp's web-based fact checker [83]) and Health Alert partnership with the World Health Organization (WHO), along with misinformation labels being used on YouTube and Twitter in 2020 [18, 86].

We piloted our interview guide with lab members who were university students and had never been involved in this project. Based on our pilots, we made minor edits to clarify question phrasing and format. Following the pilots, we continued to the main study with the finalized interview script. Our interview questions are available in our Appendix.

Recruiting: We restricted study participation to those over the age of 18, who used WhatsApp at least multiple times a week, and were living in the U.S.. We sent recruiting notices via a university-based survey research center mailing list to undergraduate and graduate students enrolled at that institu-

Code	Explanation
General	
Chat Content	Participant talked about what they usually talked about in the chats, broadly
Foreign (non-U.S.) vs. domestic communication	Participant uses WhatsApp to communicate with people in or out of the U.S.
Relationship with others in the group (with whom they interact with most often)	Participants identified relationships with others in their group chats
Misinformation Encounters	
Most recent misinformation encounter	Participant recounts most recent misinformation counter (info content, who sent it, their reaction, etc.)
Frequency of encountering misinformation	How often does a participant encounter misinformation? (e.g., once a week, month, year, etc.)
Misinformation indicators	Participant describes factors they consider when deciding to trust (and distrust) information
Design Rec.'s & Fact-Checking Strategies	
Fact-checking strategies	Participant describes how they fact-check information (Google search, literature, consulting others, etc.)
Efficacy of current WhatsApp features that combat misinformation	Participant describes the efficacy of WhatsApp features in fact-checking and limiting the spread of misinformation
Concerns about the trade-off between combating misinformation and privacy/security	Participant raises concerns that fact-checking measures (e.g., information censorship) may undermine the privacy and comfort associated with end-to-end encryption

Table 1: A subset of our qualitative code book that is most relevant to the paper with codes and code explanations, organized by topic.

tion, by posting on class Facebook pages at both institutions, and posts on Twitter. The messages did not specifically target users who were aware of misinformation. Note that around 50% of WhatsApp users in the U.S. fall into the typical age range of undergraduate and graduate students in the U.S. [19]. After screening for our filtering criteria, participants completed a demographics survey and were scheduled for interviews. We also used snowball sampling but only recruited one additional participant using this technique. Many participants were in the same geographic region as their university but not necessarily on campus owing to pandemic lockdowns. Each interview lasted 30 minutes to 1 hour and was conducted virtually over Zoom by at least one member of the research team. We interviewed participants in English even though some participants did communicate in other languages. Examining the role of language in the spread of misinformation is beyond the scope of this paper. Note participants were not required to examine their chats during our interviews. Participants were compensated with a \$20 Amazon gift card for their time. All interviews were audio-recorded and then transcribed.

Data Analysis: We analyzed our data using deductive coding and thematic analysis [62]. We created a codebook based on our interview guide and our research questions as well as insights from team discussions about emerging points of interest while interviews were being conducted. For instance, we included codes for how participants encounter misinformation and for when they encounter different forms of misinformation such as images or URLs. Our codebook was organized into 3 broad categories, ‘General Usage’, ‘Misinformation Encounters’, and ‘Design Recommendations and Fact Checking Strategies’. A portion of the codebook is displayed in

Table 1, while the full codebook is available in the Appendix. Once we finalized the codebook by consensus in our regular weekly team discussions, each interview transcript was coded by two members of the research team with four coders overall. In total, we ended up with 33 codes and 1183 coded segments across the four coders. Once all the data was coded, we used our weekly research meetings to discuss codes of interest and each of the four coders wrote a detailed summary for a subset of codes resulting in summaries for all of our main codes. These summaries included performing a breakdown of sub-themes within the code and describing each of the sub-themes with representative participant quotes. Each team member then reviewed all the summaries in depth for our thematic analysis [62]. Since we performed coding as input to a thematic analysis, we did not calculate inter-rater reliability as this is not required [44]. However, we still built team consensus through weekly Zoom meetings to decide on the final themes emerging from the data based on the team’s reading and discussion of all the thematic summaries.

3.2 Participants

Participants’ demographics and WhatsApp usage are summarized in Table 2. Our participants had an almost even gender split with 7/16 participants identifying as male, while 9/16 identified as female. Participants were also younger overall, 14/16 were in the age range of 18-24, while 2/16 were 25-34. Participants were mainly based in the Midwestern U.S. (8/16) and Northeast (6/16) with exceptions of 2/16 based in the West and the Southeast. All participants completed at least high school. The majority (14/16) were students (undergrad-

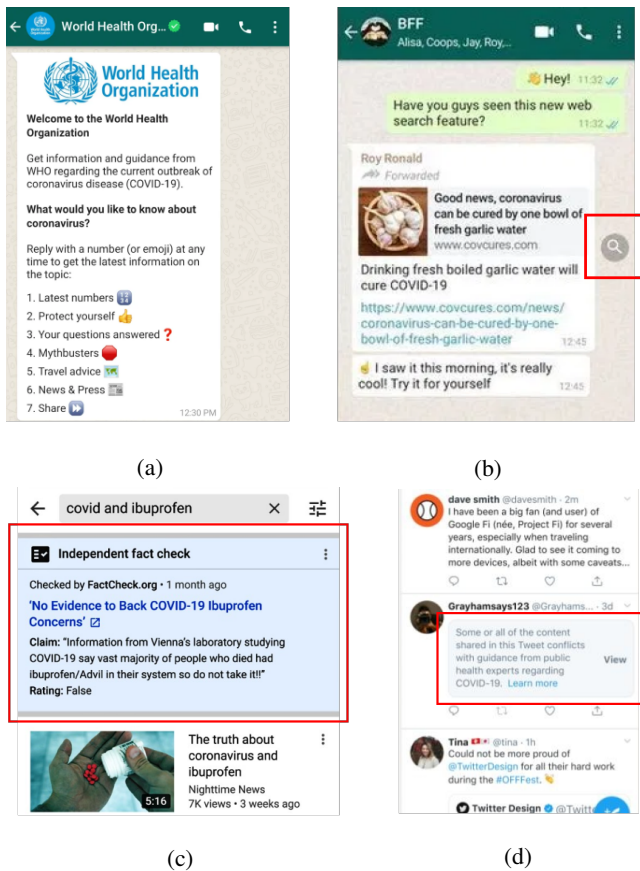


Figure 1: WhatsApp’s WHO Health Alert (a); WhatsApp’s search icon fact-checker (b); YouTube’s misinformation panel (c); and Twitter’s misinformation warning label (d).

uate or graduate) or recent graduates (2/16) including one full-time employee. Seven out of 16 reported annual incomes of <\$10,000 per year, 5/16 reported \$10,000-\$69,999, and 4/16 declined to disclose income. Participants had used WhatsApp for 1-11 years with a median of 7 years.⁴ The majority of participants self-reported that they used the app daily.

The number of contacts participants stated they had on WhatsApp varied greatly, ranging from 3 to 1015, with 20-30 being a commonly mentioned range. There was also a significant difference between the total number of contacts a user had and the number of contacts they interacted with on a regular basis. For example, P12 had 1015 total contacts on WhatsApp but was in regular contact with only about 5 of them, while Participants 11 and 15 stated that they had between 100-150 and 20-30 contacts respectively but were in touch regularly with about 20 and 10, respectively. We left the frequency term “regular” up to the definition of the participant. We also asked participants to provide us with the number of people in their chat groups (if they were comfortable doing so) and to estimate the average size of the groups they were

⁴At the time of this study, WhatsApp was more than 11 years old [81].

in otherwise. Most of the group chats were between 3 and 10 people, which were commonly mentioned sizes for private group chats consisting of family members.

4 Findings

Our analysis of the interviews yielded three main findings: how users are currently using WhatsApp (including their concerns about misinformation on the platform, how often they encountered it, and how it can spread); what misinformation indicators users look for and how they respond to misinformation on the platform; and finally, how users would like the platform to respond to misinformation.

4.1 Misinformation Perceptions And Responses

In research question one, we asked how university students currently perceive and encounter misinformation on WhatsApp. Our participants mostly used WhatsApp to communicate with others abroad, were concerned about frequently encountered misinformation on the platform, and noted that misinformation senders were often well-intentioned relatives.

4.1.1 WhatsApp Usage And Misinformation Encounters

All of our participants stated that they used WhatsApp to communicate with families and/or friends outside of the U.S. as WhatsApp was convenient to stay in touch with people abroad. This is hardly surprising as a significant number of WhatsApp users in the U.S. have non-U.S. family members [42]. Only two of our participants (P6 and P11) used WhatsApp to communicate domestically. Participants told us that they used WhatsApp primarily to share happenings in everyday life with family and friends. Interactions with family groups tended to be more regular than communications with friends.

Although participants praised the pros of WhatsApp, they also expressed concerns towards misinformation and nonsensical content circulating on WhatsApp—the main concern expressed was misleading information on COVID-19 cases and cures. For instance, at least 3/16 participants talked about how easy it is for misleading content to spread on WhatsApp since it was so easy to forward links in general. For example, P6 said that it is also “almost too easy” to select many people or groups to send a message to upon tapping the forward button, and that misinformation from families can have a layer of intimacy attached to it that makes it especially harmful:

“I know [many] have their families in WhatsApp, and people tend to trust things that come from people close to you. So, I feel like it adds almost a level of genuineness to this misinformation, and then it causes people to panic, which I think is the biggest con [of using WhatsApp].” — P6

#	Gender	Age Range	Region	Occupation	Frequency of Use (/week)	Duration of Use (years)
P1	Female	18 – 24	Midwest	Student	Daily	7
P2	Female	18 – 24	Midwest	Student	2 – 3	7
P3	Female	18 – 24	Northeast	Student	2 – 3	1
P4	Male	18 – 24	Midwest	Student	Daily	11
P5	Male	18 – 24	Midwest	Student	Daily	8
P6	Male	25 – 34	Northeast	Student	Daily	8
P7	Female	25 – 34	Midwest	Developer	Daily	8
P8	Male	18 – 24	Southeast	Student Researcher	Daily	6
P9	Female	18 – 24	Midwest	Student	Daily	3
P10	Male	18 – 24	Midwest	Student	Daily	2
P11	Male	18 – 24	Northeast	Student	Daily	8
P12	Male	18 – 24	Northeast	Student	Daily	6
P13	Female	18 – 24	Midwest	Student	4 – 6	4
P14	Female	18 – 24	Midwest	Student	2 – 3	6
P15	Female	18 – 24	Northeast	Student	2 – 3	3
P16	Female	18 – 24	West	Student	Daily	7

Table 2: Participant demographics (gender, age, region, occupation, frequency of WhatsApp use, and duration of use).

Another participant, P5, described how they have gotten so used to skeptical content on the platform that they treat it as a medium for conversation rather than relying on it for news; they also expressed the caveat that older generations trust it more. The majority of the participants (14/16) received misinformation almost every other day or multiple times a week. These participants recognized that false or misleading messages were most frequently seen in group chats possibly because “*people like to keep busy with sending messages.*” These false or misleading messages most commonly came in the form of conspiracy theories or potential cures for diseases (particularly when COVID had first entered the U.S.). For instance, P13 recalled an instance of having received a post about how “*juice made out of coriander stems and raw egg and tomato theory helps cure cancer*” in spring of 2020. The 2/16 participants who never encountered misinformation on WhatsApp attributed the lack of encounters to communicating primarily with friends (i.e., in their age range) who they know well—as opposed to family members. We also asked participants about whether or not they forwarded content to their contacts on WhatsApp to better understand how misinformation or any information may travel on the platform. Many participants (8/16) claimed to have either “*rarely*” or “*never*” forwarded any links or posts that they received on one chat to another chat. For instance, participant (P9) shared “*No, I do not because, as I mentioned, I’m guarded when I look at some of these headlines. I feel like we’re living in such a weird time.*” The 8/16 participants who did share or forward links told us that they first fact-checked the links and then sent the information only if it seemed reliable to them.

4.1.2 Misinformation Senders

We asked participants about who or what entity was sending them misinformation on WhatsApp. The 14/16 participants who had a high frequency of encountering misinformation (approximately every other day or multiple times a week), revealed that the senders were typically close family members. These family members sent (mis)information in a range of formats (from “*copy pastas*”—long, often joking texts distributed through copy and paste—to texts, images and links). Our participants felt that this information ultimately did not harm them because they were either cognizant of these groundless claims or the information itself did not pose a severe threat to anyone who believed it. In the words of P3:

“The sender for me was just my mom, and I did speak to her about it, and she was definitely of a different mindset. She was more of the mindset that we should do whatever we can even if it’s not true, even if it’s just helping your immune system at this point, we’ll do anything. So, I wouldn’t say she necessarily believed that it makes you immune to COVID, or protects you or anything, but she also didn’t consider it misinformation. She was like “As long as it’s helping everyone.” She also sent it to people. . . I mean, it’s up to you to do whatever you want with it.” - P3

Participants also expressed that these family members were often sending messages without malicious intent of sharing information that could prove dangerous. Another participant (P10), reflecting this sentiment, perceived that:

“[her mom and aunts] find it very easy to essentially forward a message from another group chat

to another, essentially spamming the group chat with all sorts of massive, long text messages about something, or a web link that is pretty much misinformation.” - P10

Contrary to having malicious intent, our participants also described how, oftentimes, their family members sent misinformation with the intention of keeping others safe and informed in the midst of a pandemic. For example, P10 also described how half of her family believed *“that we should rinse our noses with saline solution to prevent COVID”* and when asked if she followed this protocol, she would merely respond by saying yes so as to avoid getting into a lengthy argument of whether and why this approach to combating the virus is ineffective.

4.2 Misinformation Indicators and Responses

In our second research question, we asked how users identify whether content is misinformation on the platform and how they respond to misleading content. Participants told us they had four main indicators that a message was misinformation and had developed strategies for fact-checking content. In response to misinformation, not everyone was comfortable with confronting senders, often owing to family dynamics.

4.2.1 Indicators Of Misleading Content

Generally, participants told us about four main indicators that they relied on to decide whether to trust information sent to them via WhatsApp: 1) the credibility of the information source, 2) their relationship with the misinformation sender, 3) the format and framing of the message, and 4) personal politics and values. Many of these strategies, aside from relationship with the sender, echo indicators developed by Jahanbakhsh et al. [32] on reasons people believe or disbelieve claims, as well as textual misinformation indicators for automated detection specified by Resende et al. [57]. These strategies also echo findings on studies of other social media platform users such as Facebook [22, 26, 47], i.e., using the source of a news article to evaluate its credibility.

Source Credibility and Name Recognition. The majority of participants paid attention to the source’s credibility when deciding to trust information sent to them (15/16). Participants focused on the reputability of the organization when analyzing information, most often news media content. Established media and news corporations carried greater credibility and legitimacy compared to smaller, more obscure media outlets; e.g., participants mentioned The New York Times and MSNBC. Participants generally expected the source to be linked to an established news platform as opposed to a random individual’s social media account. Additionally, participants considered government organizations and links that forwarded to .org and .gov, e.g., www.cdc.gov, as reliable.

Relationship with Sender. Complementary to Geeng et al.’s finding that Facebook and Twitter users may trust certain poster’s content because they trust the individual [26], we found that the opposite can be true as well; participants may inherently mistrust content because they have deemed the sender to be unreliable and untrustworthy.

Since participants primarily used WhatsApp to communicate with friends and family, they told us they measured the trustworthiness of information based on their relationship and perception of the sender. If a sender was known to consistently share misleading information, participants were more likely to be skeptical of them. This theme was most prevalent when participants described their relationship with older relatives; 9/16 expressed concern that their older contacts were unable to distinguish between credible and untrustworthy news content and were less prone to fact-checking before sharing on WhatsApp. Over time, P2 felt increasingly suspicious when receiving messages from their grandparents and older relatives in large family group chats:

“Just because they are not as able to filter out fake news from real news. I mean, obviously it’s presented in a more and more realistic way every single day and they just lap it up and believe in it, and also, they are not as tech savvy to be able to go and Google immediately and do a quick check on what’s actually happening” — P2

Participants described how these contacts would frequently spam family group chats with information they received in other group chats and channels. Five out of 16 participants described ignoring messages from particular senders since they automatically assumed false or misleading content. However, there were a few exceptions where participants trusted their contacts when sharing information on unfamiliar topics. For example, in the midst of school and university closings in response to the early COVID-19 outbreak, P15, a graduate student, said she was bombarded with news stories that contradicted each other. This participant reached out to her sister who told her to expect her school to cancel all in-person activities. Because P15 had a close relationship with her sister, she trusted her sources.

Format and Framing. Six of the 16 participants reported distrusting and avoiding messages that: urged users to spam forwards, shared without context, were overly sensational and attention-seeking, had inflammatory language, and were opinion-based. Three out of 16 participants expressed mistrust of forwarded messages because these messages often followed a template that explicitly asked users to forward the message to their contacts. Further, participants believed if someone did not dedicate time to writing their own messages, they probably did not verify it either. Participants also took the visual layout and format of a message into account as well; two participants avoided messages that displayed excessive use of colors, advertisements, capitalized and bold

texts, emoticons, and other eye-catching designs apart from the text itself. Participants also told us they were wary of poorly spliced pictures that may have been edited beforehand or messages framed with inflammatory, opinionated content that were seen as biased and misleading (2/16). In the case of COVID-19 news, these participants trusted sources that presented numerical data (e.g., number of cases, growth rate) in a neutral tone without underlying agendas.

Political ideology. A few participants (4/16) expressed political ideology as an important factor when deciding to trust information. They said they were less likely to trust content, as credible as it may be, from news organizations or their personal contacts with conspicuous political views out of concern of an underlying political agenda. For instance, P9 expressed having conservative political values and criticized left-leaning news sources sent from contacts with opposing political ideologies because they automatically considered them biased and misleading. Likewise, P11, a self-described liberal, disregarded any news articles sent from conservative family members.

4.2.2 Fact-Checking Using Google And Intuition.

Thirteen out of 16 participants were asked about fact-checking strategies, and two main approaches were found as participants' primary fact-checking approaches: 1) searching on Google and 2) relying on personal judgment. Apart from these, reading scientific papers was mentioned once by a graduate student (P11) and directly asking other contacts such as friends by one other participant (P6). It is worth noting that, in reality, these strategies are not mutually exclusive and are often employed together by an individual in a single fact-checking attempt.

Google. 12/14 participants told us their most common way to fact-check information sent to them on WhatsApp was to search on Google to verify its accuracy. When a source's reliability was unknown, P15 stated they usually "*click on the links, maybe read some other articles that have been published by the same website or author and see if those are accurate*". If participants found multiple sources corroborating each other, they felt this was an extra piece of evidence that the information was accurate, therefore trustworthy. Participants told us that their process of verifying the information with other sources, especially those considered authoritative, was not exclusive to Google. They checked the information from any source that they usually consulted for information and trusted.

Prior Knowledge. Eight out of 16 participants relied on their intuition, prior knowledge, and understanding of current affairs to determine whether or not a message, image, text, or URL was intentionally misleading or false. This finding echoes that of Flintham et al. [22], for Facebook users who looked for 'fake news' in an experiment on fake news articles only and sometimes relied on their own judgement for determining veracity. In our study, which occurred in the first year

of the COVID-19 pandemic, most participants expressed prior knowledge of COVID-19 cases, precautions, and myths that informed them outside of their WhatsApp channels. For example, myths about COVID-19, such as gargling warm salt water or drinking lemon juice twice a day, sounded completely outlandish to some participants given their understanding of the properties of the virus and the vaccine. In another related example, P10 described a misinformation encounter where their aunt claimed eating ice cream and other cold foods increased the chances of contracting the coronavirus:

"If I had to think about basic biology, it's pretty hard to link ice cream to a virus that caused a global pandemic, I would say. I'd say, yes, maybe if you eat ice cream a lot and don't dress up in cold months, your immune system may be more vulnerable to the flu, to the virus. But it wouldn't be a direct cause of COVID" — P10

4.2.3 Dealing With Misinformation Senders

Out of 15 participants who allegedly encountered misinformation via WhatsApp, 9 people mentioned past experiences of confronting senders of misleading information, 8 people mentioned scenarios where they were passive and didn't challenge the senders—even when they recognized there were something incorrect with the content shared, and 2 others confessed they didn't always stick to one strategy.

Actively Confronting Misinformation Senders. When encountering misinformation, "active" participants confronted the sender, especially if they were on close terms with them. However, most of them recognized that "*there is no point*" in repeatedly resisting and reminding the sender to check the sources of any information they forward, prior to sharing, especially when the sender continues not to do so. In one canonical example, P3 actively confronted their mother by asking a question along the lines of "*Do you also believe this? Do you think it's believable?*" The participant also explained that they were able to confront the sender (in this case their mother) since the participant was a) close with the sender and b) they knew that the sender had no malicious interest in sending incorrect information. Other "active" participants, who fact-checked a topic by doing further research, shared that whenever they received any information that they had not yet encountered, they ventured to ask the sender questions like "*where did you find this?*". In one example, P1's mother sent her sensational and misleading information on COVID cases in the U.S.. Although P1 personally thought that the U.S. could do better in curtailing the virus, she recognized that her mother's sources made the problem worse than it was. Recognizing that she was simply worried and did not purposely share misinformation, P1 confronted her mother to comfort her:

"Yes, we did talk about this quite often during the

video chatting. I would just try to assure her, “Oh, Mom. This is okay,” and regardless how the numbers surge in America, like myself, at least I can protect myself. I just wear masks and I do hand sanitizing very often, so I’m trying to point out to her, “Mom, this is misinformation. America is actually doing fine.” Well, it’s not. So, yeah, I don’t counter the source directly, but I am trying to comfort her on speaking for my personal level.” – P1

Passively Ignoring Misinformation Senders. While these “active” participants did not let these qualms prevent their confronting of senders, “passive” participants acknowledged that they would simply ignore anything shared via WhatsApp based on the contents and sender of the post (e.g., if the content concerned the 2020 Black Lives Matter protests or COVID-19). At least 2 out of the 6 passive participants expressed explicitly that they did not want to upset any family relations due to a “trivial” post shared on social media. Other participants echoed this sentiment and told us they often reacted passively about misinformation, not taking the time to correct others’ misaligned opinions or views as it would lead to an “hour long argument” which the participants did not want to face. In another anecdote, P2 recalled having received information from her family members regarding unfounded steps of precaution to take against COVID involving gargling with “warm saltwater every time” they came back into their home from being outside to “kill off all COVID particles and be safe.” This participant did not correct their family members as they did not want to cause any unfriendliness for a harmless piece of information:

“I’m not interested in trying to correct people because it’s just not going to work, they’re going to believe what they want to believe. I had a phase a couple of years ago where I was trying to correct people and I was like, it’s not going to happen, it’s not going to work. So now I’m just like, ‘Sure, you do you and I’m just going to ignore.’” – P2

In another representative example, P8, reported that it was easier to delete group chats which they had flagged as one of main mediums of misinformation without reading any content sent. P8 accepted that “There was just a point where there was so much going around it was easier to just, honestly, stop reading things.” To summarize, participants often did not want to strain family relationships by correcting misinformation, especially given that, in many cases, they perceived the misinformation to be harmless.

4.3 Views on Existing Mechanisms To Combat Misinformation on WhatsApp

To answer research question three, we asked how aware and confident participants were of current features to combat mis-

information on WhatsApp and their opinions on how to improve how the platform handles misinformation, particularly around COVID-19 as shown in Fig. 1. In general, participants showed little to no awareness towards the features probed and expressed varying opinions on efficacy of these features and concerns around the privacy dilemma of combating misinformation in the context of end-to-end encryption.

Of all the existing features shown or discussed with all participants (WhatsApp forwarding limits, WhatsApp search icon, and the WHO health alert), on average only about 4 participants had heard of at least one or more of these features. Generally, participants mentioned that the forwarding limit could be circumvented if a sender manually copied and pasted it or by sending the message one at a time or via another platform. Participants also thought the search icon could link to multiple search engines rather than one and felt the WHO alert did not look professional owing to the use of emojis.

4.3.1 Privacy and Security Concerns

Not only were participants unaware of existing anti-misinformation measures, they also voiced concerns on whether or not WhatsApp should even be responsible for designing preventative measures against misinformation.

Content Moderation Concerns. At least 6/16 participants believed that WhatsApp, as a platform, should not be accountable for curbing any misinformation, arguing that it is up to the user’s discretion whether or not they believe what they see. Even if the content is explicitly false, they felt that users are entitled to share anything they want and believe to be true. On the other hand, participants agreed that WhatsApp definitely bears a responsibility in fact-checking and regulating any misleading content, rather than burdening the user to determine what is trustworthy.

Other participants expressed major concerns about the trade-off between users’ privacy and WhatsApp’s efficacy against misinformation (3/16). They felt these features infringed upon users’ privacy and therefore preferred if WhatsApp did not explicitly flag or censor misinformation. Should WhatsApp ever flag or censor direct messages, it would need to clarify any privacy-preserving techniques and the methods used to identify any inflammatory or misleading content.

Misinformation Warnings And Labels. When asked to suggest design recommendations to limit the spread of misinformation, only 5/16 participants thought that WhatsApp should adopt the misinformation warning labels similar to YouTube’s and Twitter’s warnings [18, 86]. They liked the idea of warning users not to trust certain sources while still giving them the option to share. As P13 said, “they should be allowed to view it because of free speech, but they should be aware that it is incorrect, it’s misinformation.” An alternative suggestion was for WhatsApp to record known misinformation sources such as websites (4/16) or to generate a credibility rating for websites for when senders share links (2/16).

5 Discussion and Design Suggestions

Our study suggests that WhatsApp is uniquely situated in the misinformation space based on the following three key findings:

- **F1:** Our participants' group-based WhatsApp communications with close family and friends make it especially effective in disseminating misinformation out of good intention. Previous studies observed the efficacy of WhatsApp as a misinformation pipeline in large public chats [31,41,46,49], but our study suggests this may also be the case in private chats. Future studies are needed to confirm if it is mainly older adults spreading content.
- **F2:** The peer-to-peer nature of communication on WhatsApp adds intimacy and complicates users' ability and/or willingness to deal with misinformation they encounter. Because we focused on gathering deep user experiences in private chats over collecting data using automated methods as in prior studies [25,45,57], we were able to surface significant social power dynamics within chats that pose challenges to countering misinformation.
- **F3:** Participants were unaware of current mechanisms on WhatsApp to combat misinformation. Moreover, privacy and information accuracy, both desirable in communication apps, can be seen as conflicting traits on WhatsApp. Such a tradeoff has been a common technical assumption known to experts in the field [43], but our study revealed that everyday users are also well aware of this trade-off.

We think it is particularly important to engage with **F3** when addressing misinformation in end-to-end encrypted environments. While some participants told us they would appreciate more effort on WhatsApp's part to flag misinformation, they also acknowledged that WhatsApp's inability to read messages will hinder its ability to do so. However, no participant mentioned that encryption should be sacrificed to offer more robust fact-checking services, implying that they still hold privacy on the platform in high regard. This tension offers rich avenues for future work.

In addition to privacy, dealing with misinformation in private chats is complicated by social relations. We found that the more personal nature of communication on WhatsApp integrated social dynamics that discouraged a user from actively confronting misinformation senders. Our observed social dynamics include cultural emphases on respect and deference to elders: many of our participants feared correcting older family members' misinformation out of concern for coming across as rude or disrespectful, despite having a justifiable and legitimate reason. Therefore, younger users, who our participants claim to be more adept at identifying misinformation, may not be able to signal the misleading nature of a piece of information to others if it is sent by older family members or relatives. Further, many participants recognized that misinformation

often resulted from well-intentioned family members who sent it out of care for others (e.g., bogus COVID-19 cures), supporting preliminary research suggesting that information dissemination on WhatsApp follow familial, communal, and ideological ties [7]. This is worthy of further study in the U.S. as it may be of particular relevance to a rising body of work around digital communication and misinformation within American immigrant diaspora communities [68,78].

These findings point to a need for alternate approaches to combating misinformation in end-to-end encrypted, private group chats, as conventional moderation techniques often rely on examining content and do not take into consideration sociocultural dynamics between group chat members. For example, educational campaigns around misinformation may include tips and suggestions for dealing with relatives but ground this in terms of caring about others.

5.1 Design Suggestions

Our participants were for the most part unaware of anti-misinformation features on WhatsApp, suggesting that even when a platform is actively trying to combat misleading content, users may not know about these measures. Assuming a platform can overcome the hurdle of raising user awareness of new anti-misinformation features, based on the insights above, we propose the following design approaches to improve the ways users can deal with misinformation on end-to-end encrypted platforms. These features may be useful to users within our study demographic, but generalizations to a broader user base cannot be made without additional studies.

5.1.1 Empowering the user to better fact-check or flag misinformation for themselves.

WhatsApp cannot analyze content to identify misinformation due to the platform's encryption policies. Another platform-controlled measure, forwarding limits, has been seen as ineffective by participants in our study as well as previous work [46]. Based on our findings, we suggest designing to *empower the user* with tools to combat misinformation. For misinformation senders, we suggest reminding users of the value of fact-checking before forwarding content. For misinformation receivers, designs should: 1) respect the user's ability to classify misinformation for themselves, and 2) make it easier for the user to organize and track their misinformation encounters so they can later fact-check and better learn from them. This can be translated into features for both the information sender and receiver.

- **Sender:** By adding friction using a popup dialogue box that asks the user whether they have fully read the contents of a link, users can be prompted to reflect on information they are sharing before forwarding content. This kind of friction is already being deployed by other platforms to reduce sharing without context [24] and

is shown to be effective in obstructing access to disinformation [33]. However, the friction should not be too high, as it can then be seen as censorship, [59].

- **Receiver:** An option to mark a message as dubious and decrease its visibility in their chat screen may help users mitigate the sight of misleading content. This can protect the user as previous work in psychology indicate that repetition of a message can increase believability in it despite one’s initial judgements [20, 21, 77] from believing deal with the constant flow of misinformation. Note that this feature is distinct from WhatsApp’s current option to delete a message, which can result in disparate versions of the same chat across different users. [79].
- **Receiver:** To help users track and fact-check messages, users may store messages that have been flagged as dubious in a “quarantine” bin for later inspection. The bin can be equipped with tools to help users surface trends, such as common language or links, across dubious messages. Users can then use these trends better identify misinformation in future messages.

5.1.2 Helping users deal with misinformation in ways that mitigate power dynamics in groups.

Our findings suggest social dynamics in family group chats can make it difficult for users to confront and correct misinformation senders. We propose the following features to allow users to subtly alert others about potential misinformation.

- **Selectively applying the fact checker icon to messages:** We can let users anonymously apply WhatsApp’s fact checker⁵ to particular messages for everyone in the chat to see and use. This offers resources to group members without accusing anyone of sending misinformation.
- **Anonymous suggestions of alternative resources:** One suggestion is to allow users to anonymously suggest a link to an alternative information resource to the sender. Once the resource is suggested, the sender can receive a notification with the anonymous suggestion and choose whether to accept it. If accepted, the link can be sent into the group as a reply to the original message to update others and gently nudge the group towards discussion.

6 Limitations and Future Work

Our study sample was limited to 16 university students and recent graduates who were mostly in a younger age bracket of 18-35 years. By its nature, our qualitative study is not intended to be generalizable [62, 63]. Future work could expand

⁵WhatsApp has already rolled out to some users its own web-based fact-checker [83]. However, since the platform cannot read message contents, it applies the fact checker to all links, which may not always be desirable.

our study to a broader sample of young users who are not students or to a larger sample of more age-diverse U.S. based participants across the country. Also, while we asked participants about misinformation around topics such as Black Lives Matter protests and U.S. elections, we did not collect sufficient data to report on it. Future work could thus investigate topics beyond COVID-19. Additionally, even though our participants were based in the U.S., we observed that most communication on the app was international. Studies that specifically investigate misinformation within domestic interactions on WhatsApp may also complement our work since the language of communication may affect the perceptions of misinformation. Studying WhatsApp users in other countries would also expand on our study. Finally, future studies could implement and test our design recommendations or study other end-to-end encrypted chat-based platforms, such as Telegram [73], Signal [65], and iMessage [2].

7 Conclusions

We interviewed 16 U.S.-based university students and a recent graduate about their experiences with misinformation related to COVID-19 in private WhatsApp group chats. We were interested in filling in two gaps in previous literature: the lack of qualitative user interviews to understand younger adults’ misinformation experiences on end-to-end encrypted messaging platforms such as WhatsApp, and the lack of studies on how WhatsApp is used in the U.S. Our findings suggest that there is a need to differentiate the nature of misinformation on WhatsApp compared to other popular American social media apps such as Twitter and Facebook. Namely, WhatsApp’s popularity as an international communication tool used with close family or friends can unknowingly turn good intentions into misinformation-sharing frenzies and hinder the ability of those who identify misinformation to notify others about it. Additionally, WhatsApp’s staunch commitment to end-to-end encryption can present limitations to the techniques the platform is able to deploy to combat misinformation. Our findings offer implications for design approaches to both mitigate the sharing of misinformation and improve experiences of users who receive misinformation. These findings and suggestions may help WhatsApp users outside the U.S.—and even users on similar platforms—handle similar issues and spark new discussions around information moderation with privacy-preserving techniques more broadly.

Acknowledgments

We thank our participants. This work was partially supported by the Princeton Council for Science and Technology and a Facebook ‘Secure The Internet’ award.

References

- [1] Zara Abrams. Controlling the spread of misinformation. *American Psychological Association*, 52:44, 03 2021.
- [2] Apple Inc. Use imessage apps on your iphone, ipad, and ipod touch. <https://support.apple.com/en-us/HT206906>. Accessed: 2021-07-09.
- [3] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. Acting the part: Examining information operations within #blacklivesmatter discourse. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.
- [4] Brooke Auxier and Monica Anderson. Social media use in 2021. <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/>, 04 2021. Accessed: 2021-06-28.
- [5] Daniel Avelar. Whatsapp fake news during brazil election ‘favoured bolsonaro’. <https://bit.ly/3tqVz61>, 10 2019. Accessed: 2021-10-21.
- [6] Ahmed Balami and Hadiza Umar Meleh. Misinformation on salt water use among nigerians during 2014 ebola outbreak and the role of social media. *Asian Pacific Journal of Tropical Medicine*, 12:175, 01 2019.
- [7] Shakuntala Banaji, Ram Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin. WhatsApp Vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. page 62.
- [8] Jay Amol Bapaye and Harsh Amol Bapaye. Demographic factors influencing the impact of coronavirus-related misinformation on whatsapp: Cross-sectional questionnaire study. *JMIR Public Health Surveill*, 7(1):e19858, 01 2021.
- [9] Zapan Barua, Sajib Barua, Najma Kabir, and Mingze Li. Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8:100119, 07 2020.
- [10] Shashank Bengali. How whatsapp is battling misinformation in india, where ‘fake news is part of our culture’. <https://www.latimes.com/world/la-fg-india-whatsapp-2019-story.html>, 02 2019. Accessed: 2021-10-21.
- [11] Porismita Borah, Bimbisar Irom, and Ying Chia Hsu. ‘it infuriates me’: examining young adults’ reactions to and recommendations to fight misinformation about covid-19. *Journal of Youth Studies*, pages 1–21, 2021.
- [12] Jeremy Bowles, Horacio Larreguy, and Shelley Liu. Countering misinformation via whatsapp: Preliminary evidence from the covid-19 pandemic in zimbabwe. *PLOS ONE*, 15:e0240005, 10 2020.
- [13] Anna Brosius, Jakob Ohme, and Claes H de Vreese. Generational gaps in media trust and its antecedents in europe. *The International Journal of Press/Politics*, page 19401612211039440, 2021.
- [14] Carole Cadwalladr. The great British Brexit robbery: how our democracy was hijacked. <https://bit.ly/3MCpdvE>, 2017. Accessed: 2021-06-08.
- [15] A. Chadwick and Cristian Vaccari. News sharing on uk social media: misinformation, disinformation, and correction. 2019.
- [16] Adélie Chevée. Mutual aid in north london during the covid-19 pandemic. *Social Movement Studies*, pages 1–7, 2021.
- [17] Mitchell Clark. Facebook wants to make sure you’ve read the article you’re about to share. <https://www.theverge.com/2021/5/10/22429174/facebook-article-popup-read-misinformation>, 2021. Accessed: 2021-06-07.
- [18] COVID-19 misleading information policy. Covid-19 medical misinformation policy. <https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy>. Accessed: 2021-07-02.
- [19] Statista Research Department. Whatsapp usage penetration in the united states 2020, by age group. <https://www.statista.com/statistics/814649/whatsapp-users-in-the-united-states-by-age/>, 10 2021. Accessed: 2021-10-26.
- [20] Lisa Fazio, Nadia Brashier, B Payne, and Elizabeth Marsh. Knowledge does not protect against illusory truth. *Journal of experimental psychology. General*, 144, 08 2015.
- [21] Lisa Fazio and Gordon Pennycook. Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, 26, 08 2019.
- [22] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. *Falling for Fake News: Investigating the Consumption of News via Social Media*, page 1–10. Association for Computing Machinery, New York, NY, USA, 2018.
- [23] Fiona Gabbert, Amina Memon, Kevin Allan, and Daniel B Wright. Say it to my face: Examining the

effects of socially encountered misinformation. *Legal and Criminological Psychology*, 9(2):215–227, 2004.

- [24] Vijaya Gadde and Kayvon Beykpour. Additional steps we’re taking ahead of the 2020 us election. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html, 2021. Accessed: 2021-06-07.
- [25] Kiran Garimella and Dean Eckles. Whatsapp and nigeria’s 2019 elections: Mobilising the people, protecting the vote. *Harvard Kennedy School (HKS) Misinformation Review*, 07 2020.
- [26] Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don’t) investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.
- [27] Amira Ghenai and Yelena Mejova. Fake cures: User-centric modeling of health misinformation in social media. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.
- [28] Nathaniel Gleicher. Removing Coordinated Inauthentic Behavior and Spam From India and Pakistan. <https://about.fb.com/news/2019/04/cib-and-spam-from-india-pakistan/>, 2019. Accessed: 2021-06-06.
- [29] Rachel Greenspan. Whatsapp fights fake news with message forwarding restrictions. <https://time.com/5508630/whatsapp-message-restrictions/>, 01 2019. Accessed: 2021-07=07.
- [30] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019.
- [31] Jamie Hitchen, Jonathan Fisher, Nic Cheeseman, and Idayat Hassan. Whatsapp and nigeria’s 2019 elections: Mobilising the people, protecting the vote. 07 2019.
- [32] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.
- [33] Ben Kaiser, Jerry Wei, Elena Lucherini, Kevin Lee, J Nathan Matias, and Jonathan Mayer. Adapting security warnings to counter online disinformation. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021.
- [34] Masato Kajimoto, Yenni Kwok, Yvonne Chua, and Ma Labiste. Information disorder in asia and the pacific: Overview of misinformation ecosystem in australia, india, indonesia, japan, the philippines, singapore, south korea, taiwan, and vietnam. *SSRN Electronic Journal*, 03 2018.
- [35] Khalid Khaja, Alwaleed Alkhaja, and Reginald Sequeira. Drug information, misinformation, and disinformation on social media: a content analysis study. *Journal of Public Health Policy*, 39, 08 2018.
- [36] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie Akl, and Khalil Baddour. Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12, 03 2020.
- [37] David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. Combating fake news: An agenda for research and action. <https://shorensteincenter.org/combating-fake-news-agenda-for-research/>, 05 2017. Accessed: 2021-06-22.
- [38] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
- [39] Stephan Lewandowsky, Ullrich Ecker, Colleen Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13:106–131, 12 2012.
- [40] Eugène Loos and Jordy Nijenhuis. Consuming fake news: A matter of age? the perception of political fake news stories in facebook ads. In *International Conference on Human-Computer Interaction*, pages 69–88. Springer, 2020.
- [41] Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. In *Companion Proceedings of The 2019 World Wide Web Conference, WWW ’19*, page 1013–1019, New York, NY, USA, 2019. Association for Computing Machinery.
- [42] Farhad Manjoo. For millions of immigrants, a common language: Whatsapp. <https://nyti.ms/39fwjZv>, 12 2016. Accessed: 2022-04-24.

- [43] Jonathan Mayer. Content moderation for end-to-end encrypted messaging. https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf, 10 2019. Accessed: 2021-10-22.
- [44] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.
- [45] Philippe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. Whatsapp monitor: A fact-checking system for whatsapp. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):676–677, 07 2019.
- [46] Philippe Melo, Carolina Vieira, Kiran Garimella, Pedro Vaz de Melo, and Fabrício Benevenuto. *Can WhatsApp Counter Misinformation by Limiting Message Forwarding?*, pages 372–384. 01 2020.
- [47] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. Social and heuristic approaches to credibility evaluation online. *Journal of communication*, 60(3):413–439, 2010.
- [48] Ryan C Moore and Jeffrey T Hancock. Older adults, social technologies, and the coronavirus pandemic: Challenges, strengths, and strategies for support. *Social Media+ Society*, 6(3):2056305120948162, 2020.
- [49] Vidya Narayanan, Bence Kollanyi, Ruchi Hajela, Ankita Barthwal, Nahema Marchal, and Philip N. Howard. News and information over facebook and whatsapp during the indian election campaign. *Project on Computational Propaganda*, 02 2019.
- [50] Khulekani Ndlovu, Dani Madrid-Morales, Herman Wasserman, Melissa Tully, and Emeka Umejei. Motivations for sharing misinformation: A comparative study in six sub-saharan african countries. *International Journal of Communication*, 15:1200–1219, 02 2021.
- [51] Office of the Director of National Intelligence. Assessing Russian activities and intentions in recent US elections. National Intelligence Council. https://www.dni.gov/files/documents/ICA_2017_01.pdf, 2017. Accessed: 2021-06-08.
- [52] Jonathan Corpus Ong and Jason Vincent A Cabañes. Architects of Networked Disinformation. The Newton Tech4Dev Network. <https://bit.ly/3aIvoRu>, 2018. Accessed: 2021-06-08.
- [53] Bella Palomo and Jon Sedano. Whatsapp as a verification tool for fake news. the case of ‘b de bulo’. *Revista Latina de Comunicacion Social*, 73:1384, 11 2018.
- [54] Sora Park, Caroline Fisher, Jee Young Lee, and Kieran McGuinness. Covid-19: Australian news and misinformation. 2020.
- [55] Sarah Perez. Report: Whatsapp has seen a 40% increase in usage due to covid-19 pandemic. <https://tinyurl.com/bdcw29ct>, 03 2020. Accessed: 2021-06-07.
- [56] Kunal Purohit. Misinformation, fake news spark india coronavirus fears. <https://tinyurl.com/yde9n8sj>, 03 2020. Accessed: 2021-10-21.
- [57] Gustavo Resende, Philippe Melo, Julio C. S. Reis, Marisa Vasconcelos, Jussara M. Almeida, and Fabrício Benevenuto. Analyzing textual (mis)information shared in whatsapp groups. In *Proceedings of the 10th ACM Conference on Web Science, WebSci '19*, page 225–234, New York, NY, USA, 2019. Association for Computing Machinery.
- [58] Gustavo Resende, Philippe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference, WWW '19*, page 818–828, New York, NY, USA, 2019. Association for Computing Machinery.
- [59] Margaret Roberts and Margaret E Roberts. *Censored*. Princeton University Press, 2018.
- [60] Henry L Roediger III and Lisa Geraci. Aging and the misinformation effect: A neuropsychological analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):321, 2007.
- [61] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. Susceptibility to misinformation about covid-19 around the world. *Royal Society Open Science*, 7(10):201199, 2020.
- [62] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE, Los Angeles, 2nd ed edition, 2013.
- [63] Irving Seidman. *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. Teachers College Press, 2013.
- [64] Michael Seufert, Tobias Hofffeld, Anika Schwind, Valentin Burger, and Phuoc Tran-Gia. Group-based communication in whatsapp. pages 536–541, 2016.

- [65] Signal. Speak freely. <https://signal.org/>. Accessed: 2021-07-09.
- [66] Lisa Singh, Leticia Bode, Ceren Budak, Kornraphop Kawintiranon, Colton Padden, and Emily Vraga. Understanding high- and low-quality url sharing on covid-19 twitter streams. *Journal of Computational Social Science*, 3:1–24, 11 2020.
- [67] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.
- [68] Wanning Sun. Chinese diaspora and social media: Negotiating transnational space. In *Oxford Research Encyclopedia of Communication*. 2021.
- [69] Edson C Tandoc Jr and James Chong Boi Lee. When viruses and misinformation spread: How young singaporeans navigated uncertainty in the early stages of the covid-19 outbreak. *New Media & Society*, page 1461444820968212, 2020.
- [70] H. Tankovska. Countries with the most whatsapp users 2019. <https://www.statista.com/statistics/289778/countries-with-the-most-facebook-users/>, 01 2019. Accessed: 2021-06-23.
- [71] H. Tankovska. Most popular global mobile messenger apps as of january 2021, based on number of monthly active users. <https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/>, 02 2021. Accessed: 2021-06-23.
- [72] Mazumder Hoimonty Tasnim Samia, Hossain Md Mahubub. Impact of rumors and misinformation on covid-19 in social media. *J Prev Med Public Health*, 53(3):171–174, 2020.
- [73] Telegram. Telegram. a new era of messaging. <https://telegram.org/>. Accessed: 2021-07-09.
- [74] Mayowa Tijani. How to spot covid-19 misinformation on whatsapp. <https://factcheck.afp.com/how-spot-covid-19-misinformation-whatsapp>, 04 2020. Accessed: 2021-10-21.
- [75] Dragana Trninić, Anđela Kuprešanin Vukelić, and Jovana Bokan. Perception of “fake news” and potentially manipulative content in digital media—a generational approach. *Societies*, 12(1):3, 2022.
- [76] Melissa Tully. Everyday news use and misinformation in kenya. *Digital Journalism*, pages 1–19, 2021.
- [77] Christian Unkelbach and Rainer Greifeneder. Experiential fluency and declarative advice jointly inform judgments of truth. *Journal of Experimental Social Psychology*, 79:78–86, 2018.
- [78] Ben Gia Minh Vo. Vietnamese america: On ‘good refugees’, fake news, and historical amnesia. *Asian American Research Journal*, 1(1), 2021.
- [79] WhatsApp Help Center. How to delete messages. <https://faq.whatsapp.com/android/chats/how-to-delete-messages/?lang=en>. Accessed: 2021-10-29.
- [80] WhatsApp Help Center. About end-to-end encryption. <https://faq.whatsapp.com/general/security-and-privacy/end-to-end-encryption/?lang=en>, 2021. Accessed: 2021-06-08.
- [81] WhatsApp LLC. Whatsapp 2.0 is submitted. <https://blog.whatsapp.com/whatsapp-2-0-is-submitted>, 2009. Accessed: 2021-03-14.
- [82] WhatsApp LLC. Keeping whatsapp personal and private. <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private>, 04 2020. Accessed: 2021-06-23.
- [83] WhatsApp LLC. Search the web. <https://blog.whatsapp.com/search-the-web>, 08 2020. Accessed: 2020-09-20.
- [84] Sam Wineburg and Sarah McGrew. Evaluating information: The cornerstone of civic online reasoning. 2016.
- [85] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90, November 2019.
- [86] YouTube Help. Covid-19 medical misinformation policy. <https://support.google.com/youtube/answer/9891785?hl=en>, 05 2020. Accessed: 2021-07-02.

Appendix A: Interview Questions

General WhatsApp usage

- Why do you use WhatsApp? (vs. other social media or messaging apps like iMessage, Facebook Messenger, etc.)
- Is WhatsApp your primary communication app?
- How often do you use WhatsApp?
- How long have you had WhatsApp?
- What do you think are the pros and cons of WhatsApp?
- How many contacts do you have on WhatsApp?
- What relationship do you have with your contacts? Are they friends? Family? Work colleagues? Acquaintances? Others?
- What do you usually talk about on WhatsApp? Do you share links when you talk?
- Are most of your conversations on WhatsApp direct messages or group chats?
 - Can you give a ballpark percentage of the conversations that happen in private messages vs. in group chats?
 - How large are your group chats? Who are in them?
- Do you know anything about WhatsApp's end-to-end encryption?

Encounters of doubtful information

- What concerns do you have about false, inaccurate, or misleading information in WhatsApp? If none, why?
- Have you ever seen or received any information on WhatsApp that you thought was false or misleading? If so, what happened? What did you do?
 - Who sent it to you?
 - Did you forward it?
 - Did the information consist of images, text, articles, or videos that you thought weren't accurate? Why did you think they were inaccurate?
 - How often do you see this type of content?
 - Has similar content ever appeared on another social media/messaging platform (e.g. Facebook News Feed)?

- What factors do you consider when deciding to trust information sent to you via WhatsApp?
- Do you forward information to your contacts?

Misinformation and recent events (COVID-19, BLM protests, U.S. election etc.)

- What kinds of information on COVID-19 have you received around WhatsApp?
- When was the last time you got a message on WhatsApp about COVID-19? What was it about? Did you think it was accurate? Why/why not?
- Have you seen more information sharing around COVID-19 on WhatsApp compared to before December 2019?
- Have you seen false, inaccurate, or misleading information around COVID-19 on WhatsApp? If so, can you give an example?
 - What did you do?
 - How did the information affect you?
 - Did you talk to the sender about it?
 - Did you fact-check it?
 - Did you ignore it?
- How has the information you've seen on WhatsApp affected your view/opinion on the country's (U.S.) situation with the pandemic (e.g. reopening phases, how COVID-19 affects youth, number of reported cases, conspiracy theories about origins of the virus)?
- How has the information on mask wearing/quarantine/social distancing affected your viewpoint with the COVID-19 information you receive?
 - How has the information on mask wearing + protests affected your viewpoint with the COVID-19 information you receive?
 - What about stay-at-home?
 - What about social distancing?
- What other messages about recent events have you received so far (BLM, elections, schools reopening)?
 - How have they affected your views on these issues?
 - How about your views on COVID-19, if at all?

Technology + fact-checking strategies

App features referenced are shown in Fig 1 (in the main paper).

- Have you used the WHO Health Alert on WhatsApp?
If not, why?
 - If yes, what did you think of its helpfulness/usefulness? How easy was it to use?
- The CDC has a bot on WhatsApp you can text to give you information on what to do if you think if you have symptoms. Have you ever used this? If not, why?
 - If yes, what did you think of its helpfulness/usefulness? How easy was it to use?
- Have you seen a new magnifying glass icon pop up beside some of your messages recently?
 - If so, have you tapped on it?
 - What did it lead you to and what did you think of it?
- How do you know what information given to you on WhatsApp can be trusted (or in general)?
 - What do you use to fact-check, if anything at all?
- What's your opinion on WhatsApp limiting the number of forward messages to lessen the spread of false information?
 - What led you to that opinion?
 - The limit is that one can only forward a message to 5 chats at a time.
 - When message is forwarded in a chain 5 times, it can only be forwarded to one chat (indicated with double arrow).
- Do you think WhatsApp can be improved to help address these issues with false, inaccurate, or misleading information? Why or why not?
- With other resources like Twitter's COVID-19 misinformation warnings (Fig. 1(d) in the main paper) and YouTube's information alert boxes (Fig. 1(c) in the main paper), would you want a better way to fact-check information in WhatsApp? Do you think these are enough? Why or why not?

Conclusion

- How has anything you said been vastly different from how you send or receive messages on other social media platforms you use?

- Is there anything else regarding WhatsApp that you want to talk about?
 - Desired technology?
 - False/inaccurate information?

Appendix B: Codebook

Code	Explanation
General	
Reason for using/liking WhatsApp	Participant explained why they like or use WhatsApp
Reason for disliking WhatsApp	Participant explained why they dislike WhatsApp, if they dislike it in any way
Chat Content	Participant talked about what they usually talked about in the chats, broadly
Foreign (non-U.S.) vs domestic communication	Participant uses WhatsApp to communicate with people in or out of the U.S.
Size of groups/chats they're in	Participants estimated the average size of the group chats they are in. They also gave exact numbers if they remember, or if they were in very few groups
Relationship with others in the group (with whom they interact with most often)	Participants identified relationships with others in their group chats
Active contacts/chat groups	Participants estimated the number of WhatsApp contacts they interacted with on a regular basis
Misinformation Encounters	
Information format (image/video/audio/text/links)	Participant describes the format of the information presented to them
Most recent misinformation encounter	Participant recounts most recent misinformation counter (info content, who sent it, their reaction, etc.)
Frequency of encountering misinformation	How often does a participant encounter misinformation? (e.g. once a week, month, year, etc.)
Who sends them misinformation content	Participant describes relationship with the misinformation sender (relative from abroad, immediate family member, etc.)
Frequency of forwarding links	Participant describes how often they forward links to their chats and messages
Misinformation indicators	Participant describes factors they consider when deciding to trust (and distrust) information
Reason for being active (talking with sender, fact-checking) about receiving misinformation	Participant explains how and why they are proactive when receiving misinformation (confronting sender, fact-checking)
Reason for being passive (ignoring) about receiving misinformation	Participant explains how and why they are passive/inactive when receiving misinformation
How WhatsApp content impacted their opinion on how the U.S. handled the pandemic	Participant explains how what they read on WhatsApp has impacted their opinion of how the U.S. handled the pandemic
How WhatsApp content impacted their opinion on BLM, 2020 elections, school reopenings	Participant explains how what they read on WhatsApp has impacted their opinion on other recent events: BLM, U.S. elections, U.S. school reopenings
Design Recommendations and Fact-Checking Strategies	
Willingness to use existing WhatsApp technology from reliable sources	Participants share their awareness of existing resources on WhatsApp from reliable sources designed to combat COVID-19 misinformation, namely the CDC bot
Fact-checking strategies	Participant describes how they fact-check information (Google search, literature, consulting others, etc.)
Efficacy of current WhatsApp features that combat misinformation	Participant describes the efficacy of WhatsApp features in fact-checking and limiting the spread of misinformation
Suggestions for improvement	Participant suggests improvements of current WhatsApp in bettering misinformation prevention/clarification

Concerns about the trade-off between combating misinformation and privacy/security	Participant raises concerns that fact-checking measures (e.g. information censorship) may undermine the privacy and comfort associated with end-to-end encryption
Features of other platforms	Participants share their opinions of existing features on other social media platforms (YouTube, Twitter, etc.) to combat misinformation.

Table 1: Our codes and corresponding explanations, organized by topic.

Anti-Privacy and Anti-Security Advice on TikTok: Case Studies of Technology-Enabled Surveillance and Control in Intimate Partner and Parent-Child Relationships

Miranda Wei, Eric Zeng, Tadayoshi Kohno, Franziska Roesner
Paul G. Allen School of Computer Science & Engineering, University of Washington
{weimf, ericzeng, yoshi, franzi}@cs.washington.edu

Abstract

Modern technologies including smartphones, AirTags, and tracking apps enable surveillance and control in interpersonal relationships. In this work, we study videos posted on TikTok that give advice for how to surveil or control others through technology, focusing on two interpersonal contexts: intimate partner relationships and parent-child relationships. We collected 98 videos across both contexts and investigate (a) what types of surveillance or control techniques the videos describe, (b) what assets are being targeted, (c) the reasons that TikTok creators give for using these techniques, and (d) defensive techniques discussed. Additionally, we make observations about how social factors – including social acceptability, gender, and TikTok culture – are critical context for the existence of this anti-privacy and anti-security advice. We discuss the use of TikTok as a rich source of qualitative data for future studies and make recommendations for technology designers around interpersonal surveillance and control.

1 Introduction

“Is my partner cheating on me?” “What is my teenager doing right now?” “How do I access something my parents restricted?” Questions like these have long existed in interpersonal relationships, and to answer these questions, some people turn to methods of surveillance and control. In recent years, the availability and accessibility of new technologies have enabled lay users to implement increasingly invasive surveillance and control over others. For example, tracking apps like Life360 facilitate precise location tracking of other individuals, and Apple AirTags can be misused to enable the same. These tools enable violations of security and privacy boundaries through unauthorized or unintended use of technology, or by otherwise transgressing others’ expectations.

In this work, we investigate a novel source of advice on

how to surveil and control others’ through technology: the social media platform TikTok. We find that on TikTok, users post detailed tutorials for surveilling their partners or children. Consider this suggestion to turn on the auto-answer call accessibility feature on a partner’s phone to detect cheating:

welcome to toxic tiktok 🤔🤔 i promise this isn’t me anymore! but lemme help you out!! if he’s not picking up, change this setting, it will automatically pick up all his calls! and if you hear stuff you didn’t want to hear... i’m so sorry bb 🥺 (TT45)

We call such videos “anti-privacy advice” or “anti-security advice”: *anti-privacy* or *anti-security* because the techniques often involve violating privacy or breaking device and account security, and *advice* because the videos are presented as guidance intended to be widely seen (more examples in Figure 1). We sought to answer the following research questions:

1. What information or systems are being targeted in anti-privacy or anti-security advice on TikTok and by whom? How are these attacks carried out and for what reasons?
2. How do anti-privacy or anti-security advice videos fit into the ecosystem of videos on TikTok, and how do they relate to a broader societal context?

To scope our study to a meaningful yet manageable size, we use case study methods to identify two interpersonal relationships as the contexts for our investigation: intimate partner and parent-child. We collect a dataset of 98 English-language TikTok videos and use qualitative methods to answer our research questions. First, we use a deductive approach to thematic analysis to apply a threat modeling framework to understand the assets, stakeholders, techniques, and motivations. Second, we use an inductive approach to thematic analysis to generate themes about how these videos are situated in the broader TikTok and societal context.

We find that surveillance in the intimate partner context is usually surreptitious and for the purposes of detecting cheating. Techniques used include leveraging tracking apps, obtaining unauthorized access to messages, and manipulations via physical access. In the parent-child context, surveillance

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

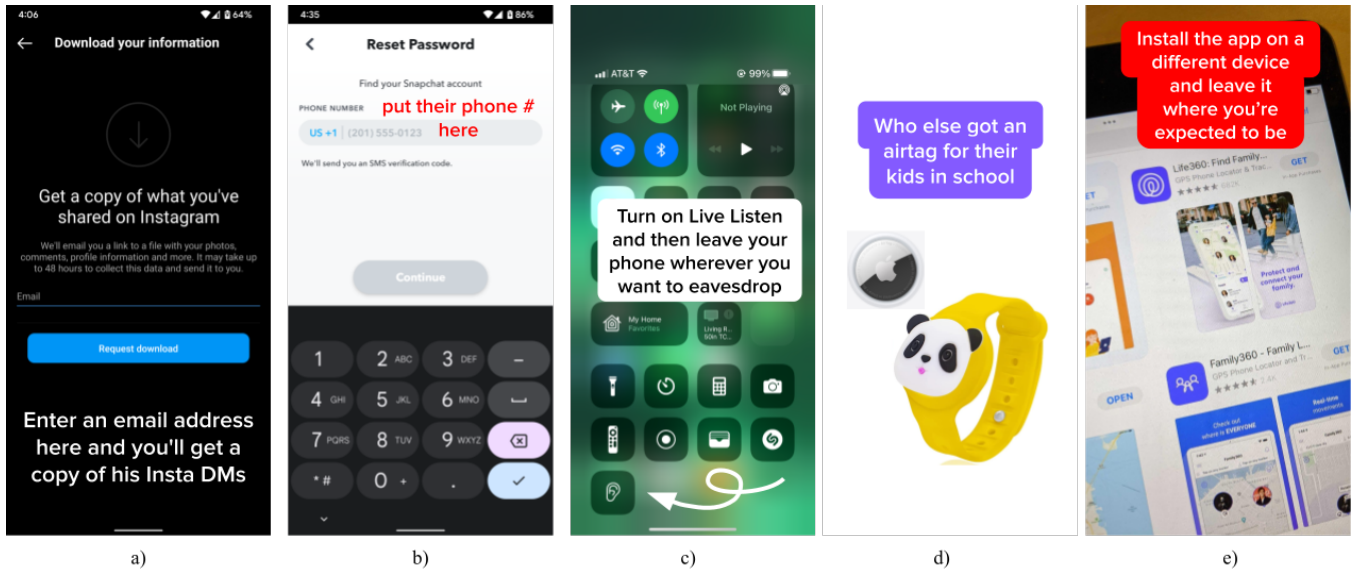


Figure 1: Examples of TikTok “Anti-Privacy and Anti-Security Advice,” recreated to protect creators’ anonymity. a) Surveil an intimate partner’s (“his”) Instagram messages by requesting a data download from the target’s phone, and sending it to the attacker’s email. b) Hijack an intimate partner’s Snapchat account to view their messages by recovering targeted account from the attacker’s phone, selecting “phone call” to verify the identity, and picking up the call on the target’s phone without unlocking it. c) Using AirPods’ Live Listen accessibility feature to surveil someone in another room. d) A parent using an AirTag necklace to track their child’s location. e) A teenager evading the Life360 app by installing it on an iPad that remains at home.

by the parent used family tracking apps and parental controls, is typically overt, and for ensuring child safety or restricting access to certain types of content. Meanwhile, teenagers in particular tended to resist these measures, and manipulated settings or broke authentication measures to evade tracking.

We generate themes about three social factors contextualizing the anti-privacy and anti-security advice we found. First, we identify that social acceptability influences framing of such advice: videos in the intimate context joked about being “toxic” because surveillance of other adults is transgressive, while videos in the parenting context framed techniques as helpful “#momhacks” for child safety. Second, we examine the influence of gender, given that a majority of language in the videos was feminine-coded, and how gender expectations could have contributed to the motivations of detecting cheating and protecting children. Third, we associate the engaging, easy-to-follow, and sometimes controversial characteristics of the anti-privacy and anti-security advice videos with TikTok’s competitive culture of creating viral content.

Our investigation sheds light on an ecosystem of people sharing anti-privacy and anti-security advice on TikTok. We close by discussing our findings’ implications for the computer security and privacy community and surfacing opportunities to address the risks introduced by anti-privacy and anti-security advice, while also recognizing that technical fixes will not fully address the associated social and societal challenges. We also reflect on the benefits and challenges of TikTok as a qualitative data source.

In summary, we make the following contributions:

1. We identify TikTok as a source for rich qualitative data about “anti-privacy” or “anti-security” advice, and conduct case study investigations of 98 videos about technology-enabled surveillance and control. We study two interpersonal contexts: intimate partner and parent-child relationships.
2. We identify assets, stakeholders, techniques, and motivations in anti-privacy and anti-security advice.
3. We generate themes about how these videos are situated in the broader TikTok and societal context.
4. We discuss our findings’ implications, identifying opportunities in security and privacy research and practice.

2 Related Work

2.1 Interpersonal Security and Privacy

Most closely related to our work are other studies of security and privacy as indexed by specific interpersonal relationships.

Intimate Partners. A growing body of scholarship studies adversaries and their methods in intimate partner relationships. Freed et al. categorize attacks into four categories based on the resources abusers leverage and their intentions [26]. Other studies investigate spyware apps for intimate partner surveillance (IPS) [15], as well as creepware for interpersonal attacks [51]. Tseng et al. [59] create a taxonomy of IPS tools discussed on IPS forums. In our work, we do not know if the TikTok creators giving anti-privacy or anti-security advice

actually use such techniques to abuse, but we highlight the potential for such advice to do so. Our context of study is also different: TikTok is an open platform, compared to narrower populations in prior work, e.g., survivors contacting Family Justice Centers [26] or those on dedicated forums [59].

Other work examines how to effectively design interventions supporting survivors [30,60,68], particularly by working in consultation with survivors to map concerns [26]. Complementing these intentional efforts, the observational nature of our work allows us to see attacks organically discussed on TikTok, for informing countermeasures and support.

Many scholars studying the intimate context draw attention to its complexities. For example, intimate partner violence (IPV) targets must negotiate tensions such as seeking distance despite social, financial, or other connections to abusers [25]. Levy & Schneier highlight common privacy assumptions made by computer scientists that do not hold in intimate relationships [37]. We join these scholars by diving into the murkiness of interpersonal relationships through the content that perpetrators and targets themselves create and post on TikTok.

Contrasting prior IPV and IPS work, our dataset includes social media stalking techniques used *before* a relationship begins, perhaps more akin to the privacy of online dating [16] or online status indicators [17]. This may speak to the normalization of intimate surveillance [38] with new technologies.

Parent-Child. Many scholars have also investigated familial privacy boundaries. One body of work interrogates the information sharing that some parents engage in — sharenting — when children are younger and unable to consent [3,9,10], as well as the normalization of parental monitoring [36,55]. Some scholars draw attention to the increased risk of “dataveillance” from parents [42,67]. Studies of parental control apps find that apps are purportedly for safety, but may favor parents’ desires at the cost of childrens’ [65], contributing to negative experiences [29], especially if designed incorrectly [63].

Between parents and their teenaged children, user studies of privacy boundaries find different technology understandings and preferences for monitoring or autonomy [19,20], but also expectations that parents and children will collaborate to find the right balance [56]. The tension between parents’ desire for information and control to ensure safety with teens’ desires for autonomy and privacy has also been documented in the context of specific technologies, e.g., IoT entryways [31,61], smart speakers [35]. The openness of TikTok creators allowed us to observe parents’ opinions and suggestions for surveillance and control, as well as the teenagers’ countermeasures.

2.2 Security Advice

Security and privacy researchers have studied what *pro*-security advice exists, its sources, and its quality [47–50]. Other work also investigated advice for specific communities, e.g., queer individuals [28], or contexts, e.g., in workplaces [21,22], after “triggers” [23], during civil rights

protests [5,62]. In this work, we instead study *anti*-security advice, or advice on how to compromise others’ security and privacy through methods of surveillance and control.

Aside from Tseng et al.’s work on IPS forums [59], we are aware of little academic work studying how security and privacy adversaries learn. Some low-tech techniques in videos we study call to mind advice from other contexts, e.g., social engineering and low-tech hacking guides [41].

2.3 TikTok

As TikTok is only 5 years old, TikTok research is still in its early stages. Some study specific subcommunities, e.g., populations with disabilities [24], healthcare workers [53], or aspects of TikTok’s culture, e.g., authorship practices [34], visibility [1]. Other work leverages TikTok as a repository for specific content, e.g., public health messaging [2,4,40], social activism [18], science memes [66], political communication [52]. We add to this growing body of work by studying anti-privacy and anti-security advice: content that teaches how to surveil or control others through technology. De Leyn et al. study tween privacy perceptions, but in conjunction with parents [39], whereas this work studies when parents may pose the privacy risk.

3 Background

TikTok is a social media platform on which users post short-form videos (also called “TikToks”). In early 2020, TikTok became the most downloaded app in the world, and reached 1 billion monthly users in late 2021 [43], demonstrating enormous growth relative to older social media platforms. As of early 2022, 35% of TikTok’s users are between 19 and 29 years old and an additional 28% are under 18; only 18% are between 30 and 39, and 19% are over 39 [33].

Usage. TikTok’s primary interface is the For You Page (FYP), an infinite scroll feed of autoplaying videos. The FYP serves videos using a recommender system, which personalizes recommended videos based on engagement metrics such as dwell time, likes, and comments. Content can also be viewed in the Following tab (to see content from previously followed creators) or the Discover tab (to search for videos or see trending topics). TikTok displays videos full screen (on mobile), and it is only possible to watch TikToks one at a time, swiping up to display the next video.

In addition to the video (often showing the creator in portrait mode), TikToks frequently include overlaid text (which may be read aloud by a built-in voiceover feature), TikTok’s own set of sounds (including licensed music), and various visual effects. Users can interact with content by liking, commenting, or sharing videos; following TikTok creators; or remixing other TikToks.

TikTok subcommunities. Subcommunities on TikTok are loose associations of creators and followers interested in a specific topic, often organizing around certain hashtags, e.g.,

#egirl (rebellious women gamers turned fashion aesthetic), sometimes with a play on the platform name, e.g., #momtok (moms on TikTok), #fittok (fitness TikTok). Relationships are one such subcommunity, with users posting anything from inspirational relationship content, to giving advice, to calling out toxic behaviors. The top relationship-related hashtag is #relationship with 90.1 billion views. Another subcommunity discusses various aspects of parenting, including sharing advice or personal experiences. The top parenting-related hashtag is #parenting with 13.0 billion views.

4 Methods

We investigate anti-privacy and anti-security advice on TikTok through case studies of two interpersonal contexts. We selected these contexts informed by case study methods and collected a total dataset of 98 TikTok videos (see 4.1). For data analysis, we performed procedures from the qualitative methods family of thematic analysis (see 4.2). Although our research did not directly recruit participants, and as such, our institution’s IRB determined our work not to be human subjects research, we still recognize that we are studying real people: we carefully made ethical considerations to protect the subjects of our research (see 4.4). We conclude by contextualizing the goals of this work with its limitations (see 4.5).

4.1 Case Selection and Data Collection

We summarize our overall approach to data collection, which occurred between November 2021 and February 2022.

We used progressive focusing [54], an approach from case study methodology, to iteratively narrow our research questions as well as select which cases we used. In his influential 1995 book, *The Art of Case Study Research*, Stake describes progressive focusing to place a high emphasis on interpretation that allows for flexibility during the research process because “the aim is to thoroughly understand [the case]. If early [research] questions are not working, if new issues become apparent, the design is changed.” [54]

In this work, our case was centered on English-language TikTok videos that described technology-enabled techniques for harming others’ digital security or privacy, i.e., anti-privacy or anti-security advice. Our criteria for inclusion of a TikTok video as anti-privacy or anti-security advice were: (a) does the video describe a technique that requires technology,¹ (b) does the technique involve violating privacy or security measures or boundaries, and (c) does the technique implement (or evade) surveillance or control?²

Initially, we tried searching for security and privacy related terms using the built-in TikTok search interface to surface relevant videos: e.g., “hacking,” “security,” “violate privacy,” “surveillance.” These terms are meaningful to the computer science community, but we discovered they were not to Tik-

Tok creators nor viewers. Instead, we realized that we would need to first identify contexts in which anti-privacy or anti-security advice could be common, and then find videos in those contexts that included technology-enabled techniques.

We conducted a literature search to identify contexts in which anti-privacy or anti-security advice could be common. We considered the following contexts (that we did not include): smart homes, proctorware, hidden cameras in vacation rentals. We searched for videos in these contexts, finding the most qualitatively rich videos in intimate partner and parent-child relationships, which we finalized as our cases.

We collected more data by adding context-specific search terms to our original set: in the intimate partner context, e.g., “toxic,” “relationships,” “cheating,” and in the parent-child context, e.g., “parental controls,” “life360,” “kid tracking.” Data collection was an iterative process between two members of the research team, who recorded relevant search terms and frequently met to discuss data collection efforts.

The majority of data collection concluded when we felt that we had exhausted the relevant search terms and could not find more videos, and that we had a rich enough dataset for analysis. Drawing from case study methods, we continued triangulating — “working to substantiate an interpretation or to clarify its different meanings” [54] — throughout our analysis and writing. By iteratively searching for relevant videos to confirm or deny our findings and interpretations, we continued to make refinements and added 21 videos in this manner. Our final dataset consisted of 98 anti-privacy or anti-security advice videos: 66 videos in the intimate partner context, 27 videos in the parent-child context, and 5 relevant to both. Altogether, our dataset accounts for 60 minutes and 14 seconds of audio-visual content, with a total of over 16 million likes (mean = 171K, median = 4.5K, max = 3.2M). For reporting, we abbreviate the *x*th TikTok in our dataset to TT*x*. We note that our dataset is a case study, and prioritizes qualitative depth over quantitatively measurable claims.

4.2 Data Analysis

We conduct thematic analyses of our data, a broad family of methods that is flexible with respect to conceptualization of the data and its meanings, inductive or deductive orientations, and the procedures that can be used [7, 8].

Deductive Thematic Analysis. The first part of our analysis focused on our first research question about (a) what information or systems are being targeted, (b) by whom, (c) using which techniques, and (d) for what reasons. We used a codebook approach [7, 8] to deductively (theory-driven) apply a security threat modeling framework to our data. Because of the significant theoretical value of this framework to security and privacy researchers and practitioners, the codebook approach permitted us to develop these questions early in the research process. First, two coders familiarized themselves with the videos by watching them multiple times, taking notes separately (this initially began concurrently with data collection).

¹Thus, we excluded videos without a technology element.

²Thus, we included videos where the technique was been demonstrated in the video with consent, but could also be used without consent.

They then met multiple times to develop four codebooks: stakeholders, assets, motivations, and techniques. Using these codebooks, one coder coded intimate partner videos, the other coded parent-child videos. Lastly, both coders reviewed each others' work, discussing and resolving concerns.

Inductive Thematic Analysis. For the second part of our analysis, we used a less structured approach to inductively (data-driven) generate themes about the social factors that contextualize the anti-privacy and anti-security advice we collected on TikTok. We did this by continuously meeting with all members of the team to discuss higher-level observations we made about the data, and drafted memos about these broader ideas. Through this iterative process [45], we developed three themes about the social context of such advice (Section 7).³ To ensure thoroughness, we also triangulated [54] these themes by going back to do more data collection, or add new elements of analysis, as necessary. For example, to triangulate our findings about the gender in Section 7.2, we went back to the data with a gendered lens.

4.3 Positionality Statement

In the process of our inductive thematic analysis in particular, as well as our overall research approach and perspective, we acknowledge our active role as researchers in the process of knowledge production [6] and regard our “subjectivity as analytic resource” [8]. Our research analyses and interpretations are the result of our particular social, cultural, historical, disciplinary, political, and ideological positionings [8]. Here, we describe our identities and how they relate to the interpersonal contexts (i.e., intimate partner and parent-child) and research data (i.e., TikToks) we study. Our research team is composed of two cisgender women and two cisgender men. Two researchers are in their 20s, one is in their 30s, and one is in their 40s. All researchers have experience with intimate partner relationships and two are parents. One researcher has 24 months of experience with TikTok, another has 6, and another has 3 at the time of these analyses.⁴

4.4 Ethical Considerations

We consulted with our institution's IRB, which determined that our study did not require review as human subjects research because the videos that we analyzed were publicly available at the time that we collected them. However, we recognize that IRB review is not sufficient to guarantee ethical research. In particular, there are ethical considerations with studying public data that was created and shared for purposes other than research [12], even if many of the videos we study have reached large audiences in the context of TikTok (and beyond — we observed some news articles about creators in

³Due to the deductive thematic analysis approach we used for applying the threat modelling framework to our data, as well as the observational nature of TikTok videos, we did not conduct a fully reflexive thematic analysis [6].

⁴The other co-author first heard about TikTok through his collaborators and only accesses it through links provided by the other three.

our dataset). To mitigate potential harms that may come from exposure of the content we study to unexpected audiences, we paraphrase creator quotes and recreated screenshots of the videos in this paper, to preserve semantic meaning while obscuring the original source. We also aim to present our data in broadly descriptive or interpretive, rather than individually judgmental, ways — we recognize that there is additional context behind the motivations and situations of creators and viewers of the content we study that we may not fully understand. Ultimately, our goal is not to study the specific people who post or engage with this content, but rather to use this data as a window into popular use of interpersonal control and surveillance techniques more generally.

Our research also surfaces complicated social ethics considerations. The surveillance and control techniques we study have a tangled relationship with the interpersonal situations they are embedded in, including non-consensual surveillance, cheating, child safety, and fostering trusting familial relationships. Our work cannot resolve these ethical questions, but as security and privacy researchers, our goals are to enable an informed conversation about security and privacy risks, and hope that our findings contribute to a better understanding of the use of surveillance and control techniques.

4.5 Limitations

Our investigation necessarily considers only a slice of data from TikTok, focusing on specific subcommunities, at a specific point in time, and limited by the videos we were able to surface via our data collection methodology and TikTok's search capabilities. There are likely relevant videos on TikTok that are not included in our dataset, so there may be motivations or techniques that we missed. Moreover, there may be other related subcommunities that our searches did not surface, e.g., communities who respond to the videos we analyze or create similar videos in other contexts. Accordingly, our analysis focuses on surfacing the breadth and depth of interpersonal surveillance and control motivations and techniques that the videos we study cover, not on understanding TikTok as a whole or on comparisons with different subcommunities.

Additionally, content on TikTok is, as on any social media platform, created and edited in order to present people and the topics they are discussing in a certain way. Our study uses TikTok data as a window into people's motivations, techniques, and responses to interpersonal surveillance and control, but (of course) does not give us information about the creators' actions or opinions beyond what is projected in the videos.

Finally, we come to TikTok and to our research questions as observers, not as TikTok content creators ourselves. There are likely unique aspects of content creation that we do not understand. However, as mentioned, several of us have significant experience immersed in TikTok as passive users.

5 Findings from the Intimate Partner Context

We collected a total of 66 TikTok videos in the intimate context. Of these, 64 were about implementing methods of surveillance and control, while 2 were about defenses. These videos were created by 25 unique TikTok creators: 18 came from Creator A, the most prolific creator; 9 came from Creator B, the second most prolific; 8 came from Creator C; and 1 video each came from seventeen creators.

5.1 Stakeholders, Assets, and Motivations

We present a summary of the stakeholders, assets (and associated technologies), and motivations in Table 1.

Explicit and Implicit Concerns about Cheating. In the videos we collected in the intimate partner context, **instigators** are interested in obtaining information about **targets**, primarily to detect cheating. Cheating concerns were sometimes made explicit by using the words “cheating” or “suspicious” (or variants thereof). We observed that many videos began with this motivation, e.g., “Do you wanna find out if your partner cheats?” (TT36), potentially to capture a viewer’s attention. Sometimes this motivation arose later, e.g., the instigator in TT18 says, “keep watching if you wanna find all Twitter conversations between your partner and someone you’re suspicious of.” The creators also made their motivation as instigators explicit by naming an audience member’s relationship to a target, e.g., “How to figure out if your partner is cheating on you” (TT10).

In other videos, concerns about cheating were implicit: for example, by implying a target’s identity by their gender: “Trying to get into his Snapchat?” (TT38). Some videos included techniques that were substantively similar to those in videos explicitly motivated to detect cheating, or sought to find evidence of cheating behaviors (e.g., communicating with someone else, being at certain locations) or contained context clues about catching a target, e.g., “Heh you can’t hide from me dummy 😏” (TT34).

Targeted Assets. Instigators sought to compromise a variety of targets’ assets: aligned with the motivation of detecting cheating, instigators creatively postulated all the digital traces that could be treated as proof, including sexually explicit photos or emails from hookup websites. Location in particular was treated as more conclusive proof if instigators used technology to verify that targets had been at suspicious locations. Social media assets, such as who targets followed or messaged, were used sometimes as less conclusive evidence, e.g., “as a preliminary step to confirm or deny my suspicions, before I get into a full investigation” (TT40).

Other Motivations. A minority of videos were not motivated to detect cheating, and were instead about general behaviors of surveillance and control in intimate relationships. These behaviors may cross targets’ personal boundaries, breaking their existing security measures or invading their privacy, either because a target would reasonably assume certain information

to private, or in some cases, because a target had explicitly set that boundary. Some instigators sought to surveil targets at all hours of the day, even absent suspicions of cheating, or generally spy on as many of their target’s digital activities as possible. Targets’ motivations were to maintain autonomy, especially in the face of potential surveillance.

5.2 Intimate Surveillance and Control

Next, we break down the specific surveillance goals and techniques of instigators. We observed at least 24 distinct techniques for surveillance and control, underscoring the variety and creativity of instigators in this context. Though we do not pose this is an exhaustive list of all techniques discussed on TikTok, we detail these techniques to surface the breadth of how instigators surveil and control their targets. The full set of goals and their associated techniques are in Appendix A.

5.2.1 Goal: Surveil Digital Communications

Instigators were interested in learning who targets were communicating with, and what those communications contained, (presumably) to determine whether they were texting with an affair partner. Several methods were suggested for obtaining information about the targets’ SMS or social media messages.

Technique: Exploit Data Downloads. One method for obtaining a target’s messages and communications was through the data download feature of social media platforms: GDPR’s Right of Access requires data subjects to be able to download archives of their data. Instigators noted that on platforms like Instagram, Snapchat, and Facebook, these data downloads can be used to obtain a copy of their messages, allowing them to search for evidence of cheating (Figure 1a). Three separate creators made tutorials for locating the data download in the settings interfaces of the above platforms. This attack relies on having physical access to the device or account access.

Technique: Gaining Direct Account Access. Another method for obtaining a target’s messages was to obtain direct access to the target’s social media account to view the target’s messages in the app. One video describes hijacking the target’s Snapchat account through the account recovery process, which only requires physical access to their phone (Figure 1b). The instigator attempts to recover the account password on their phone. Snapchat sends an authentication code via phone call, which the instigator can pick up without unlocking the phone. After confirming, the instigator can reset the target’s password, accessing the target’s Snapchat messages. Another approach suggested is to add the instigator’s phone number to the target’s iCloud account, which may enable the instigator to get a copy of their messages.⁵

Technique: Emoji Side Channel. Two TikToks suggest the target’s frequently used emojis in their keyboard as a side channel for detecting cheating. If sexually suggestive emojis

⁵This technique does not work without also enabling message forwarding, which requires additional authentication.

Table 1: A summary of the stakeholders, assets (and their associated technologies), and motivations we observed in our dataset. This table is intended to give a sense of the broader context and attack space; we note that our methods were qualitative and thus these results are not able to make exhaustive claims about what attacks are possible, nor quantitative claims about frequency.

	Intimate Partner Context	Parent-Child Context
<i>Stakeholders</i>	Instigators surveil targets' data or digital footprint, or otherwise exert control on targets' digital activities	Parents are the caretakers of children ; childrens' ages ranged from early school age to teenagers
<i>Assets</i>	Location; social media accounts; social media data (who targets followed, messaged, or content targets posted); web browsing history; photos; live audio; dating app usage	Location and location privacy; access to specific types of content; access to communications; privacy about digital activities
<i>Technologies Targeted or Used</i>	Apple software and devices (iOS, iPhones, AirTags, AirPods, Apple Watches); Android (Google Maps); social media platforms (Instagram, Facebook, Twitter, Snapchat, Tinder); email; phone calls; family monitoring apps	Apple devices (AirTags); Life360; Bark; FamiSafe; parental control features; VPNs
<i>Motivations</i>	Instigator: Detect cheating; general surveillance; control contact with targets Target: Evade surveillance; maintain autonomy	Parent: Child safety in the physical world and online Child: Autonomy; privacy

(e.g., 🍊, 🍆, 🍷) were present and the target did not use them while communicating with the instigator, it suggests the target is sexting with someone else. This technique only requires non-privileged physical access to the phone: TT40 suggests opening an iPhone's "Today's View," accessible from the lock screen and containing a keyboard in the search bar.

5.2.2 Goal: Stalk on Social Media

Another goal for instigators was to stalk a target's activities on social media, either for generally monitoring their online presence, or for specifically finding evidence of cheating.

Technique: Read Twitter Conversations. One video suggests using Twitter's advanced search to find conversations between two specific people, to look for evidence of cheating.

Technique: Anonymous Viewing of Instagram Profiles. Instigators may be interested in viewing their targets' Instagram profiles; however, activity like following or viewing stories is visible to the target. To view stories anonymously, one video suggested creating a fake Instagram account to watch stories, while another suggested using a third-party site that claims to allow anonymous viewing. A different third-party site was suggested for enlarging a target's profile picture, which are usually only shown in a small size through the app.

Technique: Side Channels in Social Media Platforms. Other videos highlight side channels that leak information about the target's activity. For example, an instigator could determine the order in which a target follows other accounts, by viewing their "following" list on the web version of Instagram, which shows follows in chronological order.⁶ Another video suggests that instigators can infer whether a target is sending Snapchat messages (e.g., sexts) to a large number of people or to an individual, by tracking the target's Snapchat score over time, and observing how much it increases.

⁶This is no longer works as of the writing of the paper.

Technique: Track Online Status Indicators. Instigators may want to know when a target is online on a messaging app to infer other aspects of their behavior (e.g., are they actually asleep, or did they lie about it?). One instigator names a third-party app that specifically sends notifications each time a WhatsApp contact signs on or off.

Technique: Contact Someone Who Blocked You. One video demonstrates texting someone who blocked you by sending from an associated iCloud email address.⁷

5.2.3 Goal: Surveil Dating App Usage

Instigators presented techniques to infer whether targets were using dating apps despite being in a relationship with them.

Technique: Find Target's Profile on Dating App. One approach is to find the target's profile on the dating app. One video suggests creating a fake account on the dating app, and swiping through profiles manually. They also suggest setting the search radius to the minimum while physically near the target narrow down the available profiles as much as possible. Another suggests a paid third-party service called "CheaterBuster" that will look for the target automatically.

Technique: Infer Dating App Usage. Other videos suggest more indirect approaches. One video suggests attempting to create a dating app account with the target's email address to see if the email address is already in use, indicating they are signed up for that service. Another suggests looking through the App Store for dating apps — the list of downloaded apps shows not only which apps were installed, but when they were first purchased or installed. This would indicate if they recently installed a new dating app.

5.2.4 Goal: Surveil Other Digital Activities

Instigators also aimed to surveil targets' other digital activity, including monitoring their browsing history for watching

⁷According to many comments, this technique does not seem to work.

porn, and searching their phones for sexually explicit content.

Technique: Searching for Explicit Content. Some videos instructed viewers to look for explicit photos in the photo gallery, as well as explicit content in the target’s email and web browsing history. One video warned viewers of an app that could hide explicit photos while appearing to be a calculator, and noted that observing a target’s reaction to being asked about whether they had this app might be informative enough.

Technique: Photo Metadata. One video suggested an app that automatically parsed EXIF data to show when a photo was originally taken, which allows inferring whether a sexually explicit photo had been, according to the instigator, “reused”: “let’s say you get a pic of their nuh-uh today, but if the pic was taken five months ago, who else might’ve gotten that pic, hm?” (TT16).

5.2.5 Goal: Manipulate Social Media

Instigators creatively manipulated the functionality of social media and messaging apps to obtain outcomes they desired.

Technique: Restrict and Unrestrict. Two videos advocate reading an Instagram direct message by blocking the sender, which then sends the message to a request inbox that does not send read receipts. Similarly, another advocates manipulating a target’s Instagram story feed by hiding a story from the target, and then unhiding, which makes the story appear first.

Technique: Fake Tags. One video describes creating a fake “tag” with the poll feature in an Instagram story that appears to be tagging another user, but instead tallies how many people clicked on the fake tag.

Technique: Message Deletions. One video describes how to delete WhatsApp messages more than an hour old: changing the system time to within an hour of the message timestamp.

5.2.6 Goal: Surveilling Physical Activities

Instigators were also interested in surveilling targets’ physical-world activities, such as their physical location, or hearing their conversations, which could provide evidence of cheating.

Technique: Tracking Location with Apple Products. A very common technique described by instigators is to use AirTags, AirPods, or Apple Watches to track a target’s location. This is done by secretly hiding one of these in the target’s belongings or car (one video demonstrates hiding it in the side pocket specifically). TT25 acknowledges that this would be “super toxic,” but one could “forget, on accident of course, an Apple device in their car and then track their every move.” In another notably overt example, an instigator makes an AirTag necklace with a customized design, names the AirTag “Cutie pie 🍪”, and gives it as a present to her boyfriend. We also observe one instigator discussing an unsuccessful attempt, as Apple’s mitigation alerted their target that they were being tracked, and later found the AirTag discarded in a bush.

Technique: Abusing Accessibility Features to Spy on Au-

dio. Instigators developed techniques for surreptitiously listening to their targets’ conversations. Some videos advocated for using Live Listen, an accessibility feature which enables an iPhone or iPad to act as a microphone to send sound to AirPods (intended for use with hearing aids, or in a noisy location). An instigator could leave their phone with the target, leave the room, and listen via AirPods (Figure 1c). Others suggested taking the targets’ phone, enabling Auto-Answer for phone calls (intended for Touch accessibility), and calling them whenever they wanted to listen to what they were doing.

Technique: Use Tracking or Monitoring Apps. Three videos advocate installing location monitoring apps (e.g., Life360) or using OS-level tracking features (e.g. Find my Friends) on partners’ phones. These videos report the location of a target in real time. Another strategy suggested by instigators was to use the iOS Significant Locations feature or Google Location History to identify locations that the target visited in the past, which could reveal if the target had been dishonest about where they had been.

5.3 Countering Intimate Surveillance

We now review targets’ strategies. In the 2 videos we collected, targets’ goals were to counter surveillance. These defenses do not counter any of the instigator techniques we found, which could be a result of our methods (Section 4.5), and does not necessarily mean such content is not on TikTok.

Technique: Detect call surveillance. Two TikToks described checking phone carrier settings to check for call forwarding or redirection. However, the videos did not suggest purposeful next steps if found: “if any are enabled... scream” (TT54).

6 Findings from the Parenting Context

We collected a total of 27 videos in the parent-child context; 16 from parents, and 11 from children. These videos were posted by 25 unique TikTok creators, distinguishing this context from the intimate partner context where three creators accounted for over half of videos.

To facilitate comparison with the intimate context, we standardized our terminology to use “surveillance” and “control” for methods used by parents to track, monitor, or restrict their children’s activities. In the parent-child context, these methods are more ethically ambiguous than the intimate partner context, and may not always be adversarial. The appropriateness of certain methods may depend on the age of a child or the overall nature of the parent-child relationship. Though some creators shared techniques with positive intentions, viewers may not necessarily share those intentions. Further, such videos may contribute to the normalization of parental surveillance [55].

6.1 Stakeholders, Assets, Motivations

In the parenting context, we observed videos from **parents** and **children**, primarily teenagers (old enough to have a smartphone and a TikTok account). Tensions centered around par-

ents having the right level of information and control to ensure childrens' safety, while children wished to have enough autonomy to ensure their own privacy. A summary of the stakeholders, assets, and motivations is again in Table 1.

Parent Perspective. When children were younger, parents were concerned about physical safety and leveraged technologies to track their location, especially when not in their supervision, e.g., riding the bus to school. Some captions alluded to more general concern: "Extreme measures are essential these days. Track kids with #airtag bracelets" (TT57). As children got older, concerns centered more on access to certain content, so some parents relied on family tracking apps, parental control features, or other technologies made for these concerns. Parents were concerned about children accidentally downloading malware or making purchases, messaging strangers, using rude or profane language, encountering explicit material, and having excessive screen time.

Child Perspective. Children's videos were motivated to evade tracking or restrictions by a desire for greater autonomy, particularly in the face of restrictions (e.g., on internet and app usage) and tracking software (e.g., for location) on their phones. Children were also motivated to hide their apps and texts from low-tech monitoring, like manual inspection by parents.

6.2 Parental Surveillance and Control

We now describe the specific goals parents had regarding child safety, and the techniques and tools used to reach those goals. Again here, we do not pose this is an exhaustive list of all possible techniques, but rather detail them to surface their breadth. Generally, parents used commercially available tracking and parental control tools, or parental control features built into mobile operating systems. Compared to the intimate partner context, parents typically used these features as intended, rather than abusing features. The full set of goals and their associated techniques are in Appendix A.

6.2.1 Physical Surveillance

Parents were interested in knowing the exact physical location of their children, for emergencies or general peace of mind.

Technique: Location Tracking with AirTags. Many of the videos from parents advocated using AirTags in order to keep track of their children's location, touting how cheap, accessible, and effective they were: "#Apple #AirTag this is so smart, only \$30, so worth it ❤️👉" (TT60). Essentially all of these were made by moms for younger children (younger than pre-teen) and a few described this technique as a "mom hack." As noted above, the motivations were to keep children safe. The parents mainly showed their personal experiences of making an AirTag bracelet, keychain, or necklace and putting it on their child (Figure 1d), while a few also showed putting (or hiding) an AirTag in their child's bag or shoes. One in particular noted that a keychain attached to their child's belt loop, instead of backpack, was the best option "because backpacks

are always left behind when something happens" (TT65). We suspect that parents chose to use AirTags with younger children because they do not yet have smartphones with which tracking apps can be used.

Technique: Location Tracking with Apps. For older children, parents described using specialized mobile apps, especially Life360, to monitor their activities. Life360 is advertised as a family location sharing app, which also provides emergency assistance alerting and digital safety tools to monitor identity theft or credit scores. One parent described using Life360 to monitor their kids while they went to school and extracurriculars (TT63).

6.2.2 Goal: Online Safety and Monitoring

Parents are also concerned about kids' online safety, and employed a variety of apps and tools to restrict access to the internet and apps, and to monitor communications.

Technique: Monitoring and Parental Control Apps. Some parents described using third party apps to impose parental controls and monitoring to their kids' smart phones. Apps mentioned include FamiSafe and Bark, which are advertised as online safety apps that monitor social media content for appropriateness as well as time limits on certain apps. Bark alerts them if profanity was detected: "privacy with a safety net" (TT97). Another set of parents created a sponsored video where they describe using FamiSafe's app download allow list to restrict their kids to trusted apps (fearing that their child might install malware on their phone).

Technique: Fully Locking Down Phone. One parent advocated for a fully locked down phone from Gabb Wireless, which had built-in parental control tools for screen time restrictions and content filters (including no access to any social media platforms), while still allowing for some phone functionalities like calling and texting.

Technique: Monitor Messages with System Features. Parents could also use built-in operating system features to perform monitoring of their children. One video explained how to monitor a child's text messages: parents can add their phone number to the child's iCloud account, and then update the settings to forward all messages to the parents' device(s).

6.3 Children's Defenses

Teenagers' primary goal in our dataset was to evade surveillance or restrictions placed on their phones by the parents; such as location tracking apps or parental controls. These techniques were generally reactive, not proactive, to parents' usage of certain commercial products or device features.

Technique: Disrupting Location Tracking Apps. Children described a number of ways to evade location tracking apps like Life360, e.g., disabling cellular data and motion and tracking permissions for Life360, while leaving location and WiFi permissions on. This prevents the app from reporting back real time location updates, but does not notify parents that

the location permission was disabled. Another technique was to install the Life360 app on another device that could be left at home (Figure 1e). Another video claims that putting the iPhone in Do Not Disturb mode would disable tracking, though commenters disputed this method.

Technique: Bypassing Parental Controls. Teens also found techniques to bypassing parental controls, which may restrict screen time, app downloads, or access to certain websites, depending on the software and how the parents configure it.

Two children described guessing the parental control passcode by examining the fingerprints left by their parents. One suggested wiping a screen perfectly clean, and another by getting a screen very dirty, and then asking parents to unlock or temporarily allow access to apps. Then, by looking at the location of the fingerprints, they systematically guessed the possible combinations. For parental controls that use a VPN to intercept web and message history, like Bark, one video suggested removing the VPN in the system settings. Lastly, to bypass App Store restrictions on which apps can be downloaded, one user suggested signing out of their iCloud account, logging into a new iCloud account to download the app, and then signing back into their usual account.

Technique: Hiding Digital Activity with OS Features. Two children advocated for a technique specifically for when parents ask to see their phone. To hide certain apps, the children described an iOS feature that hides certain homepage screens, so that the parent would not see certain apps.

7 Social Context of Anti-Privacy and Anti-Security Advice

We now present themes from all 98 videos across both settings, stepping back to consider broader social contexts.

7.1 Social Acceptability

Though on a technical level, videos in our dataset all contain advice on breaking or potentially misusing computer security and privacy features, we saw notable differences in how socially acceptable the creators perceived their advice to be, and whether the techniques were meant to be covert.

Intimate Partner Hacking: Socially Unacceptable, Covert. In the intimate partner context, creators often demonstrated performative self-awareness about how their videos were taboo, transgressive, or could be illegal or considered violations of privacy. Captions for these videos often included hashtags or phrases like “#toxic”, “#stalker”, “#crazygirlfriend” (referring to self), or “#hacks”. Some creators put disclaimers at the beginning of videos or in their account profiles, declaring that their videos were not to be taken seriously:

Disclaimer: Techniques shown here should not be replicated. If you are actually crazy, you should probably get medical help. These videos are only for entertainment and informational purposes. Use this as you will. (TT19)

Techniques used by instigators in the intimate context often had covert objectives, such as viewing content anonymously, secretly getting unauthorized access to a device or account, or abusing existing features like platform user blocking.

Parental Surveillance and Restrictions: Socially Acceptable, Overt. In contrast, videos about anti-privacy or anti-security advice in parent-child relationships were not framed as deviating from social norms. For parents’ videos, because the motivations of child safety are widely accepted, creators tended to frame their videos as helpful tips: “I really strongly recommend using AirTags if you have a kid going to school on public transit” (TT64). The techniques and tools used by parents, such as Apple AirTags, parental controls on smartphones, and apps designed for family tracking or child safety, like Life360, are commercially available, and used for their intended purpose, rather than covertly used or misused. Rather than secret surveillance methods, parents openly put AirTags on their childrens’ wrists or clothing or enabled parental controls on their childrens’ phones.

Teens Evading Surveillance and Control: Socially Acceptable, Covert. In teenagers’ videos on evading restrictions and tracking, although their techniques were often intended to be covert and undetectable by parents, none of the creators framed their videos as socially unacceptable. For example, multiple videos gave advice for disabling location monitoring in the Life360 app so they could leave the house without alerting their parents. The techniques were intended to be discreet, but the creators did not portray doing so as ethically wrong.

Why These Differences? The norms around privacy in the intimate partner context differ substantially from the parent-child case. In the intimate relationships, both people involved are adults with autonomy and reasonable expectation of privacy, and many of the suggested techniques seem to overstep social and legal norms among adults (especially without consent). Meanwhile, by biological, social, and legal norms, parents are responsible for the care of their children. So techniques for parental controls and surveillance fall within the norms for parenting, even if individual parents would disagree on the balance between control vs. autonomy, and safety vs. privacy. Similarly, teenage children rebelling against parents is well within social norms, even if done in secret.

7.2 Gender

We observed that TikTok creators framed their videos from a feminized and heteronormative perspective. The videos we collected predominantly used feminine language and were targeted to a feminine audience. Given the limitations of our method, which is observational about TikTok videos, we refrain from assuming the gender identities of creators. Instead, we qualitatively discuss the *feminine* (as opposed to *masculine*) coding of the video content, in alignment with scholarship on gender performativity [13] and in particular, gendered language (e.g., [27, 44]).

Specifically, we observed that many creators in the intimate partner context used feminized language towards *themselves*, e.g., #crazygirlfriend, “she’s back,” and masculinized language to describe the *targets* of their strategies, e.g., “the boys aren’t gonna like what I’m about to share with you” (TT23). Additional videos presumed the audience to be women in relationships with men: “ladies, the goal here is to manipulate the algorithm, sorta like the way men manipulate us” (TT39).

In the parent-child context, most creators used feminized language when referring to themselves, e.g., #momhack. One creator described using AirTags to track her daughter’s location on the weekends when her ex-husband had custody of the daughter. Many implicitly associated their motherhood with the role of ensuring their children’s safety, calling for other mothers (and not fathers) to follow their advice.

Why Feminine-Coded? We propose two explanations: First, society prescribes gendered dynamics for the relationships in which these tutorials exist (romantic relationships, parenting). Historical gender roles place significant burdens on women to do emotional labor in sustaining heterosexual relationships and to compromise or make behavioral changes whenever relationship issues arise [64]. Similarly, childcare and other domestic labor typically falls on mothers [32]. Further, the predominant motivations in these interpersonal contexts were to prevent cheating and ensure child safety, implying that if women did not carry out their gendered responsibilities, negative consequences should be blamed on the women (instead of on the men or children also in these relationships) or that men default to infidelity and children to danger.

Second, there could be selection bias in our data collection. It is possible that our search keywords or hashtags were somehow biased to mainly find videos containing gendered language or performative displays associated with women. However, even when we returned to data collection to find more videos containing gendered language or performative displays associated with men — to triangulate (see Section 4.1) this finding — we were not successful in surfacing them.

7.3 TikTok Culture

The aesthetics and substance of the videos in our dataset are strongly shaped by TikTok’s attention economy dynamics: there is significant pressure to make viral content, optimized for TikTok’s recommendation system.

Strong Emotional Appeals. The creators in our dataset tend to make the stakes or potential outcome of listening to their video clear from the very start of the video. On TikTok, getting to the next piece of content only takes one quick swipe, so creators very often say or show something engaging in the first few seconds of a video, e.g., “Think he’s a cheater? I got u girlie” (TT6) or “PROTECT YOUR CHILDREN!!! ALWAYS WATCH THEIR LOCATION!” (TT65).

Controversial Content. Another established way to increase popularity is to be controversial, and indeed, the very nature

of anti-privacy and anti-security advice is controversial. This can be seen in the comments to videos we studied, where some disagreed with the creator, e.g., “not good in any way, this is super toxic” (comment to TT3) or otherwise passed judgement: “say you’re controlling and have low self-esteem without actually saying it” (comment to TT5).

Multi-Modal Content. On TikTok broadly, as well as within the videos in our dataset, content is intensely multi-modal. Videos often have music and captions that support the overall message of the video, as well as concurrent audio speech and text overlaid on the screen. Anti-privacy and anti-security advice videos further contained screenshots and screen recordings, overlaid with annotations. This means that a viewer needs to take in multiple streams of content at once, sometimes watching the video multiple times to catch everything.

Subcommunities. Creators and influencers seek to cultivate a unique (and large) audience, which can lead to the development of subcommunities. For example, the creator of one series began the videos with, “Welcome to [name of video series]”, asserting that the viewer had entered an established digital space. In another video, a creator referred to populations of their viewers: “junior toxics” who needed to learn from “senior toxics” about the “toxicity basics,” because after all, the senior toxics had a “legacy to uphold.” Unlike structured communities on platforms like Reddit or Facebook, TikTok subcommunities exist fluidly and organically, using the same hashtags, commenting on videos, and responding to each other (e.g., in the forms of TikTok “stitches” or “duets”).

8 Discussion and Conclusion

Our work sheds light on a part of TikTok where creators give anti-privacy and anti-security advice around surveillance and control in interpersonal relationships. We believe that studying, documenting, and describing how people use (or misuse) technology today, and exploring ecosystems like the ones we see here within TikTok, is intrinsically interesting and valuable. We also draw from our findings concrete implications for security and privacy research and practice.

8.1 Implications and Recommendations

The surveillance and control techniques used by stakeholders in our case studies show ways that existing solutions are insufficient for preventing harm. What can or should be done?

Designing for strong interpersonal adversaries with physical access. Our work provides additional evidence and concrete examples of how adversaries with physical access to devices are a realistic threat for regular people, occurring commonly in both contexts we studied. Threat models should take physical access seriously for assets like location and communications privacy — these are not just at risk for people who expect to be targeted by (for example) intelligence agencies.

To raise the bar for attacks relying on physical access, apps and operating systems could require additional authentication

at privacy and security sensitive points, such as for data downloads. But while such mitigations may make some attacks more difficult — e.g., preventing “casual” or opportunistic surveillance — they do not address cases where interpersonal control or access goes further. For instance, password sharing is common in romantic relationships [46]. In more opportunistic surveillance contexts, audit logs may be helpful to surface unexpected activity, but in more extreme intimate partner abuse situations, the situation is likely more complex. As other work studying intimate partner surveillance has discussed as well, novel and thoughtful approaches are required.

Mitigating risks of location tracking hardware. Our work surfaces examples of real users openly discussing (surprisingly openly, to us) the abuse of location tracking hardware like AirTags to non-consensually track peoples’ location. Though Apple has implemented some protections, including playing audible alerts if an AirTag has followed you for too long, our data and other anecdotes suggest that these mitigations are insufficient. As of early 2022, Apple is designing modifications to make AirTags louder and improve the alerting system for unrecognized AirTags [11]. Is it possible to develop technologies or policies that prevent the use case of tracking individuals at all?

Anticipating deeply personal motivations. We note that the motivations for the surveillance and control techniques we see in our data are deeply personal and emotional (and common): romantic partners worried about their partners cheating, parents worried about their childrens’ safety, and children wishing to assert their independence. The underlying social phenomena motivating people to “hack” others are thus unlikely to go away. Developers of any apps or hardware used in these interpersonal contexts must consider how their product might be used or misused for these reasons. Our work complements other work which seeks to draw attention to these motivations and challenges [37, 57, 59].

Monitoring TikTok by researchers and developers. Given the popularity and openness with which we found anti-security advice on TikTok, continued monitoring of TikTok for these topics (including comments left on these videos, which we did not investigate) might be useful for those researching or providing support to victims of intimate partner surveillance, as well as to the companies whose technologies are being potentially misused or exploited. Future research could also evaluate the risks posed by the advised techniques.

Managing problematic viral content. Finally, we draw attention to the potential for TikTok to virally spread anti-privacy and anti-security advice to large audiences. Unlike in other contexts, like forums discussing how to do intimate partner surveillance [59], the nature of TikTok is such that its users may not be searching for specific content but rather receive content pushed to their feeds by TikTok’s recommendation algorithm. And unlike ethical security vulnerability reports, these videos explicitly suggest exploiting vulnerabilities to

violate the security and privacy of others (especially in the intimate partner context).

Thus, we must consider TikTok’s role in moderating, recommending, and perhaps limiting the spread of this type of content. TikTok’s community guidelines already forbid videos from providing instructions on how to conduct illegal activity [58], which may apply to some of the videos in our dataset. Even for content that should not directly be prohibited, there may be a role for TikTok to display additional information (e.g., pointers to resources for all parties in interpersonal relationships), similar to misinformation-related notices on social media platforms. Whether and how such notices should be designed to be helpful is a question for future work.

8.2 TikTok as a Qualitative Data Source

Benefits. Our work demonstrates how TikTok can be used as an alternative source of qualitative, observational data for security and privacy-related topics, especially in contexts where traditional usable security methods such as interviews and surveys might be challenging to recruit for or conduct. For instance, recruiting and asking people to discuss the techniques they use to surveil or control intimate partners may not have surfaced as rich results due to social desirability bias. TikTok’s user and creator base also has different demographics (e.g., skewing younger) than other social media platforms commonly studied in research (e.g., Twitter, Reddit) [14].

TikTok videos contain rich information in a short video: individual videos in our dataset often contained a multi-modal combination of video of the creator, speech, music, or other audio, text overlaid on the video, and screenshots or screen recordings. Additional context is provided through the video’s caption, which often includes hashtags.

Challenges. A major challenge we faced was identifying relevant TikTok videos to study. The utility of text-based search is limited, and the emergence of different subcommunities on the platform (e.g., “toxics”) meant that we had to discover specific terminology to find additional relevant videos.

We also could not easily investigate TikTok’s features for remixing and responding to content. Creators can “duet” videos by adding their own video to an existing one, or “stitch” videos by clipping and integrating clips into their own video. Unfortunately for our data collection, TikTok’s platform does not offer a feature to find all duets and stitches.

Future work. This paper has just scratched the surface of the types of security and privacy questions that we might investigate via TikTok content. For example, future work might investigate *pro*-security advice on TikTok. Anecdotally, we have also observed rich content on the topic of “sharenting”. There may also be other sub-communities of interest, such as people conducting more technically sophisticated exploits.

Acknowledgments

We thank our reviewers for their helpful feedback. We are grateful for the many insights of Chris Geeng 🤗, Kentrell Owens 🤗, Tina Yeung 🤗, Sudheesh Singanamalla 🤗, and Os Keyes ❤️ during this research. We thank Kaiming Cheng 🙏 for assisting with the screenshots. This work was supported in part by the U.S. National Science Foundation under Awards CNS-1565252 and CNS-2114230, and by a gift from Google.

References

- [1] Crystal Abidin. Mapping Internet celebrity on TikTok: Exploring attention economies and visibility labours. *Cultural Science Journal*, 12(1):77–103, 2021.
- [2] Corey H. Basch, Grace C. Hillyer, and Christie Jaime. COVID-19 on TikTok: Harnessing an Emerging Social Media Platform to Convey Important Public Health Messages. *International Journal of Adolescent Medicine and Health*, 2020.
- [3] Alicia Blum-Ross and Sonia Livingstone. “Sharenting,” parent blogging, and the boundaries of the digital self. *Popular Communication*, 15(2):110–125, 2017.
- [4] Dannell D. Boatman, Susan Eason, Mary Ellen Conn, and Stephenie K. Kennedy-Rea. Human Papillomavirus Vaccine Messaging on TikTok: Social Media Content Analysis. *Health Promotion Practice*, page 15248399211013002, 2021.
- [5] Maia J. Boyd, Jamar L. Sullivan Jr., Marshini Chetty, and Blase Ur. Understanding the Security and Privacy Advice Given to Black Lives Matter Protesters. In *ACM Conference on Human Factors in Computing Systems*, CHI ’21, pages 1–18, 2021.
- [6] Virginia Braun and Victoria Clarke. *Successful Qualitative Research: A Practical Guide for Beginners*. SAGE, 2013.
- [7] Virginia Braun and Victoria Clarke. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, 21(1):37–47, 2021.
- [8] Virginia Braun and Victoria Clarke. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, 2021.
- [9] Anna Brosch. When the child is born into the Internet: Sharenting as a growing trend among parents on Facebook. 2016.
- [10] Anna Brosch. Sharenting: Why do parents violate their children’s privacy? 2018.
- [11] Kellen Browning. Apple says it will make airtags easier to find after complaints of stalking. <https://www.nytimes.com/2022/02/10/business/apple-airtags-safety.html>, 2022.
- [12] Amber M. Buck and Devon F. Ralston. I didn’t sign up for your research study: The ethics of using “public” data. *Computers and Composition*, 61:102655, 2021.
- [13] Judith Butler. *Gender Trouble*. Routledge, 1990.
- [14] Pew Research Center. Social media fact sheet. <https://www.pewresearch.org/internet/fact-sheet/social-media/>, 2021.
- [15] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, , and Thomas Ristenpart. The Spyware Used in Intimate Partner Violence. In *IEEE Symposium on Security and Privacy*, SP ’18. IEEE, 2018.
- [16] Camille Cobb and Tadayoshi Kohno. How Public Is My Private Life? Privacy in Online Dating. In *The World Wide Web Conference*, WWW ’17, pages 1231–1240, 2017.
- [17] Camille Cobb, Lucy Simko, Tadayoshi Kohno, and Alexis Hiniker. A Privacy-Focused Systematic Analysis of Online Status Indicators. *PoPETS*, 2020(3):384–403, 2020.
- [18] Daniel Le Compte and Daniel Klug. Poster: “It’s Viral!” A Study of the Behaviors, Practices, and Motivations of TikTok Users and Social Activism. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW ’21, pages 108–111, 2021.
- [19] Lorrie Faith Cranor, Adam L. Durity, Abigail Marsh, and Blase Ur. Parents’ and Teens’ Perspectives on Privacy In a Technology-Filled World. In *Symposium on Usable Privacy and Security*, SOUPS ’14, pages 19–35, 2014.
- [20] Alexei Czeskis, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. Parenting from the Pocket: Value Tensions and Technical Directions for Secure and Private Parent-Teen Mobile Safety. In *Symposium on Usable Privacy and Security*, SOUPS ’10, pages 1–15, 2010.
- [21] Duy Dang-Pham, Siddhi Pittayachawan, and Vince Bruno. Impacts of Security Climate on Employees’ Sharing of Security Advice and Troubleshooting: Empirical Networks. *Business Horizons*, 59(6):571–584, 2016.
- [22] Duy Dang-Pham, Siddhi Pittayachawan, and Vince Bruno. Why Employees Share Information Security Advice? Exploring the Contributing Factors and Structural Patterns of Security Advice Sharing in the Workplace. *Computers in Human Behavior*, 67:196–206, 2017.
- [23] Sauvik Das, Laura A. Dabbish, and Jason I. Hong. A Typology of Perceived Triggers for End-User Security and Privacy Behaviors. In *Symposium on Usable Privacy and Security*, SOUPS ’19, pages 97–115, 2019.

- [24] Jared Duval, Ferran Altarriba Bertran, Siying Chen, Melissa Chu, Divya Subramonian, Austin Wang, Geoffrey Xiang, Sri Kurniawan, and Katherine Isbister. Chasing Play on TikTok from Populations with Disabilities to Inspire Playful and Inclusive Technology Design. In *ACM Conference on Human Factors in Computing Systems*, CHI '21, pages 1–15, 2021.
- [25] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders. In *CSCW*. ACM, 2017.
- [26] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. “A Stalker’s Paradise”: How Intimate Partner Abusers Exploit Technology. In *ACM Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, 2018.
- [27] Danielle Gaucher, Justin Friesen, and Aaron C. Kay. Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. *Journal of Personality and Social Psychology*, 101(1):109, 2011.
- [28] Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. “Like Lesbians Walking the Perimeter”: Experiences of U.S. LGBTQ+ Folks With Online Security, Safety, and Privacy Advice. In *USENIX Security Symposium*, 2022.
- [29] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr., and Pamela J. Wisniewski. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control.
- [30] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical Computer Security for Victims of Intimate Partner Violence. In *USENIX Security Symposium*, 2019.
- [31] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlene Fernandes, and Blase Ur. Rethinking Access Control and Authentication for the Home Internet of Things (IoT). In *USENIX Security*, pages 255–272, 2018.
- [32] Bell Hooks. *Feminism Is for Everybody: Passionate Politics*. Pluto Press, 2000.
- [33] Mansoor Iqbal. TikTok Revenue and Usage Statistics. <https://www.businessofapps.com/data/tik-tok-statistics/>, 2022.
- [34] D. Bondy Valdovinos Kaye, Aleesha Rodriguez, Katrin Langton, and Patrik Wikstrom. You Made This? I Made This: Practices of Authorship and (Mis) Attribution on TikTok. *International Journal of Communication*, 15:3195–3215, 2021.
- [35] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2, November 2018.
- [36] Tama Leaver. Intimate Surveillance: Normalizing Parental Monitoring and Mediation of Infants Online. *Social Media & Society*, 3(2), 2017.
- [37] Karen Levy and Bruce Schneier. Privacy Threats in Intimate Relationships. *Journal of Cybersecurity*, pages 1–13, 2020.
- [38] Karen E.C. Levy. Intimate Surveillance. *Idaho Law Review*, 51:679, 2014.
- [39] Tom De Leyn, Ralf De Wolf, Mariek Vanden Abeele, and De Lieven Marez. In-Between Child’s Play and Teenage Pop Culture: Tweens, TikTok & Privacy. *Journal of Youth Studies*, pages 1–18, 2021.
- [40] Yachao Li, Mengfei Guan, Paige Hammond, and Lane E. Berrey. Communicating COVID-19 Information on TikTok: A Content Analysis of TikTok Videos From Official Accounts Featured in the COVID-19 Information Hub. *Health education research*, 36(3):261–271, 2021.
- [41] Johnny Long. *No Tech Hacking: A Guide to Social Engineering, Dumpster Diving, and Shoulder Surfing*. Syngress, 2008.
- [42] Deborah Lupton and Ben Williamson. The Datafied Child: The Dataveillance of Children and Implications For Their Rights. *New Media & Society*, 19(5):780–794, 2017.
- [43] Kim Lyons. Tiktok says it has passed 1 billion users. <https://www.theverge.com/2021/9/27/22696281/tiktok-1-billion-users>, 2021.
- [44] Michael A. Messner, Margaret Carlisle Duncan, and Kerry Jensen. Separating The Men From The Girls: The Gendered Language of Televised Sports. *Gender & Society*, 7(1):121–137, 1993.
- [45] David L. Morgan and Andreea Nica. Iterative Thematic Inquiry: A New Method For Analyzing Qualitative Data. *International Journal of Qualitative Methods*, 19, 2020.
- [46] Cheul Young Park, Cori Faklaris, Siyan Zhao, Alex Sciuto, Laura Dabbish, and Jason Hong. Share and Share Alike? An Exploration of Secure Behaviors in Romantic Relationships. In *Symposium on Usable Privacy and Security*, SOUPS '18, 2018.
- [47] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How I Learned To Be Secure: A Census-Representative Survey of Security Advice Sources and Behavior. In *ACM Conference on Computer and Communications Security*, CCS '16, pages 666–677, 2016.
- [48] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. I Think They’re Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *IEEE Symposium on Security and Privacy*, SP '16, pages 272–288. IEEE, 2016.
- [49] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock

- Stevens, and Michelle L. Mazurek. A Comprehensive Quality Evaluation of Security and Privacy Advice on the Web. In *USENIX Security*, pages 89–108, August 2020.
- [50] Robert W. Reeder, Iulia Ion, and Sunny Consolvo. 152 Simple Steps to Stay Safe Online: Security Advice For Non-Tech-Savvy Users. *IEEE Symposium on Security and Privacy*, 15(5):55–64, 2017.
- [51] Kevin Roundy, Paula Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, and Acar Tamersoy. The Many Kinds of Creepware Used for Interpersonal Attacks. In *IEEE Security and Privacy*, SP '20, pages 626–643. IEEE, 2020.
- [52] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok. In *ACM Conference on Web Science*, pages 257–266, 2020.
- [53] Clare Southerton. Lip-Syncing and Saving Lives: Healthcare Workers on TikTok. *International Journal of Communication*, 15, 2021.
- [54] Robert E. Stake. *The Art of Case Study Research*. SAGE, 1995.
- [55] Valerie Steeves and Owain Jones. Surveillance, Children and Childhood. *Surveillance & Society*, 7(3/4):187–191, 2010.
- [56] Marit Sukk and Andra Siibak. Caring Dataveillance and the Construction of “Good Parenting”: Estonian Parents’ and Pre-teens’ Reflections on the Use of Tracking Technologies. *Communications*, 46(3):446–467, 2021.
- [57] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *IEEE Symposium on Security and Privacy*, SP '21, pages 247–267. IEEE, 2021.
- [58] TikTok. Community guidelines. <https://www.tiktok.com/community-guidelines>, 2022.
- [59] Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infidelity Forums. In *USENIX Security*, 2020.
- [60] Emily Tseng, Diana Freed, Kristen Engel, Thomas Ristenpart, and Nicola Dell. A Digital Safety Dilemma: Analysis of Remote Computer-Mediated Computer Security Interventions During COVID-19. In *ACM Conference on Human Factors in Computing Systems*, CHI '21, pages 1–17, 2021.
- [61] Blase Ur, Jaeyeon Jung, and Stuart Schechter. Intruders Versus Intrusiveness: Teens’ and Parents’ Perspectives on Home-Entryway Surveillance. In *UbiComp*, pages 129–139, 2014.
- [62] Kandrea Wade, Jed R. Brubaker, and Casey Fiesler. Protest Privacy Recommendations: An Analysis of Digital Surveillance Circumvention Advice During Black Lives Matter Protests. In *Extended Abstracts of the Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–6, 2021.
- [63] Ge Wang, Jun Zhao, Max Can Kleek, and Nigel Shadbolt. Protection or Punishment? Relating the Design Space of Parental Control Apps and Perceptions About Them to Support Parenting for Online Safety. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW):1–26, 2021.
- [64] Jane Ward. *The Tragedy of Heterosexuality*. New York University Press, 2020.
- [65] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M. Carroll. Parental Control vs. Teen Self-Regulation: Is There a Middle Ground for Mobile Online Safety? In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 51–69, 2017.
- [66] Jing Zeng, Mike S. Schäfer, and Joachim Allgaier. Reposting “Till Albert Einstein is TikTok Famous”: The Memetic Construction of Science on TikTok. *International Journal of Communication*, 15:3216–3247, 2020.
- [67] Leah Zhang-Kennedy, Christine Mekhail, Yomna Abdelaziz, and Sonia Chiasson. From Nosy Little Brothers to Stranger-Danger: Children and Parents’ Perception of Mobile Threats. In *International Conference on Interaction Design and Children*, pages 388–399, 2016.
- [68] Yixin Zou, Allison McDonald, Julia Narakornpichit, Nicola Dell, Thomas Ristenpart, Kevin Roundy, Florian Schaub, and Acar Tamersoy. The Role of Computer Security Customer Support in Helping Survivors of Intimate Partner Violence. In *USENIX Security*, 2021.

A Summary Table

Table 2: A summary of all of the motivations, goals and techniques we observed in our dataset, across two interpersonal contexts: intimate partner relationships and parent-child relationships. We identify what goals were sought for what motivations, with which techniques.

	Goal (what?)	Motivation (why?)	Techniques (how?)
Intimate Partner Context	<i>Instigator Perspective</i>		
	Surveil digital communications	Detect cheating	Use data downloads to obtain message history (and other metadata) Check recently used emojis for sexually explicit emojis Takeover Snapchat account with 2FA vulnerability
	Stalk on social media	Detect cheating	Find public conversations between target and suspected affair partner
	Surveil dating app usage	Detect cheating	Use 3rd party site to see if on dating app See if email address already exists on dating app Create fake account to see if on dating app
	Surveil other digital activities	Detect cheating	Look at photo metadata to determine when it was originally taken Get physical access to data on phone: explicit photos, vault apps that could hide explicit photos, porn websites in browsing history, dating apps, emails from hookup sites
	Surveil physical world	Detect cheating	Use AirTags/AirPods to track target's location Use monitoring apps (Life360) Get physical access to view location on phone or in accounts (Google Maps, iOS Significant Locations)
	Stalk on social media	Arbitrary surveillance	Abuse accessibility features to listen (Live Listen, auto-answer calls) Use 3rd party site to anonymously view target's Instagram stories or display photo See order of who target recently followed on Instagram website Use app to detect when target is signing on/off WhatsApp Use app to see searched/clicked/viewed your Instagram Create fake account to view Instagram story
Manipulate social media	Exert control	Keep track of Snapchat score to see if mass sending Restrict account on Instagram, sends DM to message requests to evade read receipts and get more time to respond Change phone time to delete previously sent WhatsApp message Create fake tag in Instagram story using poll feature and see who clicks Hide and unhide story so instigator's Instagram story appears first	
Text someone who blocked you	Exert control	Message from email (does not work)	
Parent-Child Context	<i>Target Perspective</i>		
	Detect call surveillance	Evade surveillance	Check carrier settings for call forwarding or redirection
	<i>Parent Perspective</i>		
	Surveil physical world	Child safety	Hide AirTag in bag, clothing, or car Give AirTag bracelet or keychain Install tracking app (Life360)
	Surveil digital world	Child safety	Sync iCloud messages Use text forwarding
	Restrict content and usage	Exert control	Locked down smartphone Parental control apps (Bark, FamiSafe)
	<i>Child Perspective</i>		
Evade location tracking app	Location privacy	Disable app tracking cellular data permissions Put phone on Do Not Disturb Install app on another device	
Evade digital surveillance	Device privacy	Hide home screen pages	
Evade parental controls	Autonomy	Brute force passcode by detecting fingerprints on screen Use different VPN Sign out of app store and use new Apple ID	

“Fast, Easy, Convenient.” Studying Adoption and Perception of Digital Covid Certificates

Franziska Herbert
Ruhr University Bochum

Marvin Kowalewski
Ruhr University Bochum

Theodor Schnitzler 
Ruhr University Bochum

Leona Lassak
Ruhr University Bochum

Markus Dürmuth 
Leibniz University Hannover

Abstract

Digital vaccination, recovery, and test certificates play an important role in enforcing access restrictions to certain parts of the public life in Europe during the current phase of the COVID-19 pandemic. Such certificates represent an interesting showcase for digital security and privacy in the context of sensitive personal data.

In this paper, we take a look at which types of certificates and related apps people in Germany use for which purposes, which factors influence their adoption, and which misconceptions exist concerning the security and use of certificates. To this end, we report the results of a census-representative online survey in Germany ($n = 800$) conducted in December 2021, complemented with 30 qualitative street interviews.

Most participants favor digital certificates over paper-based variants due to their ease of use and seamless integration into dedicated smartphone apps – more than 75 % of participants have installed one or more eligible app(s) on their phone. We find that older age, higher privacy concerns related to apps, and not being vaccinated are factors hindering the adoption of digital certificates.

1 Introduction

Over the past two years, the COVID-19 pandemic has caused massive restrictions to many aspects of public life all around the world, including lockdowns, curfews, cancellation of public events, limitations of international travel, and many more [31]. The broader availability of vaccines against COVID-19, especially in many countries in the Americas,

Asia, and Europe starting in 2021 [32], allowed gradual releases of several restrictions and a prospective return to normality. Since vaccinations have been shown to be very effective in preventing severe COVID-19 diseases [29, 36], many restrictions were particularly released for people who are fully vaccinated or have recovered from COVID-19. Some restrictions were also eased for people who tested negative for coronavirus.

In many European countries, e. g., Germany or Italy [13, 42] but also in Israel and some US states [14, 27], people have to prove their vaccination or recovery status, or provide a negative test result in order to attend certain public events or activities. Such requirements have become a catalyst for the development of *digital covid certificates* i. e., apps that can be used to prove the required status. Israel was one of the first countries to introduce the so-called *Green Pass* app in February 2021 [18], a QR code-based certificate scheme granting access to different activities. In the US, the state of New York has also introduced a digital QR code-based certificate in March 2021 (*NYS Excelsior Pass* app and *NYS Excelsior Pass Scanner* app) that serves as a proof of vaccination or alternatively proof of a negative coronavirus test [16]. In California, a similar digital vaccination certificate was introduced in August 2021, also enabling citizens to prove their vaccination [26].

One of the most widely deployed schemes is the EU Digital COVID Certificate, introduced by the European Union in June 2021 [10]. The underlying framework allows for interoperability of nationally issued certificates across all 27 EU member countries serving up to 450 million inhabitants. Similar to other systems, the EU certificate can be shown using a QR code and can be integrated into several dedicated mobile apps (e. g., CovPass and Corona-Warn-App in Germany). The certificate contains personal information such as name and date of birth, specifics of the vaccination, recovery, or test result (whichever applies depending on the type of certificate), and digital signatures for technical verification purposes. For *verifying the correctness* of certificates, specific apps (e. g., CovPassCheck in Germany) were introduced. These apps

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

validate electronic signatures of certificates and display the name and date of birth of the person for matching these data with ID cards.

In this work, we present the results of a census-representative online survey in Germany (n=800) on the use and perception but also the knowledge and potential misconceptions of digital covid certificates. Moreover, we are interested to learn the experiences of the respondents with this process, and adherence to correct verification in practice. We complemented our online survey with 30 street interviews with people who were obliged to verify certificates for access restriction purposes, e. g., shop owners, or restaurant staff. Finally, we observed the verification process of 80 businesses with access restrictions in the wild through convenience sampling. All three surveys were conducted in December 2021.

We find 70% of participants using apps to indicate their vaccination, recovery or test status, mostly for convenience reasons and due to the ease of use of certificate apps. Reasons against using digital certificates are for example privacy concerns and security concerns, and apps are less prevalent among participants at older ages.

Digital covid certificates provide an interesting showcase for digital security and privacy in an everyday application that is widely used by a broad audience: they contain not only personally identifiable information such as name and date of birth but also sensitive health information, i. e., vaccination, recovery, or test status. To the best of our knowledge, we are the first to study use, perception, and verification of digital covid certificates while being in wide-spread use.

In summary, our work makes the following contributions:

- We shed light on the prevalence of and attitudes towards paper-based and app-based covid certificates in a phase in which they highly facilitate participation in public life in Germany.
- Our quantitative and qualitative evaluations show that ease of use is a highly significant factor for the adoption of digital covid certificates, suggesting that easy to use solutions are desirable.
- We complement results of our consensus-representative online survey with qualitative insights from street interviews and with observations in the wild (random sample).

2 Related Work

Most related to our research is the work by Kowalewski et al. [20], who study the willingness to use of different variants of covid vaccination certificates in hypothetical scenarios prior to the introduction and use of these certificates. They find privacy, prior use of a corona app, and being against a vaccination obligation to be hindering factors for (hypothetical) willingness to use a vaccination certificate. On the other hand they find worries about the coronavirus and vaccination willingness to be

factors positively influencing the (hypothetical) willingness to use a vaccination certificate.

Other studies with regard to apps against the spread of the coronavirus, i. e., contact tracing apps, also find (app related) privacy concerns to have negative influence on the adoption but not on the continued use of these apps [25, 46, 49]. Other factors fostering the adaption of these apps are performance expectancy, social influence, technological knowledge, and apps benefits [25, 49]. The latter two factors are also found important for continued app usage of contact tracing apps [25].

A large body of work investigates the broader role of mobile apps in the pandemic in the contact tracing domain [1, 19, 21, 24, 39, 43, 52]. Individual studies also cover other types of apps for different pandemic-related purposes such as symptom checking [44] or accessing information about the pandemic [53]. A study Utz et al. [46] investigates predictors for the adoption of a broad range of app types, also finding that privacy is a significant factor for adoption.

Those findings are in line with more general related work, finding privacy (concerns) a relevant factor in decision-making about digital tools and interacting with online technology in a broad range of applications [9, 23, 38], as well as when using mobile health apps [15, 51, 54]. Also other factors influencing the use of mobile health apps, like age, education level, and e-health literacy, were identified [3].

General theories on (intention) to use technology, as the technology acceptance model (TAM), TAM2 and the Unified Theory of Acceptance and Use of Technology (UTAUT) [7, 47, 48] find the intention to use technology is based on factors like *perceived ease of use*, *perceived usefulness*, *social influence processes* (e. g., subjective norm) as well as *performance expectancy* (based on perceived usefulness and others) and *effort expectancy*.

3 Study Context

Digital COVID-19 certificates [12] have been introduced to establish a standardized and securely verifiable alternative to paper-based documents, such as the internationally recognized *yellow certificate of vaccination* document standardized by the *World Health Organization* (WHO) [8]. While such digital certificates are predominantly applied in Europe, they have also found adoption in other countries such as Israel and certain US states [16, 18, 26, 50].

In this section, we introduce the concept of the EU Digital COVID Certificate, and describe the current state of the pandemic in Germany, particularly focusing on restrictions in public life to provide the context in which we conducted our study. Whenever we refer to *covid certificates*, we include proofs for being vaccinated against COVID-19, recovered from COVID-19, or having a negative COVID-19 test result.

3.1 EU Digital COVID Certificate

In the European Union (EU), a digital covid certificate framework was rolled out starting in June 2021. It uses a QR code-based system and contains a cryptographic signature to protect against misuse or forgery. The certificate proves that a person

- is fully vaccinated against COVID-19,
- recovered from COVID-19,
- or has tested negative [11].

The information contained in the certificate includes personal data (e. g., name and date of birth), information on the vaccine (e. g., type and date) and technical details (e. g., certificate issuer, expiration date, and a unique identifier) [30]. In Germany these certificates can be included in three different apps: the *CovPass* app, the *Corona-Warn-App* (CWA), and the *Luca* app. The CovPass app was specifically developed for this purpose. The other two apps were introduced before for contact tracing (CWA) and event registration (Luca) and included the digital certificate as a new feature [5, 33, 37]. For privacy reasons, only the QR code and the person’s name are displayed within the app when the certificate is presented to a third party, e. g., for verification purposes (see Figure 1a).

To correctly verify the digital covid certificate, a so-called *verifier* app, such as the *CovPassCheck* app in Germany, is needed. Within the verification process, the verifying party uses this app to scan the QR code of the covid certificate [34]. For privacy reasons, the verifying person only sees whether or not the certificate is valid, along with name and DOB of the person to be verified (cf. Figure 1b), which have to be compared with an ID document. Whatever information is additionally shown on the device of the person to be verified (e. g., green bars, check marks, etc.) is irrelevant for correctly completing the verification process. Since the digital covid certificate contains more personal information like vaccination date(s), vaccine type, or recovery status, which is why letting another person scroll through the app is not advised due to privacy reasons.

In order to raise awareness of digital covid certificates, the German government provided a website explaining in detail how the EU digital covid certificate works and how to verify it correctly [35]. Governmental advertising campaigns on television and social media also drew attention to the digital covid certificate and how to store them in either of the two government-backed apps, i. e., CWA and CovPass. Within both apps, additional information was given to explain the correct verification of the QR code, i. e., using the *CovPassCheck* app.

3.2 Pandemic Situation in Germany

Due to high infection rates in Germany, measures referred to as *G-rules*¹ were successively introduced beginning in Au-

¹e. g., 3G represents the requirement to be either vaccinated, recovered or tested negative. All German terms start with *g* (*geimpft*, *genesen*, *getestet*).

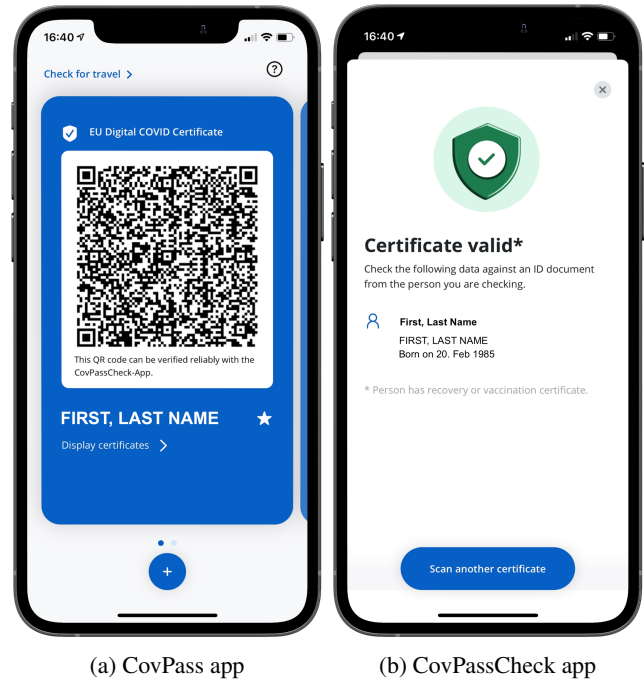


Figure 1: EU Digital COVID Certificate shown in the German CovPass(Check) app.

gust 2021 (see Table 1 for an overview). These measures were applied to certain parts of public life, e. g., to restrict attendance at professional sports events, and define what persons are eligible to access respective events [41]. In most of the 16 German states, both paper-based and digital covid certificates are accepted to prove the respective status, except for four states (i. e., Berlin, Brandenburg, Baden-Württemberg, and Saarland) that required digital covid certificates starting from September 2021 the earliest.

Over the course of the year, G-rules were continuously tightened and applied to attending large (sports) events, staying in hotels, non-essential shopping, using public transport, going to school or work, and others. Germany’s regulation obliged all these venues to verify attendants’ G-statuses. Thus, at the time of our study, certificates were required for all parts of public life except for shopping groceries and other essentials.

Table 1: Explanation of the G rules in Germany.

Access for	Fully vaccinated	Recovered	Negative Rapid Test	Negative PCR Test
3G	✓	✓	✓	✓
3G+	✓	✓		✓
2G	✓	✓		
2G+	✓ ^a	✓ ^a		

^a Additional negative covid test required

4 Method

Parts of our study are based on the work of Kowalewski et al. [20] which allows us to compare their findings about *hypothetical* willingness to use (digital) vaccination apps with the actual use of them in the wild. As one of their hypothetical scenarios (*U3: Certificates required for various aspects of public life, vaccine available for everyone*) actually reflects the current (real) situation in Germany quite well, we will later compare our findings with theirs whenever applicable.

To study user adoption, knowledge, potential misconceptions and verification processes of (digital) covid certificates, we conducted three studies: a census-representative online questionnaire (n=800), short street interviews (n=30) in a city in western Germany, and random samples (n=80) of the verification process of digital covid certificate.

The online survey was conducted between December 03, 2021 and December 09, 2021 with 800 participants, using the software *Qualtrics* and the panel provider *Respondi*. The participants of our online survey received a monetary compensation for taking part in the questionnaire. Respondi handled participant recruitment, compensation and set quotas representative for the German population for gender, age, and education. Unfortunately, people over 70 are rarely represented in online panels. Quotas were matched perfectly for age and education, and there was a maximum deviation of 2% for gender. The education classification is based on UNESCO-ISCED Levels: Low (0-2), medium (3-4), and high (5-8) [45]. We list our participants' demographics in Table 2.

The 30 street interviews were conducted in one Germany city from December 07 to 21, 2021 by three researchers shortly after (digital) covid certificates became mandatory for many parts of public life (e. g., going to the cinema, restaurants, clothing stores or the hairdresser).

Our random sampling of the certificates' verification process was conducted in the same region as the interviews from December 07 to 21, 2021. Details on the three studies are presented in the following paragraphs.

4.1 Online Survey

We designed our online survey to gain insights into user preferences, perception and motivations for the use of digital and paper-based covid certificates. We focus on the most common certificate forms available in Germany:

- yellow certificate of vaccination (paper-based)
- Corona-Warn-App (digital)
- CovPass app (digital)
- Luca app (digital)

Whereas the CovPass app was specifically developed for handling covid certificates, the other two apps had already been available and were primarily used for contact tracing (CWA) and event registration (Luca) before.

Table 2: Participant Demographics in Online Survey

	Participants		Target
<i>Gender</i>			
Female	404	(50.5 %)	49 %
Male	392	(49.0 %)	51 %
Non-binary	1	(0.1 %)	0 %
Self	3	(0.4 %)	0 %
<i>Age</i>			
18–29	160	(20 %)	20 %
30–39	152	(19 %)	19 %
40–49	144	(18 %)	18 %
50–59	192	(24 %)	24 %
60–69	152	(19 %)	19 %
<i>Education^a</i>			
Low (ISCED 0-2)	230	(29 %)	29 %
Medium (ISCED 3-4)	264	(33 %)	33 %
High (ISCED 5-8)	304	(38 %)	38 %
<i>Privacy Disposition</i>			
Mean (SD)	3.28	(0.79)	
<i>App Privacy</i>			
Mean (SD)	2.69	(1.19)	

^aEducation classification is based on UNESCO ISCED 2011 Levels [45].

4.1.1 Questionnaire

In this section, we outline the structure of our questionnaire. Due to our focus on digital covid certificates, we did not analyze all questions of the questionnaire for this paper. We will only address the questions we analyzed for this paper in this section. A complete version of the questionnaire can be found in Appendix A. All questions in the questionnaire were originally formulated in German to avoid misunderstandings. For documentation in this paper, all questions were translated to English.

General Questions and Experiences with the Coronavirus

We asked participants whether they or someone close to them has already been infected with the coronavirus (Q3–Q4), as well as their concerns of getting infected themselves (Q5) or that someone close to them might get infected (Q6).

Covid Certificates We also asked participants which COVID-19 related apps they have installed on their smartphone (Q7). Questions Q8 and Q9 list various items (i. e., purposes and activities) that may qualify for restrictions under COVID-19 measures. We selected items following related work [20] and extended the set with purposes and activities that were subject to (partially controversial) public discussions in Germany. For each item, we asked

1. whether a restriction and which restriction should apply (Q8; none, 3G, 3G+, 2G, or 2G+) and
2. which type of certificate should be required (Q9; none, paper-based, or app-based).

Moreover, we asked participants if they were already required to show their covid certificate (Q10), how effortful they perceived this (Q11), and whether they have already been vaccinated against or recovered from the coronavirus (Q12). Subsequently, we showed them a list of covid certificates (e. g., CovPass app or yellow certificate of vaccination) to indicate which of these variants they typically use to prove their vaccination, recovery or test status (Q13/Q14). Based on whether the participants have indicated to use a paper-based or digital covid certificate (Q13/Q14), we further asked them to explain their decision, i. e., deciding for or against the respective certificate variant (Q15/Q16).

Certificate Verification Process To get insights into the verification process, we asked participants to describe how their digital certificate was verified (Q17) at their latest access control situation(s). In question Q21, the participants were asked to indicate the perceived ease of use of the used covid certificate variant. To evaluate participants' knowledge of digital covid certificates, especially QR codes and the correct verification process, question Q23 consisted of various correct and false statements related to this topic. For the analysis we re-coded the answers to the false questions to compute a *knowledge score*, for which higher values indicate more knowledge. Cronbach's alpha is acceptable for the knowledge score ($\alpha = .7$), which is why we include this score in our analysis.

In question Q24, we asked participants to name the most important aspects for verifying digital certificates correctly, including what they think needs to be verified and what they perceive to be the (technical) security indicators.

Attitudes Towards Measures Against the Spread of the Pandemic In order to understand more about our participants' perceptions regarding measures taken against the spread of the pandemic, we asked them questions whether they believed that specific measures (e. g., the 3G rule, vaccinations, or contact restrictions) contributed to containing the spread of the coronavirus (Q33). Cronbach's alpha shows a good fit for these attitude items, which is why we used them as an *attitude scale* ($\alpha = .9$).

Privacy Disposition The individual vaccination but also the recovery, and test status represent personal health data. Storing this data within an app and linking it with personal information (e. g., name and date of birth) but also providing this data during a mandatory verification process may raise privacy concerns and questions regarding general data pro-

tection. To get insights into participants' privacy attitudes we used two validated three-item *Privacy Scales* [4, 22]:

1. The first part consisted of three questions to measure participants' general privacy disposition (Q34).
2. We adopted the second set of questions to our digital covid app context (Q35).

We added a fourth question to both scales covering specifically concerns related to health data. Therefore, the two scales consist of four questions each. As Cronbach's alpha ranges from acceptable (Q34, $\alpha = .7$) to excellent (Q35, $\alpha = .96$), we use both scales as described.

4.2 Street Interviews

The street interviews expand our research to the views and experiences of people *verifying* covid certificates. We interviewed people working in venues that were obligated to control certain regulations concerning the coronavirus. We spread our interviews across a variety of business sectors. 18 interviews were conducted in retail, i. e., clothing or cosmetics stores, seven were done in hotel and catering business, three in the fitness and health field, and two in cinemas. All of them required 2G rules at that time. We renounced asking demographics to keep interviews as short as possible and to protect participants' privacy. After getting the agreement to participate we asked questions regarding

- the current regulations concerning the coronavirus at the venue
- which restrictions they had to control
- how they verified covid certificates
- how thoroughly they think they performed the verification

Our complete interview included additional questions, which we do not describe here, as we consider them out of scope for this paper. The complete interview guideline including all questions can be found in Appendix B. As the interviews were really short, we refrained from transcribing them. We took notes during the interviews and later grouped them to categories for the analysis.

4.3 Sampling the Verification Process for Digital Certificates

In addition to the perception of our participants and the experiences of people verifying covid certificates, we also wanted to gain insights into how the verification process of digital covid certificates was carried out in the wild. For this purpose, three researchers entered 80 stores and businesses for which the 2G restriction applied, including fashion stores, cinemas, theaters, and restaurants. We did not interact with employees but only observed the verification process. We focused only on the verification process for digital certificates (e. g., Corona-Warn-App or CovPass app) and documented

the complete process distinguishing between the following verification levels:

- L1 No verification
- L2 Short glance (no scan and no ID card required)
- L3 Glance with ID (no scan but matching the personal data with the ID card)
- L4 Scan only (no ID card required)
- L5 Scan with ID (i. e., the correct verification process)

The correct verification of digital covid certificates consists of two factors (level L5): scanning the QR code with a suitable app (e. g., CovPassCheck) and matching the shown personal data with the data of the person’s ID card (i. e., first name, last name, and the date of birth).

4.4 Research Ethics

Our department does not have an institutional review board. Instead, our study followed best practices of human subject research and data protection guidelines. To minimize any potential adverse effects from the study we followed the ethical principles laid out in the Belmont report [28]. Specifically we sought informed consent at the beginning of the study and participants were informed about the topic of our study, data protection, data processing, and pseudonymization of their data, as well as that they could withdraw from the study without any negative consequences at any time. We did also ensure that the panel provider (Respondi) is certified according to ISO 20252:2019, relevant for consumer research.

4.5 Limitations

As Germany is organized federally, not all covid restrictions were identical for all German states. In four states, only digital covid certificates were permitted. However, we believe that the restrictions were similar enough during the time of our study (see 3.2). For our interviews, the small number of interviews and the location restriction to a single city are limitations. The same limitation applies to our random sampled verification process, which was carried out in the same region as the interviews. Additionally, we refrained from collecting demographic data for these two studies. As we used an online panel for our online survey results might tend towards app usage, as online panel works might favor digital tools. Other than that, an online survey will never be able to fully capture the complexity of interacting in real-world situations, which is why we used interviews and convenience sampling as further survey methods. Finally, most of the restrictions we asked about being already in place during the time of our survey might have biased participants to opt for them.

5 Results

In this section, we present the main results of our study, centered around the results of our online survey.

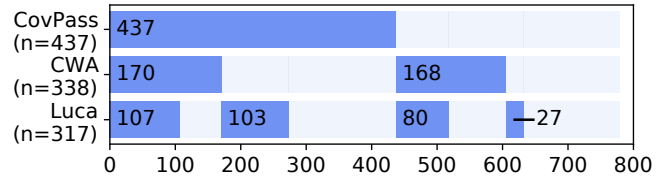


Figure 2: Overview of apps eligible to keep covid certificates installed on our participants’ smartphones (Q7). Bar segments placed below one another denote shares of participants who have installed multiple apps on their phones.

5.1 Overview of Covid Certificate Use

Figure 2 shows which covid certificate apps participants reported to have installed on their smartphones. These include the CovPass, CWA, and Luca app, with some participants using more than one app on their phone which is denoted by bar segments placed below one another in the figure. Overall, 79 % of our participants ($n = 632$) have at least one of the three apps installed. Out of the remaining 168 participants, 20 denoted to not own a smartphone (Q1).

When asked about the means primarily used to prove their vaccination or recovery status (Q13), 77 % of the respective participants (553 out of 720) named one of these three apps. Paper-based variants (e. g., the yellow WHO vaccination card) were preferably used to indicate the vaccination or recovery status by 20 % of eligible participants ($n = 142$). Interestingly, the numbers of willingness to use apps in hypothetical scenarios are much lower, as the comparison to findings by Kowalewski et al. [20] shows. They reported 37 % of participants to be willing to use a mobile app to prove their vaccination status (compared to 44 % in favour of paper-based certificates), and 12.5 % being indecisive.

For proving a negative result of a covid test (Q14), the picture is a bit different. We only asked this question to participants who were unvaccinated or did not disclose their vaccination status ($n = 93$). While 50 of these participants indicated to never use any means to provide a negative test result (e. g., when they never provide such a result at all), the remaining responses ($n = 43$) are almost evenly split across one of the apps, other digital variants, and paper-based variants. However, due to the very small subsample, we do not intend to make any claims about generalizability w. r. t. app adoption for providing negative test results. Detailed responses to these two questions are listed in Table 3.

Perceived Effort Overall, the perceived effort (Q11) required to use covid certificates was reported as rather low. The distributions of perceived effort for both paper-based and digital certificates are illustrated in Figure 3.

55 % of participants who primarily used paper-based certificates assessed the use of certificates to be *not effortful* or *a little effortful*, i. e., the lowest two levels on an equidistant

Table 3: Type of certificate primarily used (Q13/14).

Certificate Type	Vax / Recovery	Test Result
<i>Digital variants</i>		
CovPass	360 (50.0 %)	6 (6.5 %)
CWA	154 (21.4 %)	1 (1.1 %)
Luca	39 (5.4 %)	4 (4.3 %)
Other app	2 (0.3 %)	1 (1.1 %)
Other digital variant	4 (0.6 %)	12 (12.9 %)
<i>Paper-based variants</i>		
WHO certificate	112 (15.6 %)	—
Other	30 (4.2 %)	14 (15.1 %)
None of the above	19 (2.6 %)	50 (53.8 %)

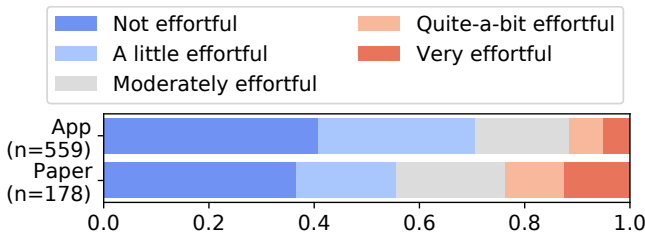


Figure 3: Perceived effort required to use digital (app-based) and paper-based covid certificates.

five-point scale ($mean = 2.44 \pm 1.94$). For digital certificates, 70 % of the respective participants assessed the required effort to be on one of the lowest two levels ($mean = 2.05 \pm 1.30$). Compared to previous findings assessing the hypothetical effort of vaccination certificates [20], the perceived effort of real usage seems to be slightly lower for both digital and paper-based covid certificates, and the difference between the two types is larger.

Access Restrictions Our online survey participants are generally in favor of restrictions applied to specific aspects of public life (Q8). For all purposes except for access to grocery stores, more than 80 % preferred one of the different types of restrictions with slight variations between purposes.

Confirming other studies [20], our participants seem to be willing to accept stronger restrictions for exceptional purposes such as *international air travel* or accessing *large events*, compared to e. g., *shopping*. When asked about the type of certificate they would use for the different purposes (Q9), we see a similar picture with participants being generally in favor of using certificates, with similar variations depending on the purpose. Paper-based certificates are preferred by approximately 25 % of participants for the majority of purposes. We observe slight deviations for access to grocery stores (20 %) and for schools (30 %). Across all purposes, the fractions of participants preferring digital certificates are in a range between 45 % and 70 %.

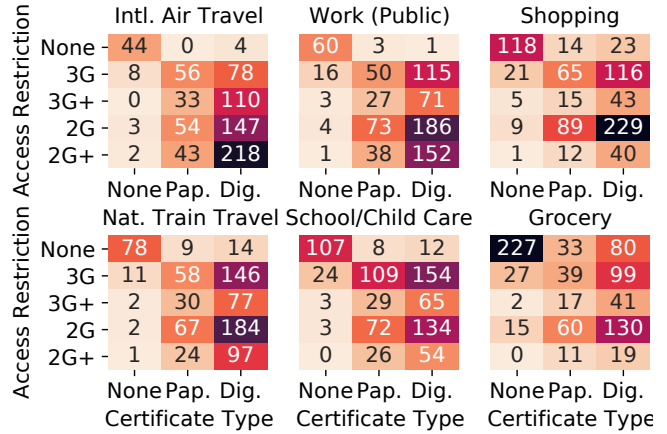


Figure 4: Numbers of participants grouped by all possible combinations of responses to Q8 (y-axis, preferred type of access restrictions) and Q9 (x-axis, preferred type of certificate) for six selected purposes.

We were also interested in whether and how answers on the strictness (Q8) and on the type of certificate likely used (Q9) might be connected. To this end, we evaluated the preferred type of certificate of participants who are in favor of a certain type of access restriction. Confusion matrices in Figure 4 show numbers of participants who responded with any combination of responses to Q8 and Q9 for six purposes. We selected these items because they represent different types of activities or have been subject to controversial public discussions in Germany. It seems that participants who are in favor of stronger restrictions (e. g., 2G or 2G+) tend to have stronger preferences for digital certificates. Most plausibly, participants who oppose access restrictions (Q8=none) by far prefer to not use any type of certificate (Q9=none).

Certificate Use in the Wild The responses we received in street interviews with business owners only partly match the data obtained in our online survey. In the interviews we found that the estimated ratio of digital covid certificates varied across business types. While in hotels and catering, we received diverse responses ranging from 50 to 90 % shares of digital certificates, cinemas, health industry, and the majority of retail reported an average of 85 to 99%. Particularly the latter seem to be higher than responses provided by online participants, among whom digital certificates were preferred by a maximum of 70 %, depending on the purpose.

5.2 Predictors for Digital Covid Certificate Use

To identify factors which foster the use of digital covid certificates, we conducted a logistic regression analysis which is the suitable method for binary outcome variables and metrical or categorical predictor variables. The use of digital covid certificates serves as outcome variable and is determined as follows: we combined responses to Q13 and Q14 into one

Table 4: Logistic regression analysis for using digital covid certificates based on the online survey data. A positive estimate and an odd ratio above one indicate higher odds of using a digital covid certificate. Significance levels are indicated with stars (* $p < .05$, ** $p < .01$, *** $p < .001$). (n = 559)

Independent variables	Est.	Odd
<i>Gender (baseline: male)</i>		
Female	0.35	1.42
<i>Education (baseline: medium education)</i>		
Low education	-0.05	0.95
High education	0.70	2.02
<i>Age (baseline: 40-59)</i>		
18-29	0.40	1.48
60-69	-0.86*	0.42
<i>[Q3/Q4]: Coronavirus infection (Baseline: no)</i>		
Yes	-0.08	0.92
<i>[Q5/Q6]: Worries about infection</i>		
	-0.04	0.96
<i>[Q12]: Vaccination status (baseline: vaccinated)</i>		
Not vaccinated	-2.81***	0.06
<i>[Q21]: Ease of use of covid certificate</i>		
	0.89***	2.42
<i>[Q23]: Knowledge about covid certificate</i>		
	0.20	1.22
<i>[Q33]: Attitudes towards measures against the spread of the coronavirus</i>		
	0.03	1.03
<i>[Q34]: Privacy disposition</i>		
	0.19	1.21
<i>[Q35]: App privacy</i>		
	-0.37*	0.69

binary variable by assigning “1” if a participant used a digital variant in either case and “0” otherwise. As we are interested in the general use of digital covid certificates, we neglect the respective certificate content, i. e., vaccination, recovery, or test result. We use the following variables as predictors for the model:

- Gender, Education, Age
- Coronavirus infection (Q3/Q4)
- Worries about coronavirus infection (Q5/Q6)
- Vaccination status (Q12)
- Ease of use of covid certificates (Q21)
- Knowledge about covid certificates (Q23)
- Attitudes towards measures against covid spread (Q33)
- Privacy disposition (Q34), App privacy (Q35)

We introduce two factors that are derived from the responses to multiple questions:

- For Q3 and Q4, we created a new factor “Coronavirus Infection” indicating “yes” for participants answering “yes” to at least one of these questions and “no” for participants answering “no” to both of these questions.
- For Q5 and Q6, we grouped the answers to a score indicating “worries about coronavirus infection”.

Table 4 shows the estimates and odd ratios of all predictors. We find four predictors that significantly influence the use of digital covid certificates. Whereas *ease of use* of covid certificates (Q21) increases the odds of using digital certificates, *older age* (60-69), *not being vaccinated* (Q12), and having more *privacy app concerns* (Q35) negatively influence the odds of using digital covid certificates. Odd ratios for *ease of use* indicate that the odds of using a digital covid certificate increase with 142% (*oddratio* = 2.42) for an increase in ease of use by 1 on a 5-point rating scale. Participants who are older (*oddratio* = .42), unvaccinated (.06), or who have higher app privacy scores (.69) are less likely to use a digital covid certificate. This confirms previous findings by Kowalewski et al. [20], who reported privacy disposition as a hindering factor for the (hypothetical) willingness to use vaccination apps, vaccination willingness positively influencing the willingness to use vaccination apps. However, age was no significant factor in their model and ease of use was not included due to the scenarios being hypothetical.

Even though only one privacy score, i. e., app privacy, is a significant predictor for the use of digital covid certificates, we conduct Wilcoxon test to see if people using digital covid certificates differ significantly in their (app) privacy dispositions from people who do not use digital covid certificates. We find significant differences between these two groups for both, privacy disposition ($p < .01$) and app privacy ($p < .001$). People not using a digital covid certificate show higher values for both privacy disposition (*mean* = 3.43 vs. *mean* = 3.22) and app privacy (*mean* = 3.32 vs. *mean* = 2.42).

5.3 Certificate Preferences

Based on whether participants have used a paper-based or digital covid certificate (Q13/Q14), we further asked them why they decided for either variant over the other one, i. e., why they chose the digital certificate (Q15) or why they preferred the paper-based alternative (Q16). To get further insights into the participants’ reasons to use either variant, Q15 and Q16 were open-ended questions.

Coding Procedure We used an iterative coding procedure to evaluate open-ended responses to these questions. The same procedure also applies to questions Q18 and Q24 presented later in this paper. Two researchers independently assigned codes for each open-ended question and each participant’s response could be assigned multiple codes. Depending on the number of responses, in a first step, an independent coding scheme was created based on a larger number of responses (approx. 100 for each open-ended question). Subsequently, a common coding scheme was agreed upon, followed by coding the remaining responses by one researcher, and finalized by a mutual validation of the responses’ codings.

Preference for Using Digital Certificates When asked about their preference for using digital certificates and deciding against using a paper-based certificate, we observed various reasons. The most common argument in the evaluated online survey responses (268 of 529 responses) is that participants carry their smartphone with them anyway (P603: “*Because I have my smartphone with me at all time*”), followed by the ease of use of digital certificates (234 of 529 responses), including the easier handling compared to paper-based certificates (P284: “*I always carry my smartphone everywhere I go. I would just forget the paper certificate*”). Ease of use also comprises statements indicating the (more) convenient use of the digital variant (P376: “*Because I find it very convenient that both partners can be stored on one smartphone. . .*”), the overall faster verification process (P657: “*More useful and works out to be more quick for me*”), and the increased practicability (P741: “*Because it [the app] is more practical*”). The fear of losing the paper-based covid certificate, especially the yellow certificate of vaccination, is also a frequently stated reason for using a digital covid certificate (60 of 529 responses). Besides fear of loss (P20: “*Fear of losing the vaccination card*”), these also include unintentionally destroying the paper variant (P581: “*. . . a paper vaccination card can get torn, smudged, or may get lost*”) and the perceived high value of the yellow certificate of vaccination, which is used for more than just vaccination against the coronavirus (P62: “*Vaccination card is too valuable for me to carry around all the time.*”). For 26 out of 529 participants, the security of the QR code-based digital covid certificates or the forgeability of paper-based covid certificates was the primary reason for choosing the digital variant (P317: “*More forgery-proof, can be scanned or should be scanned*”, P512: “*It is more secure*”, P17: “*Paper is too easy to forge, but app-based proofs are cryptographically secured*”).

Requirements imposed by some German states, events, or businesses to only recognize digital certificates or at least certificates including a QR code (12 of 529 responses) can also be a driving factor for using a digital variant (P189: “*Paper-based is not accepted everywhere*”).

Reasons to use Paper-Based Certificates To gain insights into participants’ reasons to use a paper-based variant, we asked the corresponding participants why they decided against using a digital covid certificate (Q15). We received 176 answers for this question and found similar reasons as for the preferred use of app-based certificates. Participants mentioned greater ease of use compared to digital variants ($n = 25$, e. g., P290: “*Faster to reach than the smartphone*”), carrying the paper with them anyways ($n = 11$), and fear of technical issues ($n = 23$) as reasons to use a paper version. Other reasons for using a paper-based covid certificate were not owning a smartphone ($n = 16$), regularly forgetting the smartphone ($n = 19$), or unavailability of a digital version ($n = 17$). 13 participants stated privacy concerns and 12 participants stated security

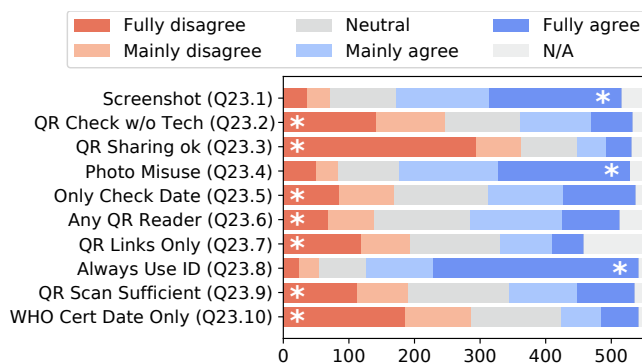


Figure 5: Participants’ agreement to statements regarding certificate QR codes and correct verification of digital covid certificates. Correct responses are labeled with an asterisk (*).

and privacy concerns for their decision against digital covid certificates (P532: “*Why should I let the apps locate me*”).

5.4 Knowledge of Digital Covid Certificates

Despite knowledge not being a significant predictor for the use of digital covid certificates, the answers to a set of knowledge questions in the survey, i. e., Q23, can provide valuable insights into users’ perception and misconceptions of certificate apps. We asked participants to indicate their level of agreement to statements mainly focusing on QR codes and the correct verification process of digital covid certificates (Q23). Here we only report answers of participants who indicated to use digital covid certificates. The distribution of their responses to all 10 statements is shown in Figure 5. In an open-ended question (Q24), we additionally asked participants about the most important aspects w. r. t. verifying digital certificates.

The knowledge score for participants using digital covid certificates ($n = 559$) is 3.45 (scale ranging from 1 to 5), indicating only moderate knowledge about these certificates, the correct verification, and especially QR codes. The statement for which we observed the most “*I do not understand the statement*” answers (18%), is Q23.7 “*QR-codes can only link to websites. URLs that simply look a bit different*”. 128 participants (23%) (rather) agree with this statement and 25% of participants indicate a neutral position to this statement. Therefore, more than half of our participants do not know that QR codes can do more than link to websites or are not sure about that.

However, 44% of the surveyed digital covid certificate users know that it is not possible to verify the validity of a certificate QR code without technical help (disagreement to Q23.2). The majority of participants (62%) also know that a picture or screenshot of a QR code can also be read by a QR code reader (agreement to Q23.1). 65% of participants know that it is not wise to publicly share a picture or screenshot

of their QR code from the covid certificate (disagreement to Q23.3). 63 % of users are aware that a picture or screenshot of their QR code could be used (maliciously) by other people (agreement to Q23.4). This shows that users are aware of possible malicious use of and cautious behavior regarding their covid certificate QR code.

Some participants also know that scanning the QR code is not sufficient for a correct verification of the digital covid certificate, as 34 % disagree with the statement “With the Corona apps, it is sufficient to scan the QR code for a correct check” (Q23.9). Interestingly, 34 % of participants agree to this statement and 28 % are not so sure, i. e., used the answer “3-neutral”. This shows that for some participants matching the information of the certificate with an identity card does not seem an important aspect, they believe scanning the QR Code is sufficient. However, the majority of digital covid certificate users (74 %) agree to the fact that a verification is only correct in combination with an ID card (agreement to Q23.8). Answers to question Q24 show that 109 participants explicitly name the QR code as an important aspect for a correct verification (code “Scan the QR code”, “QR code”). Some of them even named the correct process, i. e., “scan the QR code and match w/ ID card”. Those participants understood the importance of the QR code as a security feature. Others just named the date of the final vaccination as one of the most important aspects for verifying the certificate. Concerning the security aspects of verifying covid certificates, 226 of 442 participants describe that a complete verification is only valid in conjunction with the ID card (P768: “Name, Date of birth matching with identity card”, P551: “Comparison with ID card and scanning the QR code”). Matching the information with an identity card as an important aspect of the verification process, was agreed upon by 74 % of our participants within the knowledge questions (agreement to Q23.8). Several (92) participants stated they do not know which aspects are (most) important to verify digital certificates (P766: “Unfortunately I don’t know”, P726: “No idea”, P679: “Unfortunately, I have too little knowledge of this to give more precise information”).

5.5 Misconceptions of Digital Covid Certificates

Some answers to question Q24 in our online survey reveal misconceptions about QR code-based certificates. 18 participants incorrectly believed that showing a screenshot of the QR code to verify a covid certificate, is not valid (P65: “The code is not allowed to be a photo”). A few more participants ($n = 13$) directly mentioned that a respective app must be used, e. g., CovPass or Corona-Warn-App (P7: “It must be checked that it is not a screenshot, but is in the app”). Despite using a screenshot is perfectly fine both in terms of security and privacy. Moreover, four participants thought that scrolling (i. e., scrolling up and down the screen of the app on the owner’s smartphone) is sufficient to verify the certifi-

cates validity (P332: “You can move it [the screen] back and forth”). This is also a misconception found in the interviews (PI3, PI24: “I scroll up and down [...] like that I ensure it’s real”).

5.6 Verification Processes in the Wild

We now compare the results regarding the verification of digital covid certificates obtained in all three surveys.

Table 5 shows the frequencies of correct verification (procedure L5) for all of them: Random sampling of businesses, online survey, and street interviews. The most correct verification were named in the interviews: 50 % of the interviewees reported the procedure for the correct process – scanning the QR code *and* matching the personal data with an identity card, to make sure the certificate is valid *and* shown by the right person. In both the online survey and the random sampling, the frequencies for correct verification are lower (34 % and 37 %). Missing checks were not mentioned in either the online survey nor the street interviews but we discovered them in our sampling. Only a short glance at the digital certificate with matching the ID was observed in around 30% of all observations.

In our street interviews with business owners, responses describing checks of digital covid certificates revealed that about half of the checks are partially incorrect or missing important steps. Such checks are either missing the ID comparison or a scan with an appropriate verification app. This is also in line with our online survey, as some participants were not aware, that comparing personal data with the ID is important. A factor that seemed to positively influence the correct checks was when interviewees were provided a device for scanning purposes by their employer. This was mentioned explicitly by seven of our participants. Four participants mentioned to refrain from scanning because they *would* have to use their personal device which they did not feel comfortable with. All interviewees were sure that they conducted the verification thoroughly or very thoroughly. However, some justify this rating by stating they performed the checks *as good as they could*.

Table 5: Coding statistics – Procedure to verify digital covid certificates. Provided for the sampled checks, the open-ended responses within the online survey (Q18), and the interviews (Q7/Q8).

Procedure	Sample	Frequencies	
		Online Survey	Interviews
L1: No control	(5) 6 %	-	-
L2: Short glance	(7) 9 %	(45) 34 %	(5) 17 %
L3: Glance w/ ID	(31) 39 %	(33) 25 %	(9) 30 %
L4: Scan only	(7) 9 %	(9) 7 %	(1) 3 %
L5: Scan w/ ID	(30) 37 %	(44) 34 %	(15) 50 %
Overall responses	80	131	30

In street interviews we also identified misconceptions regarding the correct verification of digital covid certificates, that fall in line with the reported misconceptions from the survey. Some participants thought that scrolling through the app would not only be sufficient but important to determine whether the certificate is valid. Some others did not identify the QR Code as a security feature (PI10: “*We look if there are two vaccinations and check the date of the second vaccination. We only look by eye, the QR isn’t helpful for us*”). Others thought the color with which the certificate is shown indicates whether it is valid or not. One participant even mentioned “scanning” with the CWA and that it is “*odd having so much personal data of the customers on the phone*” (PI8). A similar incident has been reported in the media [17]. The CWA suits the purpose of storing the personal digital covid certificate and is not supposed to verify certificates. Scanning certificate QR codes with the CWA leads to storing the (foreign) QR code as well as the information contained within the QR code in full detail in the app. The proper verification app (CovPassCheck app) displays only the validity and basic personal information for the comparison with an ID card (1).

6 Discussion

In this study, we aimed to identify factors that influence the adoption and perception of digital covid certificates, designed to securely indicate users vaccination, test, or recovery status. As more than one app existed in Germany our results are not tailored to app design but are more broadly. We also refrain from drawing broader conclusions for other contexts, as we consider the COVID-19 pandemic as exceptional. Our results must be seen in light of restrictions in Germany during the time of our survey: Access to substantial parts of public life was only permitted with some sort of covid certificate, so one was mostly bound to use some form of covid certificates and had the choice between paper-based and digital covid certificates.

6.1 Acceptance of Digital Covid Certificates

The majority (79 %) of our participants use at least one app that offers the feature to include a digital covid certificate, which is in line with the official download numbers of the Corona-Warn-App (40 million downloads as of January 2022) and the CovPass app (23.5 million as of November 2021). The slightly higher adoption rate in our study might be due to the online panel, i. e., participants with potentially higher technology use. 70 % of survey participants usually use a digital covid certificate to indicate their vaccination, recovery, or test status when needed. The acceptance rate for app-based certificates is high, which is different from related work by Kowalewski et al. [20], finding that only 37 % of participants are willing to use a digital vaccination certificate, while 44 % would prefer a paper-based version. However, Kowalewski et

al. only surveyed usage *intention* of different implementations of vaccination apps not actual usage, as there were no vaccination apps available during the time of their study. Our results reveal that actual usage of digital covid certificates, especially when some sort of certificate is mandatory for many activities, differs from hypothetical intention to use a digital version.

Concerning access restrictions for aspects of public life, 80 % of our participants are in favor of restrictions for the mentioned purposes in this study (except for grocery shopping, see answers to Q8). As almost all of the presented purposes were restricted during the time of our study, this shows the acceptance of the measures undertaken to contain the pandemic. It seems that users favor stronger restrictions for exceptional purposes like international air travel, which confirms previous results [20]. Whereas at the workplace 3G restrictions applied in Germany, we observe high numbers in favor of stronger restrictions ($n = 449$). As also many people favor 2G or stricter restrictions for national train travel ($n = 372$), this suggests that people prefer stronger restrictions for more crowded environments like airplanes, trains, or workplaces such as offices. Most answers for no restrictions were observed for grocery shopping ($n = 384$), but opposite to German regulations 416 participants favor some restrictions (at least 3G) for grocery shopping.

Overall, digital certificates are favored over paper-based certificates by 45% to 70% across all purposes. It also seems like participants favoring stronger restrictions tend to prefer the use of digital certificates over paper-based ones.

6.2 Predictors for the Use of a Digital Covid Certificate

We find ease of use to be a significant predictor for the use of these digital certificates, not only in our logistic regression analysis but also in the open responses, in which 234 out of 529 participants use a digital certificate due to its ease of use. These findings are in line with both technology acceptance models, like the TAM, TAM 2, and UTAUT [7, 47, 48] as well as with related work researching the intention to use mobile apps [40]. Users seem to think that the easiest way to indicate their vaccination, test, or recovery status is using a corresponding app, e. g., because they carry their smartphone with them anyway (“*Because I have my smartphone with me at all time*”). Participants also stated that, by using an app, they are less likely to forget their certificate and some fear to lose their paper-based vaccination certificate, which they value as all their vaccinations (prior to covid) are included.

Another significant but hindering predictor for the use of digital covid certificates, is privacy concern related to apps. Participants with higher privacy concerns, i. e. more privacy cautious behavior, are less likely to use one of the appropriate covid apps. This is in line and conforms with related work on online technology [9, 23, 38], mobile health apps [15, 51, 54], contact tracing apps against the spread of the coronavirus [25,

46, 49] as well as (hypothetical) willingness to use mobile vaccination apps [20].

On the other hand, we observe that only 13 out of 176 participants using a paper-based certificate do so because of privacy concerns with the digital certificate. Both privacy disposition and privacy apps scores being rather moderate in our sample ($mean = 3.28$, $mean = 2.69$) indicates that participants have moderate privacy concerns at most. These privacy scores are similar to the ones Kowalewski et al. [20] observed. Therefore, our results show that when participants are not directly asked for privacy, it is only named in very few cases as a hindering aspect for using digital covid certificates. This might be due to the fact that both the CWA and the Luca App were already in use for contact tracing and event registration in the earlier phase of the pandemic, and the covid certificate functionality was added at a later point. Thus, the decision to use the app had already been made at a previous point and for a different functionality and privacy reasons were assessed already. This is in line with previous findings: It was shown earlier for contact tracing apps [25] that once the decision for using an app has been made, privacy is not a predictor for continued app usage.

People who do not plan on getting vaccinated are less likely to use a digital certificate. This might be due to the increased effort to integrate a negative test in the respective apps, as not all test facilities offer a QR code to scan test result. Older age (60 – 69) is also identified as a hindering predictor for the use of digital covid certificates, which might be due to generally lower adoption rates for technology as well as less smartphone use of older people [2, 6] (P205: “*No smartphone*”, P14: “*Because I don’t own a smartphone*”).

6.3 Knowledge and Misconceptions regarding Digital Covid Certificates

Regarding the knowledge of digital covid certificates with focus on QR codes, we observe most unsure answers for what a QR code can point to. People are not sure if QR codes are just different forms of links and can only point to websites. This might be due to users’ little exposure to QR codes, except for when they are pointing to websites. For most users, covid certificates are a new use case for QR codes. The importance and functioning of the QR code could be better explained to users, e. g., within the app. With more information maybe more users would use digital covid certificates and maybe even feel more safe using them. Out of 529 participants, only 26 mentioned the security as a reason to use the app and not the paper certificate and 12 participants use the digital certificate due to the validity of the QR code. This shows that at least a minority of users seem to understand and value the QR code as a valid security feature, but most people are not aware of that. However, most users know that sharing ones QR code publicly is not reasonable and that pictures or screenshots of QR codes can be used maliciously by others.

6.4 Perception and Misconceptions of Verifications

Regarding the correct verification process of digital covid certificates we observed one person using the CWA for the verification process in our interviews. However, the CWA is not suitable for the verification process as it extracts and stores the entire data of the digital covid certificate. The respective app to verify certificates is the CovPassCheck app.

Across all three surveys, we observed the highest estimations of correct verification processes in the interviews (50%), however these were just self-reports and the results of our sample and online survey with only 37% and 34% correct controls, hint to lower correct verification processes than self-reported by the verifiers. Such high rates of incorrect verification processes also indicate that governmental campaigns (e. g., online, TV) might have not reached all audiences in an appropriate way, or that there is lack of trust in these campaigns. However, lack of awareness and understanding may not be the only reason: Instead, interviewees did not want to use their own device for scanning the QR code and therefore refrained from scanning overall.

Therefore, a more in-depth analysis of the reasons for low adherence to correct verification is required and could be taken up by future work. For comparable situations in the future, we additionally recommend to not only provide information on specific processes, but to also allow asking for feedback and further consultation, and to actively support or assist those individuals who are in charge of executing quasi-official tasks such as verifying certificates.

7 Conclusion

Digital covid certificates are preferred by our participants over paper-based variants due to their ease of use and seamless integration into dedicated smartphone apps. Users perceive the apps as easy and convenient to use, carry their smartphone with them all the time anyway. Unfortunately, the security-related processes of scanning the QR code and matching it with the bearer’s ID card are not always followed or even known by people obliged to check certificates. Therefore, more information on security aspects of digital certificates and the correct verification process are needed, especially for people checking certificates. For further app advancement and development, we suggest to make the app as easy to use as possible, to avoid unclear design and to give users information on how to use the app, especially for verification purposes. Privacy and security indicators should be explained to users. However, our results are limited by the fact that covid certificates were mandatory for many aspects of public life in Germany, e. g., eating in a restaurant. Therefore, use and perception of these apps might be different and not directly transferable to other countries and societies.

Acknowledgments This research was supported by DFG (German Research Foundation) under Germany’s Excellence Strategy – EXC 2092 CASA – 39078197 and by the PhD School “SecHuman” by the federal state of NRW, Germany.

References

- [1] Samuel Altmann, Luke Milsom, Hannah Zillessen, Raffaele Blasone, Frederic Gerdon, Ruben Bach, Frauke Kreuter, Daniele Nosenzo, Séverine Toussaert, and Johannes Abeler. Acceptability of App-Based Contact Tracing for COVID-19: Cross-Country Survey Study. *JMIR mHealth and uHealth*, 8(8):e19857, August 2020.
- [2] Ionut Andone, Konrad Błaszkiwicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. How age and gender affect smartphone usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp ’16, page 9–12, New York, NY, USA, 2016. Association for Computing Machinery.
- [3] Nadine Bol, Natali Helberger, and Julia C. M. Weert. Differences in Mobile Health App Use: A Source of new Digital Inequalities? *The Information Society*, 34(3):183–193, April 2018.
- [4] Xi Chen and Shun Cai. Self-disclosure under social networking sites: A risk-utility decision model. In *International Conference on Electronic Commerce*. Association for Computing Machinery, 2012.
- [5] culture4life GmbH. Impf- und genesenennachweis hinterlegen, February 2022. <https://www.luca-app.de/impfnachweise-hinterlegen/> as of June 13, 2022.
- [6] Sara J. Czaja, Neil Charness, Arthur D. Fisk, Christopher Hertzog, Sankaran Nair, Wendy A. Rogers, and Joseph Sharit. Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (create). *Psychology and aging*, 21 2:333–52, 2006.
- [7] Fred D. Davis. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3):319–340, September 1989.
- [8] Christopher Elliott. What you need to know about vaccine passports, December 2020. https://www.washingtonpost.com/lifestyle/travel/yellow-card-vaccine-passport/2020/12/30/746c0558-40b7-11eb-8db8-395dedaaa036_story.html as of June 13, 2022.
- [9] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie F. Cranor. Which Privacy and Security Attributes Most Impact Consumers’ Perception and Willingness to Purchase IoT Devices? In *IEEE Symposium on Security and Privacy*, SP ’21, San Francisco, CA, USA, May 2021. IEEE.
- [10] European Commission. EU Digital COVID Certificate, June 2021. https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans/eu-digital-covid-certificate_en as of June 13, 2022.
- [11] European Commission. Eu digital covid certificate, February 2022. https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans/eu-digital-covid-certificate_en as of June 13, 2022.
- [12] European Data Protection Supervisor. Smart vaccination certificate, February 2022. https://edps.europa.eu/press-publications/publications/techsonar/smart-vaccination-certificate_de as of June 13, 2022.
- [13] Federal Foreign Office Germany. Covid-19: entry and quarantine regulations in germany, January 2022. <https://www.auswaertiges-amt.de/en/coronavirus/2317268> as of June 13, 2022.
- [14] Fisher & Phillips LLP. California boosts vaccination requirement for workers in healthcare settings, January 2022. <https://www.fisherphillips.com/news-insights/california-boosts-vaccination-requirement-workers-healthcare.html> as of June 13, 2022.
- [15] Jie Gu, Yunjie (Calvin) Xu, Heng Xu, Cheng Zhang, and Hong Ling. Privacy concerns for mobile app download: An elaboration likelihood model perspective. *Decision Support Systems*, 94:19–28, February 2017.
- [16] Heilweil, Rebecca. How new york overengineered its million-dollar vaccine passport, October 2021. <https://www.vox.com/recode/22736276/new-york-state-vaccine-passport-excelsior-pass> as of June 13, 2022.
- [17] Jochen Hilgers. Kölner barbesitzer weist auf mangelhafte kontrollen der impfnachweise hin, December 2021. <https://www1.wdr.de/nachrichten/rheinland/barbesitzer-deckt-sicherheitsluecke-corona-warn-app-auf-100.html>.
- [18] Israel Ministry of Health. What is a Green Pass?, May 2021. <https://corona.health.gov.il/en/directives/green-pass-info/> as of June 13, 2022.
- [19] Gabriel Kaptchuk, Daniel G. Goldstein, Eszter Hargittai, Jake Hofman, and Elissa M. Redmiles. How Good is Good Enough for COVID19 Apps? The Influence

- of Benefits, Accuracy, and Privacy on Willingness to Adopt. *arXiv preprint arXiv:2005.04343*, 2020.
- [20] Marvin Kowalewski, Franziska Herbert, Theodor Schnitzler, and Markus Dürmuth. Proof-of-Vax: Studying User Preferences and Perception of Covid Vaccination Certificates. In *Proceedings on Privacy Enhancing Technologies*, PETS '22, pages 317–338, Sydney, Australia, November 2021. sciendo.
- [21] Tianshi Li, Jackie (Junrui) Yang, Cori Faklakis, Jennifer King, Yuvraj Agarwal, Laura Dabbish, and Jason I. Hong. Decentralized is not risk-free: Understanding public perceptions of privacy-utility trade-offs in COVID-19 contact-tracing apps. *arXiv preprint arXiv:2005.11957*, 2020.
- [22] Yuan Li. A multi-level model of individual information privacy beliefs. *Electronic Commerce Research and Application*, 13(1), 2014.
- [23] Sebastian Linsner, Franz Kuntke, Enno Steinbrink, Jonas Franken, and Christian Reuter. The Role of Privacy in Digitalization – Analyzing Perspectives of German Farmers. In *Proceedings on Privacy Enhancing Technologies 2021*, pages 334–350. Sciendo, March 2021.
- [24] Xi Lu, Tera L. Reynolds, Eunkyung Jo, Hwajung Hong, Xinru Page, Yunan Chen, and Daniel A. Epstein. Comparing Perspectives Around Human and Technology Support for Contact Tracing. In *ACM CHI Conference on Human Factors in Computing Systems*, CHI '21, Virtual Event, May 2021. ACM.
- [25] Yannic Meier, Judith Meinert, and Nicole C. Krämer. Investigating factors that affect the adoption of covid-19 contact-tracing apps: A privacy calculus perspective. *Technology, Mind, and Behavior*, 2(3), 8 2021. <https://tmb.apaopen.org/pub/i7moqr3r>.
- [26] Miguel, Ken and Glover, Julian. How to show proof of vaccination in san francisco or anywhere in california, August 2021. <https://abc7news.com/california-digital-vaccine-card-proof-of-vaccination-mandate-my-record-cdph/10950078/> as of June 13, 2022.
- [27] Murphy, J. Kim. L.a. movie theaters, restaurants will require proof of covid-19 vaccination starting monday, November 2021. <https://variety.com/2021/biz/news/los-angeles-covid-vaccination-proof-movie-theaters-restaurants-1235105494/> as of June 13, 2022.
- [28] Office for Human Research Protections. The Belmont Report. Technical report, Office for Human Research Protections (OHRP), January 2018.
- [29] Samantha M Olson, Margaret M Newhams, Natasha B Halasa, Ashley M Price, Julie A Boom, Leila C Sahni, Pia S Pannaraj, Katherine Irby, Tracie C Walker, Stephanie P Schwartz, et al. Effectiveness of BNT162b2 Vaccine against Critical Covid-19 in Adolescents. *New England Journal of Medicine*, January 2022.
- [30] Q_PERIOR. Digital certificate of vaccination: technical eu specifications and how they are implemented nationally, February 2022. <https://www.q-perior.com/en/fokusthema/digital-certificate-of-vaccination-technical-eu-specifications-and-how-they-are-implemented-nationally/> as of June 13, 2022.
- [31] Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian, and Max Roser. Policy Responses to the Coronavirus Pandemic, 2020. <https://ourworldindata.org/policy-responses-covid> as of June 13, 2022.
- [32] Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian, and Max Roser. Coronavirus (COVID-19) Vaccinations, 2022. <https://ourworldindata.org/covid-vaccinations> as of June 13, 2022.
- [33] Robert Koch Institut. The CovPass-App, May 2021. <https://digitaler-impfnachweis-app.de/en> as of June 13, 2022.
- [34] Robert Koch Institut. Check EU COVID certificates directly via app, January 2022. <https://www.digitaler-impfnachweis-app.de/en/covpasscheck-app/> as of June 13, 2022.
- [35] Robert Koch Institute. Check eu covid certificates directly via app, August 2021. <https://www.digitaler-impfnachweis-app.de/en/covpasscheck-app> as of June 13, 2022.
- [36] Eli S Rosenberg, Vajeera Dorabawila, Delia Easton, Ursula E Bauer, Jessica Kumar, Rebecca Hoen, Dina Hoefler, Meng Wu, Emily Lutterloh, Mary Beth Conroy, Danielle Greene, and Howard A Zucker. COVID-19 vaccine effectiveness in New York state. *New England Journal of Medicine*, 386(2):116–127, 2022.
- [37] SAP Germany. Open-Source Project Corona-Warn-App, June 2020. <https://www.coronawarn.app/en/> as of June 13, 2022.

- [38] Theodor Schnitzler, Shujaat Mirza, Markus Dürmuth, and Christina Pöpper. SoK: Managing Longitudinal Privacy of Publicly Shared Personal Online Data. In *Proceedings on Privacy Enhancing Technologies 2021*, pages 229–249. Sciendo, November 2020.
- [39] Lucy Simko, Ryan Calo, Franziska Roesner, and Tadayoshi Kohno. COVID-19 Contact Tracing and Privacy: Studying Opinion and Preferences. *arXiv preprint arXiv:2005.06056*, 2020.
- [40] Carlos Tam, Diogo Santos, and Tiago Oliveira. Exploring the influential factors of continuance intention to use mobile apps: Extending the expectation confirmation model. *Information Systems Frontiers*, 22, 02 2020.
- [41] The federal Government. Vaccinations: protection for everyone, August 2021. <https://www.bundesregierung.de/breg-en/news/federal-regional-consultation-coronavirus-1949666> as of June 13, 2022.
- [42] The Local. How is italy using covid health passes compared to elsewhere in europe?, July 2022. <https://www.thelocal.it/20210728/how-is-italy-using-covid-health-passes-compared-to-elsewhere-in-europe/> as of June 13, 2022.
- [43] Simon Trang, Manuel Trenz, Welf H. Weiger, Monideepa Tarafdar, and Christy M. K. Cheung. One app to trace them all? Examining app specifications for mass acceptance of contact-tracing apps. *European Journal of Information Systems*, 29(3):1–14, July 2020.
- [44] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *ACM CHI Conference on Human Factors in Computing Systems*, CHI '21, Virtual Event, May 2021. ACM.
- [45] United Nations Educational, Scientific and Cultural Organization (UNESCO). International Standard Classification of Education (ISCED 2011), December 2012. <http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf> as of June 13, 2022.
- [46] Christine Utz, Steffen Becker, Theodor Schnitzler, Florian Farke, Franziska Herbert, Leonie Schaewitz, Martin Degeling, and Markus Dürmuth. Apps Against the Spread: Privacy Implications and User Acceptance of COVID-19-Related Smartphone Apps on Three Continents. In *ACM CHI Conference on Human Factors in Computing Systems*, CHI '21, Virtual Event, May 2021. ACM.
- [47] Viswanath Venkatesh and Fred Davis. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46:186–204, 02 2000.
- [48] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3):425–478, September 2003.
- [49] Michel Walrave, Cato Waeterloos, and Koen Ponnet. Ready or not for contact tracing? investigating the adoption intention of covid-19 contact-tracing technology using an extended unified theory of acceptance and use of technology model. *Cyberpsychology, Behavior, and Social Networking*, 24, 10 2020.
- [50] Rachel Wilf-Miron, Vicki Myers, and Mor Saban. Incentivizing Vaccination Uptake: The “Green Pass” Proposal in Israel. *JAMA*, 325(15), April 2021.
- [51] Verena M. Wottrich, Eva A. van Reijmersdal, and Edith G. Smit. The privacy trade-off for mobile app downloads: The roles of app value, intrusiveness, and privacy concerns. *Decision Support Systems*, 106:44–52, February 2018.
- [52] Baobao Zhang, Sarah Kreps, and Nina McMurry. Americans’ Perceptions of Privacy and Surveillance in the COVID-19 Pandemic. *OSF preprint osf.io/9wz3y*, 2020.
- [53] Yixuan Zhang, Yifan Sun, Lace Padilla, Sumit Barua, Enrico Bertini, and Andrea G. Parker. Mapping the Landscape of COVID-19 Crisis Visualizations. In *ACM CHI Conference on Human Factors in Computing Systems*, CHI '21, Virtual Event, May 2021. ACM.
- [54] Leming Zhou, Jie Bao, Valerie Watzlaf, and Bambang Parmanto. Barriers to and Facilitators of the Use of Mobile Health Apps From a Security Perspective: Mixed-Methods Study. *JMIR mHealth and uHealth*, 7(4):e11223, April 2019.

A Questionnaire – Online Survey

Welcome Text Study on the topic of 2G-/3G certificates. Thank you for your interest in our study!

In this study we will ask a series of questions about vaccination, recovery, and test certificates. The purpose of this survey is to get a comprehensive understanding of 2G-/3G certificates in the context of the coronavirus pandemic of the German population. By participating, you can make a valuable contribution to this purpose.

Purpose: This scientific study investigates your perception of 2G-/3G- regulations by means of (digital) proofs, as they are required (e. g., for cinema visits or other activities and events).

Prerequisites: To participate in this study, you must be at least 18 years old.

Duration: Participation in the study is expected to last 20 minutes. There are no anticipated risks for you to participate. Please answer the questionnaire as honestly as possible. If you no longer wish to participate in this study, you may discontinue at any time as long as you have not yet submitted your answers or they have not yet been evaluated.

Contact: The study is conducted by researchers [...]. If you have any questions about or problems with this research, please feel free to contact [...].

Data protection: Your responses in this study will be linked to your Respondi-ID and will be stored in pseudonymous format. We do not ask for any information that could identify you personally. This data is collected on behalf of the [...] and will not be passed to third parties. By starting the questionnaire, you agree to the collection of data for the purpose of conducting this study. The processing of your personal data is based on Article 6 (1) DSGVO and §17 DSG NRW. You have the right to revoke your consent to data processing at any time, as well as to request information, correction, restriction of processing and deletion of your personal data. To exercise these rights, please contact the e-mail address mentioned above. The competent supervisory authority is the Data Protection Commissioner of the State of North Rhine-Westphalia.

Declaration of Consent

Q0: Please confirm that you have read the above terms and conditions and that you are at least 18 years old. [single choice]

- *I hereby confirm that I accept the conditions of participation in this study and that I am at least 18 years old.*

Demographics First, we would like to obtain some information about you.

Q_A: How old are you? [single choice]

- *18-29; 30-39; 40-49; 50-59; 60-69*

Q_G: What is your gender? [single choice]

- *Female; Male; Non-binary; Describe yourself (free-text answer); Prefer not to answer*

Q_E: What is your highest level of education? [single choice]

- *No school leaving certificate; Secondary school (primary school) or equivalent leaving certificate; High school (O level) or equivalent leaving certificate; A level, vocational high school / general or university entrance qualification; Occupational or vocational training / apprenticeship; Completion of a technical college or administrative or professional academy; Bachelor's degree; Diploma university course or masters (including: teaching position, state examination, Master's course, artistic or comparable courses of study); PhD; Prefer not to answer*

Q_K: Do you have practical experience in computer science, computer technology or information technology fields (e. g., through your job or education background)? [single choice]

- *Yes; No; Prefer not to answer*

General Questions and Experiences with the Coronavirus First, we would like to ask you some general questions about your smartphone use and your experience with the coronavirus.

Q1: Do you own a smartphone? [single choice]

- *Yes; No*

Q2: [If “Yes” in Q1] Do you use an app (or smartwatch) to monitor your health or track your fitness? [single choice]

- *Yes; No*

Q3: Are you or have you been infected with the coronavirus? [single choice]

- *Yes; No; Prefer not to answer*

Q4: Is there a person in your social circle who is or has been infected with the coronavirus? [single choice]

- *Yes; No; Prefer not to answer*

Q5: How concerned are you that you will become infected with the coronavirus? [single choice]

- *1 – Not concerned; 2 – A-little concerned; 3 – Moderately concerned; 4 – Quite-a-bit concerned; 5 – Very concerned*
- *Prefer not to answer*

Q6: How concerned are you that someone you are close to may be infected with the coronavirus? [single choice]

- *same answer options as Q5*

2G-/3G Certificates

Q7: [If “Yes” in Q1] Which of the following COVID-19 apps do you have installed on your smartphone? [multiple choice]

- *Corona-Warn-App; Luca App; CovPass App; CovPass Check App; Other / Additional Corona specific apps (please specify); [exclusive answer] I have not installed any Corona specific app*

Q8: Which type of events or purposes should require a certificate? Please mark the appropriate form of certificate. [matrix table]

- *items: National flights; International flights; National railroad travel; International railroad travel; Crossing countries by car (i. e., outside Germany); Overnight stays in hotels (domestic and abroad); Participation in major events (e. g., soccer matches, concerts); Visits to restaurants, museums, and cinemas; To be allowed to carry out professional activities with public interaction (e. g., hospitals, care facilities); Sport clubs and gyms; Beauty related services (e. g., hairdressing, cosmetics); Private events (e. g., weddings, birthday parties); Retail (clothing stores, construction stores); Stores for daily needs (e. g., grocery stores, pharmacies); Facilities such as schools, daycare centers, and after-school programs; This is an attention check question. Please mark the answer “2G: vaccinated, recovered”*
- *answer options: No certificate should be required; 3G: vaccinated, recovered, or tested (rapid test); 3GPlus: vaccinated, recovered, or tested (PCR test); 2G: vaccinated or recovered; 2GPlus: vaccinated or recovered and additionally tested (rapid test)*

Q9: What variant of certificate would you want to use for the respective purpose? [matrix table]

- *items: same items as Q8 without attention check question*
- *answer options: No certificate should be required; paper-based certificate (e. g., yellow certificate of vaccination , print-out from test center); digital certificate (Corona-Warn-App, CovPass app, or email from test center)*

Q10: Have you already visit events or stores that required proof of vaccination, recovery, or test? [single choice]

- *Yes; No; Don’t know; Prefer not to answer*

Q11: How effortful do you perceive showing proof of vaccination, recovery, or test to be? [single choice]

- *1 – Not effortful; 2 – A-little effortful; 3 – Moderately effortful; 4 – Quite-a-bit effortful; 5 – Very effortful*

Q12: Have you already been vaccinated or recovered against the coronavirus? [single choice]

- *Yes; No; Prefer not to answer*

Q13: [If “Yes” or “Prefer not to answer” in Q12] Which of the following certificates do you typically use to proof your coronavirus vaccination or your recovery, e. g., when visiting a restaurant?

- *Corona-Warn-App; Luca app, Covpass app; Other Corona specific app (please specify); Other digital variant (e. g., email from your doctor; photo of your certificate); Yellow certificate of vaccination; Other paper-based certificate (e. g., print-out from test center); I do not use any of these variants*

Q14: [If “No” or “Prefer not to answer” in Q12] Which of the following certificates do you typically use to proof your coronavirus test, e. g., when visiting a restaurant?

- *Corona-Warn-App; Luca app, Covpass app; Other Corona specific app (please specify); Other digital variant (e. g., email from your doctor; photo of your certificate); Other paper-based certificate (e. g., print-out from test center); I do not use any of these variants*

Q15: [If “[any paper-based variant]” in Q13/Q14] Why do you use a paper-based certificate (instead of a digital variant)? Why did you decide against a digital certificate? [free-text]

Q16: [If “[any digital variant]” in Q13/Q14] Why do you use a digital certificate (instead of a paper-based variant)? Why did you decide against a paper-based certificate? [free-text]

Certificate Verification Process

- Q17: [If “[any paper-based variant]” in Q13/Q14] Please think about your last control(s) and describe how your paper-based certificate was verified. [free-text]
- Q18: [If “[any digital variant]” in Q13/Q14] Please think about your last control(s) and describe how your digital certificate was verified. [free-text]
- Q19: Please reflect back on the control(s) you just described. How careful did you perceive this control(s) was? [single choice]
- 1 – Not carefully; 2 – A-little carefully; 3 – Moderately carefully; 4 – Quite-a-bit carefully; 5 – Very carefully
 - Prefer not to answer
- Q20: How secure did you feel from an infection by this control(s)? [single choice]
- 1 – Not secure; 2 – A-little secure; 3 – Moderately secure; 4 – Quite-a-bit secure; 5 – Very secure
 - Prefer not to answer
- Q21: How easy do you perceive it is to prove your certificate using an app? [single choice]
- 1 – Not easy; 2 – A-little easy; 3 – Moderately easy; 4 – Quite-a-bit easy; 5 – Very easy
 - I do not use an app for this
- Q22: Please rank the following certificate variants related to their forgery resistance in descending order, i. e., the most forgery-resistant certificate comes in first place. Feel free to place several certificate variants on the same rank or in the same place. [order and rank task]
- items: *Digital certificates with QR-code (e. g., Corona-Warn-App, CovPass app); Yellow certificate of vaccination; Paper-based certificates (e. g., print-out from the test center)*
 - answer options: *Rank 1; Rank 2; Rank 3*
- Q23: Please indicate whether you agree with each of the following statements. [matrix table]
- items: *A photo or screenshot from a QR-code can also be read by a QR code reader; Even without technical devices, you can tell if a QR-code within a corona app is valid; It is harmless to publicly share a photo or screenshot of the QR-code from my corona app; A photo of a QR-code from an app (e. g., Corona-Warn-App) can be photographed and used by an unauthorized person; For a secure verification of digital vaccination certificates, it is sufficient to check the date of the 2nd vaccination within one of the available apps; The validity of a QR code for vaccination certificates (e. g., within the Corona-Warn-App) can be verified with any QR-code reader; QR-codes can only link to websites, they are just differently looking URLs; Correct verification of vaccination certificates is only possible in any case (paper-based or digital) in combination with an ID document; With the Corona apps, it is sufficient to scan the QR-code for a correct check; In the case of the yellow certificate of vaccination, it is sufficient to look for the vaccination date for a correct verification*
 - answer options: *1 – Fully-disagree; 2 – Mainly-disagree; 3 – Neutral; 4 – Mainly-agree; 5 – Fully-agree*
- Q24: What aspects do you think are the most important to verify the digital certificates? What do you think needs to be verified in the case of an app, for example? How can a forgery be detected? In your opinion, what are (technical) security indicators? [free-text]
- Q25: What aspects do you think are the most important to verify the paper-based certificates? What do you think needs to be verified within the yellow certificate of vaccination, for example? How can a forgery be detected? In your opinion, what are security indicators? [free-text]
- Q26: Please drag all the items into the box that are in your opinion necessary for a correct verification of a digital vaccination, recovery, or test certificate consisting of a QR code. Please use the order as you think the verification should proceed. [order and rank task]
- items: *Match ID document, such as ID card, with the displayed personal data within the app used to scan the QR code (e. g., CovPass Check app); Scan QR code with a suitable app, e. g., CovPass Check app; Check manually the date of the 2nd vaccination; Scroll to the 2nd vaccination date within the person’s Corona-Warn-App or CovPass app; Check the person’s Corona-Warn-App or CovPass app to verify if 2/2 vaccinations are displayed; Match name within the person’s app (or on the person’s document) with an identification document*
 - answer options: *Correct verification consists of*
- Q27: [If “[any paper-based variant]” in Q13/Q14] Please think back to the situations in which you were checked. In what percentage of cases was your paper-based certificate checked professionally, i. e.: the data within the, e. g., yellow certificate of vaccination or on the print-out was verified *and additionally* the data was compared with your ID card? [single choice]
- Please note the slider to your desired position (you can only adjust the slider in steps of 5)
- Q28: [If “[any digital variant]” in Q13/Q14] Please think back to the situations in which you were checked. In what percentage of cases was your digital certificate checked professionally, i. e.: the QR code was scanned, *and additionally* the data was compared with your ID card? [single choice]
- Please note the slider to your desired position (you can only adjust the slider in steps of 5)

Certificate Inspectors / Verifier

Q29: Have you already personally verified vaccination, recovery, or test certificates (e. g., in the course of performing your job duties)? [single choice]

- *Yes; No*

Q30: [If “Yes” in Q29] In which business area do you work? [single choice]

- *Hotel business; Gastronomy; Body-related services (e. g., hairdressing, cosmetics); artistic sector (e. g., theater, museums), Other (please specify)*

Q31: [If “Yes” in Q29] Please describe how you usually verify vaccination, recovery, or test certificates. [free-text]

Q32: [If “Yes” in Q29] How time-consuming do you perceive conducting these verifications? [free-text]

- *1 – Not effortful; 2 – A-little effortful; 3 – Moderately effortful; 4 – Quite-a-bit effortful; 5 – Very effortful; Prefer not to answer*

Pandemic Situation

Q33: Please indicate whether you agree with each of the following statements. [matrix table]

- *The 3G rule is contributing in containing the coronavirus pandemic; The 2G rule is contributing in containing the coronavirus pandemic; Contact restrictions are contributing in containing the coronavirus pandemic; School closures are contributing in containing the coronavirus pandemic; Work at home is contributing in containing the coronavirus pandemic; Most people I care about think that coronavirus vaccinations are important to contain the coronavirus pandemic; Vaccination against COVID-19 contributes to the containment of the coronavirus pandemic; Mandatory mask-wearing is contributing in containing the coronavirus pandemic*

Privacy Disposition

Q34: For each of the following statements, please indicate the extent to which you agree.² [matrix table]

- *items: Compared to others, I am more sensitive about the way other people or organizations handle my personal information; Compared to others, I see more importance in keeping personal information private; Compared to others, I am less concerned about potential threats to my personal privacy (R); Compared to others, I value health data as especially worthy of protection*
- *answer options: 1 – Fully-disagree; 2 – Mainly-disagree; 3 – Neutral; 4 – Mainly-agree; 5 – Fully-agree*

Q35: For each of the following statements, please indicate the extent to which you agree.³ [matrix table]

- *items: I am concerned that the information I submit in a corona app could be misused; I am concerned about submitting information in a corona app, because of what others might do with it; I am concerned about submitting information in a corona app, because it could be used in a way I did not foresee; I am concerned about disclosing health data in a corona app*
- *answer options: 1 – Fully-disagree; 2 – Mainly-disagree; 3 – Neutral; 4 – Mainly-agree; 5 – Fully-agree*

Demographics (German state)

Q36: In which state do you live? [single choice]

- *Baden-Württemberg; Bavaria; Berlin; Brandenburg; Bremen; Hamburg; Hessen; Mecklenburg Western Pomerania; Lower Saxony; Northrhine-Westphalia; Rhineland Palatinate; Saarland; Saxony; Saxony-Anhalt; Schleswig Holstein; Thuringia*

²The first three items are from the “Disposition to privacy” scale in the version of Yuan Li [22].

³The first three items are from the “Perceived Privacy Risk” scale in the version of Chen and Cai [4].

B Questionnaire – Interviews

Note: Textparts in red where notes for the interviewers and not necessarily asked during each interview.

Thanks a lot for agreeing to talk to us. We will note all your answers but keep them anonymous. We will solely document your industry and the position you work in. Are you ok with that?

Q1: Hence the first question: In which position do you work here?

(Meaning e. g., employee or owner)

Q2: Under what conditions are guests currently allowed to receive your services or stay with you? Please describe them.

(If necessary assist mentioning 2G or 3G, etc.)

Q3: Do you check customers' test, recovery, or immunization records as part of your job?

Q4: What do you estimate is the percentage of paper-based certificates (e. g., yellow immunization card) that you are shown?

Q5: Please think of your current certificate checks or the checks you did during the last few weeks. Please describe how you typically check **paper-based** test, recovery, or immunization records (e. g., yellow immunization card).

(Follow-up questions, if applicable: How confident are you that your checks are sufficient / "safe"? / How confident do you feel performing it?)

Q6: What aspects do you think are most important for checking **paper-based** certificates?

(For example, what do you think needs to be verified in the yellow vaccination card? How can a forgery be detected? What do you think are the security indicators?)

Q7: Please think of your current certificate checks or the checks you did during the last few weeks. Please describe how you typically check **digital** test, recovery, or immunization records (e. g., in the Corona-Warn-App or CovPass App).

(Follow-up questions, if applicable: How confident are you that your checks are sufficient / "safe"? / How confident do you feel performing it?)

Q8: What aspects do you think are most important for checking **digital** certificates?

(For example, what do you think needs to be verified in the yellow vaccination card? How can a forgery be detected? What do you think are the security indicators?)

The following three questions were asked separately for paper-based and digital certificates.

Q9: On a scale from 1 - not sure to 5 - very sure: How sure are you to recognize forged certificates?

(Follow-up question: Have you ever recognized a forgery before? If so, how?)

Q10: On a scale from 1 - not time-consuming to 5 - very time-consuming. How time-consuming do you perceive the certificate checks to be?

Q11: On a scale from 1 - not thoroughly to 5 - very thoroughly, how thoroughly do you think you execute your checks?

Q12: Thinking about the last few weeks, did you have more positive or negative experiences with checking test, recovery, or immunization records?

(e. g., sympathetic guests; Would you like to tell us/myself about those experiences?)

Q13: Do you feel adequately informed by politics (or your managers) about how to correctly check the various certificates?

(Have you been trained on how to check certificates?)

Q14: Would you have hoped for (more) education, support, or information from politics (or you managers or associations, e. g., Dehoga)?

(What kind of education, support, and/or information would you have wished for?)

Q15: Do you have any concerns regarding the verification of the different certificates?

(Difficulties e. g., to detect forgeries, scaring away guests, etc.)

Q16: Would you like to tell us anything else?

Thanks a lot for your time and our discussion!

“As soon as it’s a risk, I want to require MFA”: How Administrators Configure Risk-based Authentication

Philipp Markert , Theodor Schnitzler , Maximilian Golla* , and Markus Dürmuth‡ 
Ruhr University Bochum, *Max Planck Institute for Security and Privacy, ‡Leibniz University Hannover

Abstract

Risk-based authentication (RBA) complements standard password-based logins by using knowledge about previously observed user behavior to prevent malicious login attempts. Correctly configured, RBA holds the opportunity to increase the overall security without burdening the user by limiting unnecessary security prompts to a minimum. Thus, it is crucial to understand how administrators interact with off-the-shelf RBA systems that assign a risk score to a login and require administrators to configure adequate responses.

In this paper, we let $n = 28$ system administrators configure RBA using a mock-up system modeled after Amazon Cognito. In subsequent semi-structured interviews, we asked them about the intentions behind their configurations and experiences with the RBA system. We find that administrators want to have a thorough understanding of the system they configure, show the importance of default settings as they are either directly adopted or depict an important orientation, and identify several confusing wordings. Based on our findings, we give recommendations for service providers who offer risk-based authentication to ensure both usable and secure logins for everyone.

1 Introduction

Password-based authentication is still the dominant form of user authentication, despite severe weaknesses such as phishing attacks [41, 49], password reuse attacks [14, 23], and their guessability [47, 57]. Password alternatives such as biometric authentication [34, 67], graphical passwords [6, 56], or

security keys [12, 19] all have their own set of drawbacks that so-far have prevented their widespread adoption [9, 27].

To improve user security, services deployed additional protection mechanisms to reinforce passwords, for example, by using *multi-factor authentication* (MFA) [13, 22, 30], proactive *password-reuse checks* [36, 45, 53], and *risk-based authentication* (RBA) [16, 20, 64]. Several authorities, such as the NCSC [42], NIST [24], and others [7], all mention risk-based authentication as one of the key concepts to minimize account compromises.

RBA is a method for strengthening user authentication on the server’s side without involving the user (except for rare cases). Thus, it offers the potential to increase the security of accounts without burdening the legitimate user. However, RBA comes at the cost of being a privacy-invasive technique that requires login behavior monitoring and client-side fingerprinting [8, 66]. At the moment of password entry, RBA monitors a variety of signals, such as the source IP, user-agent, login time, and further information about the user’s machine, e.g., obtainable via client-side fingerprinting. This information is then compared with the user’s profile from past logins, as well as profiles from typical attacks. Based on this information, a risk level is computed [20, 26].

The configuration of an RBA system requires administrators to decide how the system should treat logins with different risk levels. We consider this a non-trivial configuration task as it interferes with usability and security requirements that directly impact the user. In this work, we study how administrators interact with configuration interfaces for RBA. We focus on professionals not specialized in the administration of RBA, which we assume is rather common in small and medium-sized enterprises. To the best of our knowledge, our work is the first to study the rationale behind configuring RBA systems on the administrators’ side. Thus, we keep our research exploratory and follow three broad research questions. RQ1: *How do administrators configure RBA?* (e.g., risk-level behavior, when and how to notify), RQ2: *Which obstacles and misunderstandings do they encounter?*, and RQ3: *What*

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

is the impact of previous exposure to other RBA systems and how do different requirements influence administrators?

In our two-part study, we assigned $n = 28$ administrators a configuration task for adjusting risk level behavior and RBA notification settings in an enterprise scenario. The configuration tool they worked with resembled the look-and-feel of Amazon Cognito, the system that *Amazon Web Services* (AWS) offers to its customers. Subsequently, we interviewed participants about their intentions behind the configurations they made, their interaction with the configuration system, and potential obstacles they encountered while completing their task. To facilitate the recruitment of system administrators from different continents, the study was conducted online using a video conferencing tool and an online web interface accessible to the participants.

Our results suggest that system administrators want to deeply understand how risk-based authentication systems work in order to be able to make informed decisions. For example, the tool we used hid some complexity behind generic phrases such as *low*, *medium*, and *high risk*, which was criticized by several participants. Additionally, our study identifies issues with room for improvement and other topics to be explored by future research—both in more detail and in a larger variety of RBA configuration systems. In summary, our paper makes the following key contributions:

- Through an in-depth qualitative evaluation of interviews, we complement existing knowledge about risk-based authentication by providing insights into administrators' decision processes.
- Our study shows that system administrators desire detailed information about risk levels and the ability to make fine-grained configurations in order to ensure appropriate risk level behavior.
- Our findings unveil several issues to be explored by future research, while at the same time indicating first recommendations for service providers to ensure usability and security of RBA systems.

2 Related Work

In this paper, we study how administrators configure risk-based authentication. Since there are no other studies to the best of our knowledge, we align this section along prior work about RBA and studies focusing on system administrators.

Risk-based Authentication In 2010, Google added a new feature to protect their users from suspicious account activity [16], and while, e.g., Facebook also stated to employ risk-based authentication [46], not much was publicly known about its inner working. In 2016, Microsoft started to offer risk-based conditional access to its Azure AD customers and supported risk events like unfamiliar locations, impossible travel, IP addresses with suspicious activity, and users

with leaked credentials [50]. At the same time, Hurkafa [29], Bonneau et al. [10], and Freeman et al. [20] discussed the potentials of RBA. The latter also presented a prototype and found that an algorithm based on the user's IP address and user-agent history has a recall rate of up to 89% and a false-positive rate of 10%. Later, other features like the round-trip time of IP packets were found to be useful [48, 63].

Wiefling et al. [64] showed that verification codes sent via email are the de-facto standard for login challenges enforced by RBA. In a subsequent study, they demonstrated that providing this code in the subject can reduce the login time [65]. A study by Doerfler et al. [17] evaluated the efficacy of login challenges at preventing account takeovers. They found that up to 94% of phishing-rooted hijacking attempts and even 100% of automated hijacking attempts can be prevented. As shown by Wiefling et al. [62], RBA is perceived as more secure than passwords but also more usable than multi-factor authentication. While the latter poses an even higher security standard, increasing its adoption is a research field on its own. Rates of the Google user-base from 2018 show that less than 10% have MFA enabled [40]. In response, Google decided to auto-enable MFA for 150 million users in October 2021 [31].

Studies with System Administrators Studies with system administrators as their focus group have investigated different aspects. For example, Xu et al. [68] studied how administrators resolve common "access denied" issues and found that missing feedback can cause trial-and-error approaches. Xu and Zhou [69] surveyed characteristics of common configuration errors in an attempt to support administrators in making fewer errors. Similarly, Dietrich et al. [15], who investigated security misconfigurations, found missing documentation to be one of the root causes. Studies focusing on the update process [35, 39, 54] also find that administrators struggle to find useful information about updates although they perceive them as eminent for solving their tasks. This aligns with our findings of administrators criticizing the lack of information.

Studies analyzing tools used by administrators [33, 37, 55] highlighted the importance of usability as it can have a direct impact on security. This is especially important as administrators may have a technical background, but their mental models can be incorrect [28, 32]. Verdi et al. [60] further confirmed the importance of usability: the networking monitoring tool they analyzed received an average SUS score of 49, and the surveyed administrators complained about missing help and sometimes even failed to complete the provided task. In our study, all participants finished the task. Still, the usability of the tested RBA interface was also not assessed to be perfect. One part of this is whether administrators prefer graphical or command-line interfaces to complete their tasks. Towards this end, Voronkova et al. [61] found that 60% actually prefer a graphical interface.

Table 1: Options to configure risk-based authentication offered by cloud providers and access managers.

	Service	Automated Risk Levels	Behavior Defaults	Behavior Modifiable	Notifications Modifiable	Custom Policies
Cloud Providers	Alibaba Cloud	<i>only internally</i>	-	-	-	○
	Amazon Web Services	low, medium, high	●	●	●	○
	Google Cloud Platform	<i>only internally</i>	-	-	-	●
	IBM Cloud	low, medium, high, very high	●	●	●	●
	Microsoft Azure	no risk, low, medium, high	●	●	-	●
	Oracle Cloud	low, medium, high	-	●	●	●
	Tencent Cloud	-	-	-	-	○
Access Managers	CyberArk	non detected, low, medium, high, undetermined	-	●	●	●
	ForgeRock	-	-	-	-	●
	Illantus	-	-	-	-	●
	Micro Focus	-	-	-	-	●
	Okta	low, medium, high	-	●	●	●
	Auth0 (Okta)	low, medium, high, neutral	-	●	●	●
	OneLogin	0-100	-	●	●	●
	Ping Identity	low, medium, high	-	●	●	●
	Thales	-	-	-	-	●

●: Offers the option, ○: Partially offers the option, -: Does not offer the option.

3 Real-World RBA Systems

In this section, we describe how the risk-based authentication systems of different real-world service providers are implemented and which configuration options they offer. For our analysis, which is summarized in Table 1, we considered five factors. First, we determined which *automated risk levels* are provided by the services, i.e., what are the potential output variables of the function that calculates a risk for a new login. *Behavior defaults* describes, if actions are suggested by the services that should be taken in response to the calculated risk levels, i.e., high risk login attempts are blocked by default. The third factor, *behavior modifiable*, describes if it is possible to modify the actions taken in response to the calculated risk levels. The fourth factor, *notifications modifiable*, considers whether the provider allows administrators to adjust how to inform the user about the actions taken in response, for example, by customizing notifications. Finally, we checked if the service providers allow for *custom policies*, which can be used to implement custom logic, e.g., block certain IP ranges, devices, or users. The results depicted in Table 1 are shown for two groups, *cloud providers* and *access managers*.

Cloud providers offer a range of services to enable customers to move IT infrastructure into their data centers and easily scale services. In contrast, access managers have an intentionally narrow focus on access-related services like identity management and MFA. As such, they close a gap by offering their service to enterprises that are already in the cloud but need features their cloud providers do not offer. To get an overview of a representative group of providers, we consulted the Gartner “Magic Quadrant for Cloud Infrastructure and Platform Services” [5] and the “Magic Quadrant for Access Management” [52].

Alibaba Cloud and Google Cloud Platform do offer RBA only internally, without an option for the customers to configure it. Microsoft Azure provides their customers with four risk levels, allows them to modify the behavior for each of them, and provides a default behavior which blocks all login attempts which are deemed as low, medium, or high risk. Notifications sent to users cannot be modified while custom policies based on various login information like the IP address, device, and the calculated risk level are supported. IBM Cloud offers all checked options, Oracle only does not provide a default behavior. Tencent is the only cloud provider supporting only custom policies based on the IP address, but no automated risk levels or any form of RBA in general. In contrast, most access managers like CyberArk, Okta, Auth0 (acquired by Okta [44]), OneLogin, and Ping Identity support RBA with all the described functionalities. Since they rely on custom configurations, none of them provides a default behavior. ForgeRock, Illantus, Micro Focus, and Thales do not support RBA, yet.

In this study, we decided to focus on Amazon, the market leader in cloud computing according to Gartner [5] and others [11, 51]. We tested the “adaptive authentication” feature from Amazon, which is part of its paid AWS service Cognito [4]. Cognito’s adaptive authentication provides three automated risk levels and a default behavior which is similar to IBM and Microsoft. It also allows to modify this behavior and the sent notifications. Custom policies are supported but only in the form of allow- and blocklists for certain IP ranges. Hence, based on the options it offers, AWS depicts an average representative in the group of cloud providers. To study Cognito’s adaptive authentication interface, we built a self-hosted copy of it. In Section 4.2, we provide a detailed description of the tested interface and all of its components.

4 Method

This section describes our user study design, the tested scenarios, and the recruitment process and discusses our ethical considerations and the limitations of our findings.

4.1 Study Structure

The study was designed as an online study due to the ongoing COVID-19 pandemic and to facilitate the participation of an international audience. Prior to the main study, we conducted a pilot study with four participants to ensure that the procedure works as intended. The study, which was offered in both English and German, was split into two parts. First, we sent participants a link that led them to a website hosted on our servers where they configured a risk-based authentication system using an interface based on AWS. Afterward, they answered 26 multiple choice questions. For this first part, we observed a mean completion time of 11 minutes. In the second part, we conducted an interview, which took 33 minutes on average. *Zoom* was used throughout the study with no interaction except for a short introduction during the first part. We decided not to record the hands-on task to prevent participants from feeling monitored and avoid influencing them. Below, we outline the general structure of both parts. For a detailed description, please refer to Appendix A and B.

Part 1: Hands-on Task

1. *Agenda*: After welcoming the participants via *Zoom*, we briefly summarized the structure of the study and provided them with the link to the first part. We also told participants that they could seek our help at any time during the study. Still, we asked them to only do this if they do not know how to continue and that they should rather approach and solve the task like any other task they would get at work.
2. *Consent Form*: The first page on the website contained the consent form, which contained all the basic information about the study and informed participants that they could withdraw from the study at any time.
3. *Scenario*: After consenting, participants saw information about the fictitious company *MediaShop Corporation*, which they should imagine working for, and an email from their supervisor telling them about their task. This information changed depending on the scenario (see Section 4.3).
4. *Configuration*: Using a configuration interface, participants configured the risk-based authentication (depicted in Figure 2 in Appendix D). The upper settings specified a behavior for each of the three risk levels and whether or not a notification should be sent to the user. Below, the participants could adjust the wording of the notifications. We describe this interface in more detail in Section 4.2.
5. *Usability*: After the configuration, participants filled out the 10 items of the System Usability Scale (**SUS1–SUS10**).

To ensure the quality of the data, we also included an attention check (**AC**) which all 28 participants passed.

6. *Security Knowledge*: To assess the participant's security knowledge, we asked a variant of the *Web-use Skill Measure* [25], which we expanded using common security terms from the NCSC glossary [43].
7. *Demography*: The first part concluded with the demographics (**D1–D6**). In addition to basic personal information, we also collected information about participant's employment, including their current job title, work experience, and the size of the company they work for.

Part 2: Interview

1. *Introduction*: We started the second part by describing the general outline of the interview. We highlighted that there are no wrong or right answers, and we are solely interested in perceptions and opinions. We also asked if we were allowed to record the interview. All participants agreed.
2. *Warm-up*: The interview started with two questions (**Q1 & Q2**) about the participants' job to allow them to familiarize themselves with the situation. We also used these questions to double-check participants' eligibility.
3. *Risk Level Configuration*: Questions **Q3** to **Q8** covered the part of the configuration which defines the behavior for the risk levels. We asked about the reasoning for the chosen settings and if there were any difficulties. Participants who clicked on the link to the info page were asked about their reasons and whether or not the page was helpful.
4. *Notification Wording*: We now focused on the wording of the notifications. Questions **Q9** to **Q13** were similar to the previous ones and covered the reasoning, potential issues, and any consulted help.
5. *Risk-based Authentication*: After asking participants about their settings, we intended to learn about general aspects in regard to risk-based authentication. First, we asked participants how they incorporated the scenario to understand how it affected their settings (**Q14**). Afterward, **Q15** focused on prior experience with such notifications and if it may have played a role during the configuration. This question was added after the pilot study, where three of four participants mentioned this aspect without being specifically asked about it. We concluded this block with question **Q16** about any prior experiences with risk-based authentication.
6. *Improvements*: For the last set of questions (**Q17–Q21**), we shifted the focus back to the system participants have used to make their settings. We asked participants to assess the offered granularity of the options, potential obstacles, as well as the most positive and most negative aspects of the system. Finally, we let participants describe how the system would look like if they could change it in any way.
7. *Debriefing*: We finished the interview by answering any final questions the participants had and explained the background of the study. As part of this, we also showed participants the original system, which is part of AWS.

4.2 Configuration Interface

The central aspect of the first part of the study was the configuration of the RBA system. The user interface for this can be seen in Figure 2 in Appendix D. It consists of two components, a decision matrix defining the behavior according to the risk levels and text boxes to customize notifications. The layout of this interface is modeled after the risk-based authentication system of AWS Cognito (cf. Section 3). All aspects of the risk level and notification configuration match the Cognito interface, including texts, links, tooltips, help pages, and the overall design. We only removed the configuration of the `From` and `Reply-To` email addresses, as well as the allow- and blocklists for certain IP ranges, because we wanted to focus on adjustments which are made based on personal experience and judgement.

Risk Level Configuration. The decision matrix maps the three risk levels (*low*, *medium*, *high*) to one of four actions (*allow*, *optional MFA*, *require MFA*, *block*) and a binary decision depicting whether or not the user should be notified. If a risk is set to *allow*, any correct login the system assigns to this risk level will be granted. If set to *optional MFA*, users who have set up a second factor will be challenged to provide it. For users who have not registered a second factor, the system will continue without a challenge, i.e., the login flow is identical to *allow*. If the behavior for a risk level is set to *require MFA*, users have to provide a second factor; users who have not registered a second factor are blocked. Similarly, *block* prevents all logins. The default setting, which we adopted from AWS, allows low risk logins, whereas medium and high risk are set to *optional MFA*. Notifications are sent in all three cases. In addition to the general description of the matrix, a link to a page with further information about risk-based authentication is provided. This page is again a copy of the documentation AWS provides and contains information for each of the four behaviors and the feature that the user can be notified.

Notification Configuration. By default, AWS sends a notification email after every login attempt to the user. A login is registered after entering the correct username and password and pressing the login button, independent of the successful login and risk-level configuration.

On AWS, as well as in our user study, text boxes allow to modify the subject and the body for these notifications for each of the three risk level outcomes: (1) login is allowed, (2) MFA is required, and (3) login is blocked. Note, *optional MFA* is covered by either the notification for allowed logins or those that require MFA. For the default notifications, the email subjects for allowed and MFA logins are both set to “New login attempt” while “Block login attempt” is used for blocked logins. The body of the default notifications is shown in Listing 1 and only differs in the first sentence, which describes the risk level outcome. For example, for allowed

logins the sentence is: “*We observed an unrecognized sign-in to your account with this information.*” The rest of the text includes the login time, device name, and location. The notification also instructs the user to change their password and click a link if they do not recognize the login. The email also includes another link that a user can (optionally) visit to tell the system that the login was legitimate. An administrator can add or remove template placeholders variables like `{city}` from a predefined list that can be found in the official AWS documentation [3]. To mimic this behavior, we also included a link to a self-hosted version of this message template page and observed if the participants visited it.

Listing 1: Default RBA notification message.

```
<risk level outcome>
Time: {login-time}
Device: {device-name}
Location: {city}, {country}
If this login was not by you, you should change
your password and notify us by clicking
on {one-click-link-invalid}.
If this login was by you, you can follow
{one-click-link-valid} to let us know.
```

4.3 Scenarios

We used four real-world scenarios with varying focuses to cover different circumstances system administrators may face, how they affect the configuration of the RBA, and if the tested system allows administrators to configure RBA in situations with varying requirements. Without knowing that there were four different ones, each participant randomly saw one scenario before the configuration phase. Please refer to Appendix A for the exact wording used in each scenario.

Neutral (N): In this scenario, participants were told that they are the system administrator of the *MediaShop Corporation*, where they are responsible for the online shop hosted at *dresscode.com*. An email from their supervisor Jo further informs them that it is their task to complete the configuration of the risk-based authentication.

Security (S): The background information given in this scenario is identical to the neutral scenario with one exception: the supervisor mentions a recent hack in an email that emerged from a password reuse attack. To prevent similar incidents in the future, risk-based authentication should be set up.

Usability (U): This scenario is again based on the neutral one. The only difference is given in an email where the supervisor highlights that customers should not be annoyed by the introduction of the RBA.

Neutral In-House (NI): Unlike the first three cases, participants in this scenario were not told that they administrate the online shop but “the login system ‘VPN-Guard’ that the employees use to work from home.” Apart from this, the scenario is similar to the neutral one in that it does not introduce any focus on security or usability.

Table 2: Demographic information of participants ($n = 28$).

Age		Gender	
Minimum	30	Female	2
Maximum	55	Male	26
Median	40		
Degree		Experience	
High School	5	2–3 years	3
Training	9	4–5 years	3
Bachelor’s	8	6–10 years	3
Master’s	6	11–15 years	10
		>15 years	9
Residency		Company	
Germany	17	10–49 employees	4
USA	6	50–250 employees	5
Other	5	>250 employees	19

4.4 Recruitment and Demographics

The recruitment for our study targeted a special audience in the form of system administrators. On top of that, we conducted a qualitative study with an expected duration close to an hour which we assumed would further reduce the willingness to participate. Hence, we utilized multiple channels to get in contact with potential candidates and shared the information to the study on *LinkedIn*, the German pendant *XING*, the subreddits *r/sysadminjobs*, and *r/SampleSize*, as well as personal contacts in industry. We decided not to require prior experiences with RBA to include participants who have not worked with such a system but potentially could in the future. To also include those where sysadmin tasks only make up a certain part of their daily job, which often applies to small companies, we only required participants to work at least partially in the field of system administration. In cases where the background of the participants was not obvious to us, we asked for additional information, e.g., their *LinkedIn* profile.

We recruited a total of $n = 28$ participants for the study through the described channels. While saturation was reached after 21 participants, we decided to conduct the already scheduled seven additional interviews. The study took place in December 2021 and lasted 48 minutes on average. Each participant received a \$45,- Amazon voucher as compensation. The demographics of the participants are shown in Table 2. Participants were between 30 and 55, with 40 years being the average. In terms of the gender distribution, we anticipated a shift towards male-identifying participants and tried to mitigate this by proactively contacting persons with other identities. Still, we ended up with a majority (26; 93%) who identified as male; we note this in our limitations section. Most participants resided either in Germany (17; 61%) or the United States (6; 21%). The distribution of degrees was more equal, ranging from 18% for high school to 32% for training, with the latter being the typical degree for system

admins in Germany. Two-thirds of the participants (19; 68%) have worked as a system admin for at least 11 years and work in a company with more than 250 employees.

To assess the participants’ security knowledge, we asked them to rate their familiarity with 9 security related items. The basis for this scale is the Web-use skill Measure [25], which we expanded with terms from the NCSC glossary [43]. The results of this assessment are shown in Table 4 in Appendix C. Overall, we observe high ratings ranging from 4.5 to 4.8; a Cronbach’s α of 0.80 indicates a good level of internal consistency. The term *challenge response* is the only outlier (3.9), suggesting a slightly lower understanding of this term. Still, a composite score of 4.6 demonstrates a high familiarity with security-related terms and confirms our expectations since all participants have a strong background in IT.

4.5 Ethical Considerations

Our institution does not have an Institutional Review Board (IRB) governing this kind of study. Still, we ensured that our study would meet all requirements for such an approval, e.g., participants were told upfront about the study procedure, had to actively consent to participate, and were able to withdraw at any time. To further ensure the ethics of our research, we designed it to conform to the principles described in the Menlo Report [59] and stored all data in accordance with the General Data Protection Regulation (GDPR) [18].

4.6 Limitations

We planned our study to provide a high level of ecological validity, still, there are some limitations which we note in this section. First, our demography is shifted towards male-identifying participants despite our efforts to proactively recruit a diverse sample. Still, the distribution of system administrators is disparate in general: according to the German Federal Employment Agency only 11% of currently employed system administrators identify as female [21], the U.S. Bureau of Labor Statistics puts this proportion at 17% [58].

Secondly, participants mostly resided in Germany and the USA which can be attributed to our recruiting channels. We were not able to observe any differences in the responses across the described demographics, yet, our findings may not be representative for all system administrators.

In terms of the framing and the context of the study, we are limited by the fact that participants configured the risk-based authentication for a fictional company. Hence, participants did not have to fear any negative implications, e.g., due to potentially insecure settings and may have not taken the task as serious as if they would have configured a real-world system. Still, we believe that the insights we got are valid as they align across the group of participants. Moreover, during the interview some participants even described that they spent

minutes to think about additional changes they could make on the configuration page but finally continued without any.

Finally, we studied the interface of AWS Cognito, which is only one of several available RBA systems. Thus, all findings apply primarily to AWS, and future research is needed to generally confirm them. However, as shown in Section 3, the solutions have many commonalities, so certain findings are applicable across them. For example, four services, including AWS, use the three risk levels *low*, *medium*, *high*. In response to Q16, seven participants also confirmed that they have worked with a similar solution before.

5 Results

We now present the results of our study, concentrating on how administrators configure risk-based authentication. Table 3 provides an overview of risk level behavior and notification configurations administrators chose in the first part of our study. We start with presenting configurations for each of the two blocks, followed by analyses of participants' reasoning behind the configuration based on the interviews during the second part of the study. Responses in these interviews were separately labeled by two coders who then met to resolve differences and create the codebook. An extended version of this work with the full codebook is available online [38].

5.1 Risk Level Configuration

In the default configuration, low-risk logins are always allowed. For medium- and high-risk, the user is prompted to confirm the login with MFA, if it is activated for their account (*optional MFA*). By default, there is no enforcement of multi-factor authentication, nor are any login attempts blocked.

5.1.1 Configured Risk Level Behavior

Participants' risk level behavior configurations are summarized in the first block of Table 3. Overall, only one participant (N-P6) went with the defaults here. All others configured stronger measures for at least one of the three levels. Low-risk login behavior was changed by 19 participants, most of whom selected *optional MFA*; six even increased the measures to *require MFA*. For medium-risk logins, 23 overruled the default risk level behavior (*optional MFA*) and required multi-factor authentication instead. All participants who made changes chose a stronger option for high-risk logins: 17 participants required MFA for such login attempts, 10 chose to block them. In total, 11 participants selected a configuration with incrementally stronger measures on each risk level.

Figure 1 provides an overview of the risk level behavior configuration for all three risk levels separately for our four studied scenarios. Although our study was designed for an in-depth qualitative analysis, and the group sizes do not allow conclusions about (significant) differences between the

Table 3: Summary of RBA configurations. See Table 5 in Appendix C for the configuration made by each participant.

		<i>Risk Level Behavior</i>			
		Allow	Optional MFA	Require MFA	Block
<i>Risk Level</i>	Low	9*	13	6	0
	Medium	0	5*	23	0
	High	0	1*	17	10
		<i>Notification Configuration</i>			
		Do Not Notify	Notify		
<i>Risk Level</i>	Low	7	21*		
	Medium	2	26*		
	High	1	27*		

* Default

groups, we can still observe a couple of interesting tendencies. For low-risk login attempts, the usability scenario is the only one in which none of the participants required multi-factor authentication. For login attempts classified as high-risk, more than half of the participants of the security scenario configured blocking, which is more than in any other scenario.

On the opposite, four participants who were all in one of the two *neutral* scenarios (see Table 5 in Appendix C) configured the same behavior for all three risk levels (*require MFA*).

5.1.2 Rationale Behind Configuration

When participants were asked to explain the rationale behind the configurations they made (Q4), the reasons of 14 participants revolved around multi-factor authentication and when to activate it. Six participants stated to always require MFA regardless of the risk levels. For two of them, N-P5 and NI-P5, security was a key factor for their MFA configuration. Both of them referred to the ease of use of multi-factor authentication and did not see it as a burden for their users.

"I chose to require MFA because from my experience, users don't find it that hard to use, and it really increases the security. So that's why I chose that for everyone, not just for low and medium risk." (N-P5)

Participants' personal attitudes also played a role among those requiring MFA, e.g., N-P7 expressed to be generally cautious in the light of any type of risk.

"As soon as it's a risk, I want to require MFA." (N-P7)

Two participants said they would always *offer* MFA to the users of their system (*optional MFA*) because they preferred MFA in general but refrained from requiring it due to the context being an online shop. They mainly pointed out that an online shop application was less sensitive than other systems.

"[...] it is dresscode.com, had it been my bank, maybe blocked would be more prudent." (N-P1)

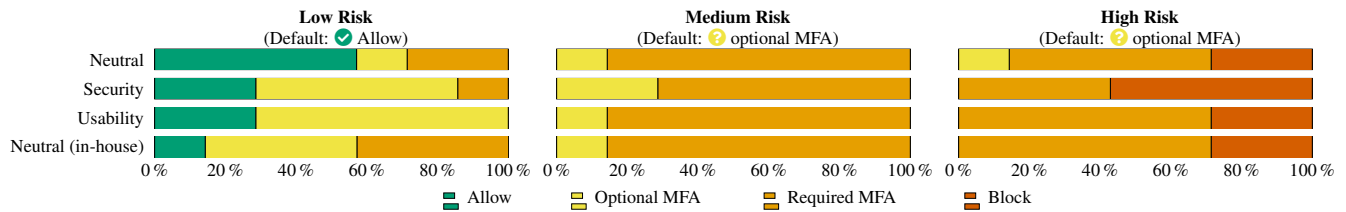


Figure 1: Overview of the risk level behavior configuration. For all risk levels, participants tend to increase the default provided by AWS. In the neutral scenario, participants chose a less strict configuration, especially in contrast to security and in-house.

The remaining participants whose justification involved MFA, basically mentioned medium or high risk to be appropriate for requiring multi-factor authentication.

In total, 11 participants have configured RBA with stronger measures for each risk level (see Section 5.1.1). For 10 of them, this incremental increase was the justification for their configuration, i.e., they wanted a stronger requirement the higher the risk was classified.

User experience when using the system was named by four participants being a reason for their configuration. This aspect is most likely connected with the considered application being an online shop, since user experience was mostly viewed in the light of customer satisfaction. That is, these participants were rather careful in bothering users with MFA or even blocking access since they feared disadvantages for their business when users preferred their less intrusive competitors.

“Blocking is of course extremely invasive. I mean, I would bounce our customers and we don’t want that. Maybe they go to a competitor.” (N-P3)

Six participants mentioned examples, e.g., situations which they had experienced before, that represent triggers for RBA events. These situations include login attempts from new geographical locations ($n = 5$), e.g., in the case of travel, and logins from previously unknown devices ($n = 3$). Participants used such examples to make risk level assessments for login attempts more tangible and reasoned what action they would require. Therefore, their configurations likely incorporate realistic scenarios that are relevant in the context but may also involve a risk of being too narrowed to specific anecdotes, losing sight of the broader threat landscape.

Four participants referred to having taken reactions into account they had when experiencing real-world RBA systems. While three of them mentioned their own experience from a user’s perspective for services such as Netflix or PayPal, participant N-P4 stated to have followed the practice of their own company from the administrator side.

“When choosing the settings, I more or less followed the way we do it at ours [company]. For example, we aim to protect external access with MFA.” (N-P4)

5.1.3 Obstacles in the Configuration

Q6 to Q8 were designed to capture obstacles participants faced during configuration and if and how they solved them. Some difficulties already became apparent when participants explained their choice in Q4. Six participants misunderstood the *optional MFA* setting when configuring the risk level behavior. For example, S-P7 interpreted optional as a decision that can be made by users in their account settings.

“I have interpreted this so that the user can decide whether they want to use it or not, so that they specify this somewhere in the settings beforehand, whether they want it or not. As a result, users can also control how secure they want to be.” (S-P7)

Four participants misunderstood the concept of risk levels which became apparent when, e.g., participant U-P2 referred to different users being categorized as different risk levels. While we must keep in mind such issues when interpreting our participants’ responses as a whole, we judged that none of the misconceptions qualified for invalidating entire responses.

Eight participants mentioned that being unsure about the risk level computation affected their choice (Q4). This is consistent with responses to Q6, in which the same participants named the unclear functionality of the levels a difficulty.

Further issues include missing specific descriptions of individual items ($n = 4$), and missing options for the risk level behavior configuration ($n = 3$). As an example, N-P4 asked for the ability to configure an MFA method (e.g., enforcing the use of a security key) to be required for confirming the login with MFA. S-P2 mentioned the lack of a test environment to simulate their configuration from a user’s perspective.

Missing information about specific items is also reflected in the use of the provided help pages (Q7). Out of 15 participants who clicked on the help link, six participants responded they were looking for information about the risk level behavior, five participants searched for information about how the risk levels work. The remaining four participants accessed the help page out of curiosity for no specific reason.

Finally, responses to Q8 indicate that the level of information provided in our study was largely appropriate and complete. Only two participants mentioned they used external help (Google and Wikipedia) for rather small issues, and the remaining 26 participants did not use any sources of information from outside our study.

5.2 Notification Configuration

By default, AWS sends a notification after every login attempt, independent of a successful login and risk-level configuration. The second block of Table 3 provides an overview of the changes to the default notification configuration.

5.2.1 When to Notify

Overall, 20 participants have not changed the defaults suggested by AWS when to notify the user. Seven participants turned off the default notifications for low-risk sign-in attempts, two of which also turned off notifications for medium-risk attempts. Most notably, one participant U-P4 turned everything upside down and opted not to notify the user for high-risk login attempts but for the two lower risk levels.

Their preference not to annoy users in the case of a negligible risk along with the danger of notification fatigue motivated seven participants to disable the notification email for low-risk login attempts:

“If you get bombarded with login notifications you get annoyed. [...] why would you look at the high risk notification unless you make it screaming? So I chose to only notify when there’s a reason.” (N-P1)

Only two of these seven participants allowed low-risk logins to proceed without MFA, while four configured optional MFA. NI-P7 even required MFA for such low-risk logins.

Out of the two participants who disabled the email even for medium-risk logins, both configured MFA to be required. The participant who turned off notifications for high-risk login attempts assumed high-risk logins to originate from “hijacked” accounts. Thus, an attacker might be able to fool the system by clicking a link in the email to report that the login was legitimate (cf. Section 5.2.2). However, they did not go into detail how such accounts could be recovered:

“I don’t know if I’m giving away information there. If I have a hijacked account, and I send a notification, which the attacker can get and click—‘Yes, it’s really me.’—How it goes on then?” (U-P4)

Interestingly, a similar scenario is mentioned by Google in a talk by Grzegorz Milka [40], where immediately deleting the “Security alert: A new login on . . .” notification, might cause an increase of the security risk score.

Across scenarios (i.e., focus on security or usability) one can observe a tendency towards sending more notifications in the security scenario, and less in the usability-focused scenario. However, due to the quantitative focus of the study, no statistical significant difference can be observed.

5.2.2 Content and Wording

The default notification text, which slightly differs by the risk level outcomes, can be found in Listing 1. All emails are also depicted in full length in Figure 2 in Appendix D.

We observed 12 participants who decided not to change the default notification or its subject. Reasons for not changing the text are either the notification being similar to those sent by popular service providers or the default is seen as sufficient in the amount of detail it contains:

“I found the mail to be basically fine. Of course you can still customize it individually, but in the end, the users get the information they need.” (N-P6)

N-P4 also gave an additional justification for not touching the notification text, namely, the fear that a change will likely cause a lot of issues in future updates:

“I know from experience that if you put software somewhere and tinker with it, it will break by the third update at the latest. [...] Especially when working with placeholders, things go wrong so easily.” (N-P4)

In contrast, 16 participants decided to change the text. The considerations when changing or tweaking the default template include: (1) adding details (e.g., username or IP), (2) improving the wording, (3) adding context (e.g., shop name), (4) preventing phishing, and (5) a distrust in the location.

One participant acknowledged that designing such notification requires a lot of time and effort and might also involve other departments and some testing.

“I’m trying to make it understandable, which can be a challenge, so in real life, I probably would have spent more time and also work with the communications people and tested it.” (N-P1)

Add Details. Noteworthy, eight participants considered adding more details to be important. Most often, participants wanted to add the following: the username to increase trust by addressing the receiver individually, the IP address or event ID, in both cases, to enable easier debugging, and some form of contact information to support the user.

“It is important to have an event ID so you can assign it afterwards.” (NI-P1)

Of course, the details participants added are influenced by the template placeholders that AWS lists in the official documentation [3]. It was accessed by eight participants of which five added details. An additional three added details but did not check the documentation and even one participant who decided against changing the notification suggested the importance of providing a lot of details.

Improve Wording. Overall, four participants noted the importance of changing the wording of the message. Here, the motivation was either to make sure the notification is understandable or to highlight certain aspects as NI-P4 describes:

“[...] I just made it a little more urgent, saying ‘hey, you have to do something’ [...]” (NI-P4)

Add Context. In total, three participants remarked (depending on their scenario) the importance of context in the email subject and/or the body. For example, N-P1 who changed the subject to “New login attempt to dresscode.com” said:

“I added some context, that it was from dresscode.com in the subject, so it stands out a little bit more.” (N-P1)

On the other hand, NI-P2 who changed the intro of the notification to “We observed an unrecognized VPN sign-in attempt” explained the motivation as follows:

“I can imagine the MediaShop has many different types of accounts and systems. [...] But here we’re specifically talking about the VPN. So that’s why I narrowed in on that.” (NI-P2)

Prevent Phishing. Interestingly, three participants were concerned about phishing, suggesting to remove the two hyperlinks and increase trust by adding the username.

“Because normal phishing emails just go out without your username.” (NI-P2).

Location Distrust. Finally, two participant wanted to add the word “Approximate” in front of the word “Location”. They explained that IP-based geolocation cannot be trusted.

“The location is never 100% accurate. That database changes far too often, and it can be changed arbitrarily. Sometimes, when I have a new IP, it goes back to somewhere in Kansas or whatever the center point of America is. So the word ‘approximate’ is important.” (S-P5).

5.3 Other Influential Factors

There are several additional factors that may have influenced participants’ RBA configurations. In this context, we are particularly interested in effects of the scenario itself (Q14), and participants’ prior experience with RBA, both from a user’s (Q15) and administrator’s perspective (Q16).

Incorporating the Scenario. When we asked participants whether they had incorporated the scenario, 16 participants stated they had done so, whereas 12 had not considered it. Among the participants who considered it, eight described that they had considered the context of a company with an online shop more generally. Four participants stated that they made a trade-off weighing the security of the online shop and its usability when configuring the RBA settings.

“When you have an online shop, you have lots of customers so it’s a balance [...] you always want to have this nice and easy experience, but at the same time you want to protect the customers.” (S-P2)

Another four participants considered the scenario when making the configurations but at the same time admitted they would have requested additional information in a real-world setting. However, S-P7 further added that even then the decision to deploy RBA would probably not have been overruled.

“I might have asked if it was certain that it really was a hack. But let’s put it this way, if the boss says turn it on, then you turn it on.” (S-P7)

From those participants who did not incorporate the scenario, the vast majority stated to have followed a rather gen-

eral approach that was not influenced by specific properties of the described scenario ($n = 10$). Two participants explained that they used experience from their current job as a background to configure the RBA appropriately.

Previous Experience with RBA. In the pilot study, three of the four participants mentioned that they followed a login notification they received, without being specifically asked about it. Hence, we decided to ask participants if their approach was similarly influenced by such real-world notifications; 22 confirmed while 6 negated. Of the former, 16 participants described that the information in the notification text should reflect the information present in real-world notifications.

“I actually think that Facebook does a pretty good job of these. If I remember correctly, their emails look a lot like this and include most of these things, you know, time, device, location.” (NI-P2)

Five participants emphasized that their configuration, i.e., the behavior in response to the risk level, was chosen such that it matches services they use.

A different aspect not directly related to the configuration, but still highlighted by five participants, is the abuse of such notifications for phishing. This risk is further enabled by the fact that even legitimate notifications, like the default text used by AWS, contain links. As we could already observe in Section 5.2.2, some participants tried to mitigate this, e.g., by removing the links. A second challenge, described by two participants, is the risk of notification fatigue caused by login notifications being sent too often.

Regarding the administrator’s perspective, 16 participants did not have experience with RBA systems before. From the remaining 12 participants who already had such experiences, seven stated that the system they worked with was similar to the one used in our study. While none of them worked with AWS, we had participants who worked with Microsoft Azure that offers a similar level of detail. In contrast, five participants reported differences, most of which were subject to variations in the levels of detail, such as the way how different risk levels are presented.

5.4 Using the System

We used the System Usability Scale (SUS) to assess the usability of the RBA system in our study. The mean score across all participants was 75 ($SD = 13$), i.e., “above average” usability (>68). Still, this shows that there is room for improvement. Hence, we will now provide insights into participants’ feedback on using the system and investigate which aspects are already satisfying and which can be improved.

Generally speaking, 13 participants rated the settings options as overall sufficient. While most responses to Q17 remained rather unspecific, five participants appreciated the simplicity of the settings, and two emphasized that the configuration granularity was a good fit for the scenario showcasing a small business environment. Simplicity aspects were again

referred to when we asked participants what they remembered most positively about the system (Q20). Here, simplicity was named 14 times in different flavours, often in conjunction with clarity of how settings were presented ($n = 7$). Other positive aspects concerned certain features ($n = 7$), e.g., the tooltips for optional and required MFA, and that settings can be adjusted to the context of the scenario ($n = 4$).

On the downside, 18 participants missed certain items in the settings (Q17). Note that the total number of mentions is larger than the number of participants as they could rate the options as collectively sufficient and at the same time state that they were missing something. Of the 12 participants who preferred to have more actions in response to risk levels (Q17), seven declared that this circumstance hindered them from configuring the RBA settings the way they wanted (Q18). In response to Q19, the same participants mentioned this lack as the most negative aspect of the system. Seven participants referred to missing descriptions when asked about obstacles. For four of them, this was the most negative aspect.

When we asked participants what they would change and how a perfect system would look like (Q21), 10 wished for adjustable risk levels. Five participants wanted to be able to configure multi-factor authentication in more detail. These responses are largely in line with comments to previous questions, e.g., with participants demanding the ability to further specify the MFA requirements (cf. Section 5.1.3). Some participants also asked for certain features, including a monitoring solution on the administrator side ($n = 4$) and a preview function of the final notification ($n = 3$).

6 Discussion

Overall, we identified several issues with the RBA system of AWS concerning key aspects like the meaning of risk levels and the configuration interface. Moreover, we saw a tendency to increase the defaults and observed a basic intuition for usability requirements. In the following, we like to discuss the implications of our findings in more detail and how they apply to Amazon Cognito and RBA systems in general.

6.1 Risk Levels

Most prominently, we highlight the need for a clear description of the risk levels in an RBA system and how many different levels there are. AWS's interface allows defining actions for three risk levels (*low*, *medium*, *high*). However, a fourth outcome is that the system assesses the login as “not risky at all” and does not enforce any additional security mechanisms. IBM, Microsoft, and CyberArk prevent this confusion by making this lowest risk level part of the configuration.

Second, administrators demand insights into the calculation of the risk levels, arguing that it is crucial for an informed decision. In our study, we saw participants overcoming this problem by guessing how the risk levels work, which may

lead to inaccurate and potentially insecure configurations. Others argued that they must treat all levels equally if they cannot distinguish them. This may not lead to an insecure decision, yet it contradicts the initial goal of RBA in limiting security prompts for users. Others emphasized that a thorough description would be a “must-have” when deciding on a solution. Hence, service providers should also be interested in providing a complete and comprehensive documentation.

Third, we observed administrators who wanted to adjust the calculation of the risk levels and configure a more fine-granular behavior. We emphasize that fewer participants brought up this aspect, which appeared to have a more in-depth understanding of RBA. The majority was able to configure RBA according to their needs and emphasized the simplicity of the evaluated system. Hence, service providers who want to offer this feature may want to provide an additional “expert mode”. This mode would allow professionals with special requirements to make more fine adjustments, while others could still use a simpler user interface.

6.2 Interface

The Amazon Cognito interface uses two terms that are crucial but, at the same time, not self-explanatory: *optional MFA* and *block*. The former defines a behavior where users who have MFA enabled are prompted, while users who have not, are still allowed to login. However, nine participants misinterpreted it such that the user is asked during the login whether or not they like to use MFA. Hence, they argued that it cannot prevent an attack because the MFA prompt can simply be skipped, and legitimate users would likely skip it for convenience reasons. We emphasize that hovering over the term “optional MFA” on the configuration interface will display a tooltip with a short explanation, just like on the original AWS implementation. Moreover, the term is also explained in more detail on the provided help page. Regarding the tooltip, none of the nine participants who misunderstood the term noticed the tooltip, as there is no visual indicator present. Seven of those nine participants noticed the information on the help page; the other two did not visit the page. To minimize the risk for misinterpretation AWS should describe the term “optional MFA” more prominently, e.g., as part of the main interface, since it is crucial for a thorough understanding of the configuration.

The term “block” also caused confusion among the administrators. In contrast to optional MFA, the general idea of denying the login was clear to all. However, details of the actual consequence were not. For example, SP-6 extensively reasoned about how long the block will last and whether it is combined with some sort of rate-limiting. The participant concluded that blocking attempts is not an option unless its consequences are fully understood, again highlighting the need for a profound documentation, similar to the risk levels. In contrast to the term “optional MFA”, which is unique to AWS, blocking logins is an option all RBA services pro-

vide. Especially since it is the most invasive outcome, service providers should describe in detail of how it is implemented.

Regarding the template placeholder variables, we had participants who wanted an easier-to-use interface. While some found the approach easy and understandable, others struggled with using the variables surrounded by curly brackets and suggested preferring a drag-and-drop-based solution. Moreover, we observed that one participant misunderstood the `{one-click-link-invalid}` variable and asked why the email should contain a “non-working link.” This also aligns with a statement by N-P4, who describes the granularity of the configuration interface as inconsistent: the risk level behavior is configured via radio buttons while the notification templates can be changed arbitrarily. When providing a single configuration page for both the risk levels and the notifications, as AWS does, one solution could again be an additional expert mode that would enable the use of placeholders. A second solution is to keep the configuration of the risk level and the modification on separate pages; this is what IBM, Oracle, and all access managers do.

6.3 Spicing Up Defaults

Interestingly, only one of the 28 participants did not increase the risk level behavior. It seems that the defaults AWS provides (low risk: *allow*, medium risk: *optional MFA*, high risk: *optional MFA*) are perceived as too lax. Especially 10 participants stand out who blocked access for high-risk logins which could also be caused by false positives, e.g., a login from another country during vacation. A user would have no other option than to contact the helpdesk (or order at another online shop). Moreover, it is distinct that many participants prefer to prompt the user for MFA even for low-risk logins: 19 went with either “optional MFA” or “require MFA”.

Our findings highlight the need for a correctly balanced RBA configuration to be able to increase security while at the same time limiting notifications to a minimum. This is also supported by AWS’s documentation, which recommends keeping “*the advanced security features in audit mode for two weeks before enabling actions*” to observe and train the login behavior before deciding on what to enforce and block [2]. In November 2021, AWS changed its defaults to “block” for all risk levels [1]. This way, enabling and using the defaults is no longer a valid option, potentially leading to more administrators who audit the logins before deciding on any actions.

6.4 Cooperation and Usability

It is pleasant to see that some administrators are aware of usability requirements, e.g., some participants took a moment to consider the impact of their work on the end-user. We noted a preference for easy-to-understand notifications, and a few participants even decided not to send notifications that could be considered unnecessary or unhelpful. While participants’

primary concern was on common tasks in their responsibility like debugging (i.e., adding an event ID), we also observed an awareness to cooperate with other departments, e.g., “*the communications people*”. Ultimately, this might lead to a more secure system. However, such an approach cannot be taken for granted as it is hard to follow for most smaller IT departments. For example, S-P5 summarized that it is most important to minimize the time spent with the configuration: “*you know, my time is forever compromised.*” Hence, it should be the goal to reduce the workload by providing useful default notifications and guidelines.

7 Summary & Future Work

In this study, we investigate how administrators configure risk-based authentication, which issues they face, and how different requirements influence their decisions. Generally, we observed an urge of administrators to increase the default security parameters of RBA systems. We learned that some of these often unnecessary changes are owed to undefined risk levels and confusing wordings like “optional MFA.” As small- to medium-sized enterprises cannot rely on trained specialists, our research reveals the need for easier-to-use configuration interfaces that support administrators in making more informed decisions, e.g., by highlighting the impact of the various configuration options. We observed that administrators are aware of potential usability issues, as some of our participants considered the impact of their work on the end-user. Still, guidance should be provided when possible.

Based on our findings, we identified multiple research directions for the design of RBA systems:

- Defaults are crucial as administrators sometimes struggle to decide which risk level behavior is reasonable and which notifications are necessary. One approach could be to have trained professionals predefine defaults based on the requirements of common scenarios, e.g., online shopping. Similarly, a guided and an expert mode could be developed to allow administrators to customize the settings according to their prior experience and knowledge.
- It needs to be investigated how terms that are open to interpretation, such as “low risk,” “optional MFA,” and “link-invalid” can be explained in a meaningful way.
- Administrators want to understand the implications of their configurations. It could be tested if a simulation that depicts the user’s perspective provides these insights.
- Regarding the notification design, we identified a lack of consensus across participants, suggesting that future work needs to explore how to design RBA notifications, i.e., which information to include.

Acknowledgments

We thank Julian Vogt for his help with the implementation of the study website. We also thank our shepherd and the reviewers for their insightful comments and feedback. This research was supported by the research training group “Human Centered Systems Security” sponsored by the state of North Rhine-Westphalia and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2092 CASA – 390781972.

References

- [1] Amazon Web Services, Inc. Amazon Cognito: Amazon Cognito Launches New Console Experience, November 2021. <https://aws.amazon.com/about-aws/whats-new/2021/11/amazon-cognito-console-user-pools/>, as of June 9, 2022.
- [2] Amazon Web Services, Inc. Amazon Cognito: Developer Guide – Audit Mode for Two Weeks, June 2021. <https://docs.aws.amazon.com/cognito/latest/developerguide/cognito-user-pool-settings-advanced-security.html>, as of June 9, 2022.
- [3] Amazon Web Services, Inc. Amazon Cognito: Developer Guide – Message Templates, June 2021. <https://docs.aws.amazon.com/cognito/latest/developerguide/cognito-user-pool-settings-message-templates.html>, as of June 9, 2022.
- [4] Amazon Web Services, Inc. Amazon Cognito: Developer Guide – Using Adaptive Authentication, June 2021. <https://docs.aws.amazon.com/cognito/latest/developerguide/cognito-user-pool-settings-adaptive-authentication.html>, as of June 9, 2022.
- [5] Raj Bala, Bob Gill, Dennis Smith, Kevin Ji, and David Wright. Gartner Magic Quadrant for Cloud Infrastructure and Platform Services. Report G00736363, Gartner, Inc., July 2021.
- [6] Robert Biddle, Sonia Chiasson, and Paul C. Van Oorschot. Graphical Passwords: Learning from the First Twelve Years. *ACM Computing Surveys*, 44(4):19:1–19:41, August 2012.
- [7] Joseph R. Biden Jr. Executive Order on Improving the Nation’s Cybersecurity, May 2021.
- [8] Joseph Bonneau, Edward W. Felten, Prateek Mittal, and Arvind Narayanan. Privacy Concerns of Implicit Secondary Factors for Web Authentication. In *Who Are You?! Adventures in Authentication Workshop*, WAY ’14, Menlo Park, California, USA, July 2014. USENIX.
- [9] Joseph Bonneau, Cormac Herley, Paul C. Van Oorschot, and Frank Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *IEEE Symposium on Security and Privacy, SP ’12*, pages 553–567, San Jose, California, USA, May 2012. IEEE.
- [10] Joseph Bonneau, Cormac Herley, Paul C. Van Oorschot, and Frank Stajano. Passwords and the Evolution of Imperfect Authentication. *Communications of the ACM*, 58(7):78–87, June 2015.
- [11] Canalys. Global Cloud Services Spend Exceeds US\$50 Billion in Q4 2021, February 2022. <https://www.canalys.com/newsroom/global-cloud-services-q4-2021>, as of June 9, 2022.
- [12] Stéphane Ciolino, Simon Parkin, and Paul Dunphy. Of Two Minds about Two-Factor: Understanding Everyday FIDO U2F Usability through Device Comparison and Experience Sampling. In *Symposium on Usable Privacy and Security, SOUPS ’19*, pages 339–356, Santa Clara, California, USA, August 2019. USENIX.
- [13] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Faith Cranor, and Nicolas Christin. “It’s Not Actually That Horrible”: Exploring Adoption of Two-Factor Authentication at a University. In *ACM Conference on Human Factors in Computing Systems, CHI ’18*, pages 456:1–456:11, Montreal, Quebec, Canada, April 2018. ACM.
- [14] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The Tangled Web of Password Reuse. In *Symposium on Network and Distributed System Security, NDSS ’14*, San Diego, California, USA, February 2014. ISOC.
- [15] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating System Operators’ Perspective on Security Misconfigurations. In *ACM Conference on Computer and Communications Security, CCS ’18*, pages 1272–1289, Toronto, Ontario, Canada, October 2018. ACM.
- [16] Pavni Diwanji. Google: Detecting Suspicious Account Activity, March 2010. <https://security.googleblog.com/2010/03/detecting-suspicious-account-activity.html>, as of June 9, 2022.
- [17] Periwinkle Doerfler, Kurt Thomas, Maija Marincenko, Juri Ranieri, Yu Jiang, Angelika Moscicki, and Damon McCoy. Evaluating Login Challenges as a Defense

- Against Account Takeover. In *The World Wide Web Conference, WWW '19*, pages 372–382, San Francisco, California, USA, May 2019. ACM.
- [18] The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119/1, April 2016.
- [19] Florian M. Farke, Lennart Lorenz, Theodor Schnitzler, Philipp Markert, and Markus Dürmuth. “You still use the password after all” – Exploring FIDO2 Security Keys in a Small Company. In *Symposium on Usable Privacy and Security, SOUPS '20*, pages 19–35, Virtual Conference, August 2020. USENIX.
- [20] David Mandell Freeman, Sakshi Jain, Markus Dürmuth, Battista Biggio, and Giorgio Giacinto. Who Are You? A Statistical Approach to Measuring User Authenticity. In *Symposium on Network and Distributed System Security, NDSS '16*, San Diego, California, USA, February 2016. ISOC.
- [21] German Federal Employment Agency. Employees by Occupation (KldB 2010) – Germany (Quarterly Figures), June 2021. <https://statistik.arbeitsagentur.de/DE/Navigation/Statistiken/Fachstatistiken/Beschaeftigung/Beschaeftigung-Nav.html>, as of June 9, 2022.
- [22] Maximilian Golla, Grant Ho, Marika Lohmus, Monica Pulluri, and Elissa M. Redmiles. Driving 2FA Adoption at Scale: Optimizing Two-Factor Authentication Notification Design Patterns. In *USENIX Security Symposium, SSYM '21*, pages 109–126, Virtual Conference, August 2021. USENIX.
- [23] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. “What was that site doing with my Facebook password?” Designing Password-Reuse Notifications. In *ACM Conference on Computer and Communications Security, CCS '18*, pages 1549–1566, Toronto, Ontario, Canada, October 2018. ACM.
- [24] Paul A. Grassi, James L. Fenton, and William E. Burr. Digital Identity Guidelines – Authentication and Lifecycle Management: NIST Special Publication 800-63B, June 2017.
- [25] Eszter Hargittai and Yuli Patrick Hsieh. Succinct Survey Measures of Web-Use Skills. *Social Science Computer Review*, 30(1):95–107, February 2012.
- [26] Cormac Herley and Stuart Schechter. Distinguishing Attacks from Legitimate Traffic at an Authentication Server. Technical Report MSR-TR-2018-19, Microsoft, June 2018.
- [27] Cormac Herley and Paul C. Van Oorschot. A Research Agenda Acknowledging the Persistence of Passwords. *IEEE Security & Privacy*, 10(1):28–36, January 2012.
- [28] Dennis G. Hrebec and Michael Stiber. A Survey of System Administrator Mental Models and Situation Awareness. In *SIGCPR Conference on Computer Personnel Research, SIGCPR '01*, pages 166–172, San Diego, California, USA, April 2001. USENIX.
- [29] Adam Hurkała and Jarosław Hurkała. Architecture of Context-Risk-Aware Authentication System for Web Environments. In *International Conference on Informatics Engineering and Information Science, ICIEIS '14*, pages 219–228, Lodz, Poland, September 2014. ACM.
- [30] Roger Piqueras Jover. Security Analysis of SMS as a Second Factor of Authentication. *ACM Queue*, 18(4):37–60, August 2020.
- [31] Guemmy Kim. Google: Making You Safer With 2SV, March 2022. <https://blog.google/technology/safety-security/reducing-account-hijacking/>, as of June 9, 2022.
- [32] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. “If HTTPS Were Secure, I Wouldn’t Need 2FA” – End User and Administrator Mental Models of HTTPS. In *IEEE Symposium on Security and Privacy, SP '19*, pages 246–263, San Francisco, California, USA, May 2019. IEEE.
- [33] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. “I Have No Idea What I’m Doing” – On the Usability of Deploying HTTPS. In *USENIX Security Symposium, SSYM '17*, pages 1339–1356, Vancouver, British Columbia, Canada, August 2017. USENIX.
- [34] Leona Lassak, Annika Hildebrandt, Maximilian Golla, and Blase Ur. “It’s Stored, Hopefully, on an Encrypted Server”: Mitigating Users’ Misconceptions About FIDO2 Biometric WebAuthn. In *USENIX Security Symposium, SSYM '21*, pages 91–108, Virtual Conference, August 2021. USENIX.
- [35] Frank Li, Lisa Rogers, Arunesh Mathur, Nathan Malkin, and Marshini Chetty. Keepers of the Machines: Examining How System Administrators Manage Software Updates For Multiple Machines. In *Symposium on Usable Privacy and Security, SOUPS '19*, pages 273–288, Santa Clara, California, USA, August 2019. USENIX.

- [36] Lucy Li, Bijeeta Pal, Junade Ali, Nick Sullivan, Rahul Chatterjee, and Thomas Ristenpart. Protocols for Checking Compromised Credentials. In *ACM Conference on Computer and Communications Security, CCS '19*, pages 1387–1403, London, United Kingdom, November 2019. ACM.
- [37] Salvatore Manfredi, Mariano Ceccato, Giada Sciarretta, and Silvio Ranise. Do Security Reports Meet Usability? Lessons Learned from Using Actionable Mitigations for Patching TLS Misconfigurations. In *Workshop on Education, Training and Awareness in Cybersecurity, ETACS '21*, pages 1–13, Virtual Conference, August 2021. IEEE.
- [38] Philipp Markert, Theodor Schnitzler, Maximilian Golla, and Markus Dürmuth. “As soon as it’s a risk, I want to require MFA”: How Administrators Configure Risk-based Authentication (Extended Version), August 2022. <https://philipp-markert.com/assets/papers/soups22-48-rba-admin.pdf>, as of June 9, 2022.
- [39] Florin Martius and Christian Tiefenau. What Does This Update Do to My Systems? – An Analysis of The Importance of Update-Related Information to System Administrators. In *Workshop on Security Information Workers, WSIW '20*, pages 1–12, Virtual Conference, February 2020. USENIX.
- [40] Grzegorz Milka. Anatomy of Account Takeover. In *USENIX Enigma Conference, Enigma '18*, Santa Clara, California, USA, January 2018. USENIX.
- [41] Katharine Murphy. Google Detecting 18 Million Malware and Phishing Messages per Day Related to COVID-19, July 2020. <https://www.theguardian.com/australia-news/2020/jul/14/google-detecting-18m-malware-and-phishing-messages-per-day-related-to-covid-19>, as of June 9, 2022.
- [42] National Cyber Security Centre. Cloud Security Guidance: Identity and Authentication, November 2018. <https://www.ncsc.gov.uk/collection/cloud-security/implementing-the-cloud-security-principles/identity-and-authentication>, as of June 9, 2022.
- [43] National Cyber Security Centre. NCSC Glossary: Definitions for Common Cyber Security Terms, December 2021. <https://www.ncsc.gov.uk/information/ncsc-glossary>, as of June 9, 2022.
- [44] Okta, Inc. Okta Completes Acquisition of Auth0, May 2021. <https://www.okta.com/press-room/press-releases/okta-completes-acquisition-of-auth0>, as of June 9, 2022.
- [45] Bijeeta Pal, Mazharul Islam, Marina Sanusi, Nick Sullivan, Luke Valenta, Tara Whalen, Christopher A. Wood, Thomas Ristenpart, and Rahul Chatterjee. Might I Get Pwned: A Second Generation Compromised Credential Checking Service. In *USENIX Security Symposium, SSYM '22*, Boston, Massachusetts, USA, August 2022. USENIX.
- [46] Christopher Palow. After Watching This Talk, You’ll Never Look at Passwords the Same Again, November 2013. <http://www.meetup.com/HNLondon/events/150289672/>, as of June 9, 2022.
- [47] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let’s Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *ACM Conference on Computer and Communications Security, CCS '17*, pages 295–310, Dallas, Texas, USA, October 2017. ACM.
- [48] Esteban Rivera, Lizzy Tengana, Jesús Solano, Alejandra Castelblanco, Christian López, and Martín Ochoa. Risk-Based Authentication Based on Network Latency Profiling. In *ACM Workshop on Artificial Intelligence and Security, AISec '20*, pages 105–115, Virtual Conference, November 2020. ACM.
- [49] Kevin Shalvey. A Hacker Stole More than \$55 Million in Crypto after a bZx Developer Fell for a Phishing Attack, November 2021. <https://www.businessinsider.com/hacker-steals-55-million-in-crypto-after-bzx-phishing-attack-2021-11>, as of June 9, 2022.
- [50] Alex Simons. Azure AD Identity Protection: Risk-Based Conditional Access Policies, March 2016. <https://techcommunity.microsoft.com/t5/azure-active-directory-identity/azure-ad-identity-protection-is-in-public-preview-whoop-whoop/ba-p/244242>, as of June 9, 2022.
- [51] Synergy Research Group. As Quarterly Cloud Spending Jumps to Over \$50B, Microsoft Looms Larger in Amazon’s Rear Mirror, February 2022. <https://www.srgresearch.com/articles/as-quarterly-cloud-spending-jumps-to-over-50b-microsoft-looms-larger-in-amazons-rear-mirror>, as of June 9, 2022.
- [52] Henrique Teixeira, Abhyuday Data, and Michael Kelley. Gartner Magic Quadrant for Access Management. Report G00740722, Gartner, Inc., November 2021.
- [53] Kurt Thomas, Jennifer Pullman, Kevin Yeo, Ananth Raghunathan, Patrick Gage Kelley, Luca Invernizzi, Borbala Benko, Tadek Pietraszek, Sarvar Patel, Dan

- Boneh, and Elie Bursztein. Protecting Accounts From Credential Stuffing With Password Breach Alerting. In *USENIX Security Symposium*, SSYM '19, pages 1556–1571, Santa Clara, California, USA, August 2019. USENIX.
- [54] Christian Tiefenau, Maximilian Häring, Katharina Krombholz, and Emanuel von Zezschwitz. Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators. In *Symposium on Usable Privacy and Security*, SOUPS '20, pages 239–258, Virtual Conference, August 2020. USENIX.
- [55] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right. In *ACM Conference on Computer and Communications Security*, CCS '19, pages 1971–1988, London, United Kingdom, November 2019. ACM.
- [56] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *ACM Conference on Computer and Communications Security*, CCS '13, pages 161–172, Berlin, Germany, November 2013. ACM.
- [57] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. “I Added ‘!’ at the End to Make It Secure”: Observing Password Creation in the Lab. In *Symposium on Usable Privacy and Security*, SOUPS '15, pages 123–140, Ottawa, Ontario, Canada, July 2015. USENIX.
- [58] U.S. Bureau of Labor Statistics. 11. Employed Persons by Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity, January 2022. <https://www.bls.gov/cps/cpsaat11.htm>, as of June 9, 2022.
- [59] U.S. Department of Homeland Security. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, August 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/, as of June 9, 2022.
- [60] Fábio Luciano Verdi, Hélio Tibagi de Oliveira, Leobino N. Sampaio, and Luciana A. M. Zaina. Usability Matters: A Human-Computer Interaction Study on Network Management Tools. *Transactions on Network and Service Management*, 17(3):1865–1874, September 2020.
- [61] Artem Voronkov, Leonardo A. Martucci, and Stefan Lindskog. System Administrators Prefer Command Line Interfaces, Don't They? An Exploratory Study of Firewall Interfaces. In *Symposium on Usable Privacy and Security*, SOUPS '19, pages 259–271, Santa Clara, California, USA, August 2019. USENIX.
- [62] Stephan Wiefeling, Markus Dürmuth, and Luigi Lo Iacono. Verify It's You: How Users Perceive Risk-based Authentication. *IEEE Security & Privacy*, 19(6):47–57, November 2021.
- [63] Stephan Wiefeling, Markus Dürmuth, and Luigi Lo Iacono. What's in Score for Website Users: A Data-Driven Long-Term Study on Risk-Based Authentication Characteristics. In *Financial Cryptography and Data Security*, FC '21, pages 361–381, Virtual Conference, March 2021. Springer.
- [64] Stephan Wiefeling, Luigi Lo Iacono, and Markus Dürmuth. Is This Really You? An Empirical Study on Risk-Based Authentication Applied in the Wild. In *International Conference on ICT Systems Security and Privacy Protection*, IFIP SEC '19, pages 134–148, Lisbon, Portugal, June 2019. IFIP.
- [65] Stephan Wiefeling, Tanvi Patil, Markus Dürmuth, and Luigi Lo Iacono. Evaluation of Risk-based Re-Authentication Methods. In *International Conference on ICT Systems Security and Privacy Protection*, IFIP SEC '20, pages 280–294, Virtual Conference, September 2020. IFIP.
- [66] Stephan Wiefeling, Jan Tolsdorf, and Luigi Lo Iacono. Privacy Considerations for Risk-Based Authentication Systems. In *International Workshop on Privacy Engineering*, IWPE '21, pages 320–327, Virtual Conference, September 2021. IEEE.
- [67] Flynn Wolf, Ravi Kuber, and Adam J. Aviv. “Pretty Close to a Must-Have”: Balancing Usability Desire and Security Concern in Biometric Adoption. In *ACM Conference on Human Factors in Computing Systems*, CHI '19, pages 151:1–151:12, Glasgow, Scotland, United Kingdom, April 2019. ACM.
- [68] Tianyin Xu, Han Min Naing, Le Lu, and Yuanyuan Zhou. How Do System Administrators Resolve Access-Denied Issues in the Real World? In *ACM Conference on Human Factors in Computing Systems*, CHI '17, pages 348–361, Denver, Colorado, USA, May 2017. ACM.
- [69] Tianyin Xu and Yuanyuan Zhou. Systems Approaches to Tackling Configuration Errors: A Survey. *ACM Computing Surveys*, 47(4):70:1–70:41, July 2015.

Appendix

A Study Part 1: Hands-on Task

Scenario

For participants in the neutral, security, and usability treatment

In this scenario, you are a system administrator of the MediaShop Corporation, a company with 300 employees. There you administrate the online shop www.dresscode.com, which sells both cheap and expensive clothing. You have just received an email from your supervisor Jo:

For participants in the neutral treatment

Hey Alex,
did you know that our login management system supports risk-based authentication? I just activated it, but not sure which settings are the best for us. Could you please complete the setup? I'm sure you will do fine.

Regards,
Jo

For participants in the security treatment

Hey Alex,
not sure if you heard it, but a hacker was able to log in to one of our customers accounts. As far as we know, the customer reused their password and the hacker got it from a hacked database. Afterwards, the hacker ordered lots of expensive jewelry using the account. My boss wants me to make sure that this should never happen again! I just activated the risk-based authentication in our login management system, could you please complete the setup for me?

Regards,
Jo

For participants in the usability treatment

Hey Alex,
did you know that our login management system supports risk-based authentication? We should give it a try. Could you please complete the setup? But make sure our customer support doesn't receive a ton of emails because of frustrated customers.

Regards,
Jo

For participants in the neutral (in-house) treatment

In this scenario, you are a system administrator of the MediaShop Corporation, a company with 300 employees. There you administrate the login system 'VPN-Guard' that the employees use to work from home. You have just received an email from your supervisor Jo:

Hey Alex,
did you know that VPN-Guard supports risk-based authentication? I just activated it, but not sure which settings are the best for us. Could you please complete the setup? I'm sure you will do fine.

Regards,
Jo

Now you open the setup...

Configuration

Page as shown in Figure 2

Usability Questionnaire

For the assessment of the configuration system you just used, please select your agreement/disagreement with the following statements. Please select the answer choice that most closely matches how you feel about the following statements:

- SUS1** I think that I would like to use this system frequently.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS2** I found the system unnecessarily complex.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree

- SUS3** I thought the system was easy to use.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS4** I think that I would need the support of a technical person to be able to use this system.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS5** I found the various functions in this system were well integrated.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS6** I thought there was too much inconsistency in this system.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- AC** Please select 'Agree' as the answer to this question.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS7** I would imagine that most people would learn to use this system very quickly.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS8** I found the system very cumbersome to use.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS9** I felt very confident using the system.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree
- SUS10** I needed to learn a lot of things before I could get going with this system.
 Strongly disagree Disagree Neither agree or disagree
 Agree Strongly agree

How familiar are you with the following terms? Please choose a number between 1 and 5 where 1 represents "Not at all familiar" and 5 represents "Extremely familiar" with the item.

	Not at all familiar (1)	Slightly familiar (2)	Somewhat familiar (3)	Moderately familiar (4)	Extremely familiar (5)
Malware	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Phishing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Two-factor authentication	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
One-time password	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Personal identification number (PIN)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Auto-fill	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Challenge-response	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Brute-force attack	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Security question	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Demography

- D1** What is your official job title?
Answer: _____
- D2** For how many years have you been working as a system administrator?
 0–1 years 2–3 years 4–5 years 6–10 years
 11–15 years >15 years
- D3** How large is the organization that you work for?
 1–9 employees 10–49 employees 50–250 employees
 >250 employees
- D4** How old are you?
 Answer: _____ Prefer not to answer
- D5** Which of these best describes your current gender identity?
 Woman Men Non-binary
 Prefer to self-describe: _____
 Prefer not to answer
- D6** What is the highest degree or level of school you have completed?
 No schooling completed Some high school, no diploma
 High school graduate, diploma, or equivalent
 Trade, technical, or vocational training Bachelor's degree
 Master's degree Doctoral degree Prefer not to answer

B Study Part 2: Interview

Introduction

- Thanks again for taking part in this study.
- The interview will take about 30 minutes.
- Are you OK with me recording our interview?
- <Start recording.>
- There are obviously no right or wrong answers here, we are just interested in your personal perceptions and your honest opinions.
- Are there any questions from your side before we start?

Warm-up Questions

- Q1** What do you like about your job as an administrator?
Q2 What are the main tasks in your job?

Behavior for the Risk Levels

We're now interested in the settings for the risk-based authentication. If we use the term "settings" in the following, we refer to the table on top of the page.

- Q3** How did you go about choosing the settings?
If not already covered by Q3
- Q4** Explain the reasons for the chosen settings.
If not already covered by Q3
- Q5** Explain the reasons for the chosen settings for notifying the users.
- Q6** Which difficulties or problems did you have when configuring the settings?
If log showed that info page was visited
- Q7** You have used the Wiki which contained more information about the settings: why did you click on the link? Was the information helpful?
- Q8** Have you used any other help, e.g., Google? If yes, why?

Wording of the Notifications

We're now interested in the notifications and their settings, i.e., the text fields on the bottom of the page.

- Q9** How did you go about when choosing the wording of the notifications?
If not already covered by Q9

- Q10** Explain the reasons for the way you worded the notifications.
Q11 Which difficulties or problems did you have when choosing the wording of the notifications?

If log showed that info page was visited

- Q12** You have used the info page which contained more information about the configuration of notifications: why did you click on the link? Was the information helpful?
- Q13** Have you used any other help, e.g., Google? If yes, why?

Risk-based Authentication

- Q14** How did you incorporate the scenario when making the configurations?
- Q15** Have you ever received such a notification? If yes, have you thought about this experience when making the configurations?
- Q16** Have you ever worked with risk-based authentication before? If yes, how did you experience the system you used compared to this one.

Potential Improvements

We're now interested in the system as a whole, i.e., both the table on the top and the text fields on the bottom of the page.

- Q17** How do you rate the current level of detail in the settings options?
Q18 Please explain anything that hindered you from the risk-based authentication in the way you wanted.
- Q19** What did you notice or remember most negatively about the system?
Q20 What did you notice or remember most positively about the system?
Q21 If you could change the system in any way you want: how would the perfect system look like?

Debriefing

- Research goal: Analyze the usability of an exemplary systems for the configuration of risk-based authentication, identify good and bad aspects to be able to make recommendations on how to improve such a system.
- Do you have any questions about the interview or the study?
- <Stop recording.>

C Additional Tables

Table 4: General security knowledge of the participants determined by rating the familiarity with 9 security-related items. The items are in the order of appearance in the questionnaire.

Item	Mean	SD
Malware	4.6	0.6
Phishing	4.8	0.5
Multi-Factor Authentication	4.8	0.4
One-Time Password	4.7	0.5
Personal Identification Number (PIN)	4.8	0.4
Auto-Fill	4.5	0.6
Challenge-Response	3.9	1.1
Brute-Force Attack	4.5	0.9
Security Question	4.6	0.6
Composite score	4.6	0.7
Cronbach's α	0.80	

Table 5: Configuration for the behavior of the risk levels (✔: allow, ? : optional MFA, ! : require MFA, - : block), notifying users (🔔: notify, 🔕: do not notify), and changes to the notification (✏️: changed, -: unchanged).

Participant	Risk Level Behavior			Notify Users			Changed Notification	
	Low	Medium	High	Low	Medium	High		
Default	✔	?	?	🔔	🔔	🔔	-	
Neutral	N-P1	✔	!	!	🔕	🔕	🔔	✏️
	N-P2	✔	!	!	🔔	🔔	🔔	✏️
	N-P3	✔	!	-	🔔	🔔	🔔	-
	N-P4	?	!	!	🔔	🔔	🔔	-
	N-P5	!	!	!	🔔	🔔	🔔	✏️
	N-P6	✔	?	?	🔔	🔔	🔔	-
	N-P7	!	!	!	🔔	🔔	🔔	-
Security	S-P1	✔	?	!	🔔	🔔	🔔	-
	S-P2	✔	?	!	🔔	🔔	🔔	✏️
	S-P3	?	!	-	🔔	🔔	🔔	✏️
	S-P4	!	!	-	🔔	🔔	🔔	✏️
	S-P5	?	!	!	🔔	🔔	🔔	✏️
	S-P6	?	!	!	🔔	🔔	🔔	-
	S-P7	?	!	-	🔕	🔔	🔔	✏️
Usability	U-P1	✔	!	-	🔔	🔔	🔔	-
	U-P2	?	!	!	🔔	🔔	🔔	-
	U-P3	?	!	!	🔕	🔕	🔔	-
	U-P4	?	!	-	🔔	🔔	🔕	✏️
	U-P5	?	!	!	🔔	🔔	🔔	✏️
	U-P6	✔	?	!	🔔	🔔	🔔	✏️
	U-P7	?	!	!	🔕	🔔	🔔	✏️
Neutral (in-house)	NI-P1	?	!	!	🔔	🔔	🔔	✏️
	NI-P2	✔	?	!	🔕	🔔	🔔	✏️
	NI-P3	?	!	-	🔕	🔔	🔔	✏️
	NI-P4	!	!	!	🔔	🔔	🔔	✏️
	NI-P5	!	!	-	🔔	🔔	🔔	-
	NI-P6	?	!	!	🔔	🔔	🔔	-
	NI-P7	!	!	!	🔕	🔔	🔔	-

D RBA Configuration Interface

Configuration

To read the Scenario again, click [here](#).

How do you want to use risk-based authentication for login attempts that are classified as low, medium and high risk?

You can use risk-based authentication to increase protection against login attempts that are considered to be at higher risk, such as login attempts from an unknown location or device. [Learn more about risk-based authentication.](#)

	Allow	Optional MFA	Require MFA	Block	Notify users
Low risk	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="checkbox"/>
Medium risk	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="checkbox"/>
High risk	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="checkbox"/>

Risk-based authentication notification messages

In the email templates, you can use placeholders to include specific details about the event such as {country}, {city}, {login-time}, {device-name}, {ip-address}, {one-click-link-valid} and {one-click-link-invalid}. To build your own one click link, you can use placeholders such as {event-id}, {feedback-token} and {username}. [Learn more about message templates.](#)

Notification for allowed logins (risk identified but MFA not required and not blocked)

Subject
New login attempt

We observed an unrecognized sign-in to your account with this information:
Time: {login-time}
Device: {device-name}
Location: {city}, {country}
If this sign-in was not by you, you should change your password and notify us by clicking on {one-click-link-invalid}.
If this sign-in was by you, you can follow {one-click-link-valid} to let us know.

Notification for logins requiring MFA

Subject
New login attempt

We required you to use multi-factor authentication for the following sign-in attempt:
Time: {login-time}
Device: {device-name}
Location: {city}, {country}
If this sign-in was not by you, you should change your password and notify us by clicking on {one-click-link-invalid}.
If this sign-in was by you, you can follow {one-click-link-valid} to let us know.

Notification for blocked logins

Subject
Blocked login attempt

We blocked an unrecognized sign-in to your account with this information:
Time: {login-time}
Device: {device-name}
Location: {city}, {country}
If this sign-in was not by you, you should change your password and notify us by clicking on {one-click-link-invalid}.
If this sign-in was by you, you can follow {one-click-link-valid} to let us know.

Figure 2: The interface of the central page in our study where participants configured the risk-based authentication. The layout of this interface is modeled after the risk-based authentication system of AWS Cognito (see Section 3). All aspects of the risk level and notification configuration match the Cognito interface, including texts, links, tooltips, help pages, and the overall design.

Let's Hash: Helping Developers with Password Security

Lisa Geierhaas
University of Bonn

Anna-Marie Ortloff
University of Bonn

Matthew Smith
University of Bonn, FKIE Fraunhofer

Alena Naiakshina
Ruhr University Bochum

Abstract

Software developers are rarely security experts and often struggle with security-related programming tasks. The resources developers use to work on them, such as Stack Overflow or Documentation, have a significant impact on the security of the code they produce. However, work by Acar et al. [4] has shown that these resources are often either easy to use but insecure or secure but hard to use. In a study by Naiakshina et al. [44], it was shown that developers who did not use resources to copy and paste code did not produce any secure solutions at all. This highlights how essential programming resources are for security. Inspired by the Let's Encrypt and Certbot that support admins in configuring TLS, we created a programming aid called Let's Hash to help developers create secure password authentication code easily. We created two versions. The first is a collection of code snippets developers can use, and the second adds a wizard interface on top that guides developers through the decisions which need to be made and creates the complete code for them. To evaluate the security and usability of Let's Hash, we conducted a study with 179 freelance developers, asking them to solve three password programming tasks. Both versions of Let's Hash significantly outperformed the baseline condition in which developers used their regular resources. On average, Let's Hash users were between 5 and 32 times as likely to create secure code than those in the control condition.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

1 Introduction

It is well known that end-users struggle with password security. Recent work in the field of Usable Security for Developers and many real-world compromises have shown that many developers also struggle when tasked with implementing password-based authentication systems [7, 18, 29, 40–44, 49]. Unlike end-users' problems that can be dangerous enough, although only one account is usually affected, millions of accounts can be affected if developers make only one mistake.

There have been multiple studies to advance the understanding of how the usability of APIs affects security during software development [4, 22, 33, 40, 42, 43, 45, 68]. However, one crucial aspect is the quality of the available documentation that developers use to solve their tasks. These are often either easy to use but insecure or secure but hard to use [5, 6, 27, 68] with many examples showing that developers copy and paste insecure code from online resources [4, 5, 24, 27]. Acar et al. write [4]: “our results confirm that API documentation is secure but hard to use, while informal documentation such as Stack Overflow is more accessible but often leads to insecurity.”

So copy and pasting of insecure code is a serious concern to software security, with Fischer et al. [27] postulating that Stack Overflow is harmful. However, studies by Naiakshina et al. [43, 44] show that only the participants who used copy and paste achieved any security. Those who did not use copy and paste did not achieve any security. So while copy and paste has been reliably identified as a serious security threat, it is also an essential method for secure solutions. Thus the goal needs to be to create programming resources that are easy to use but also help developers create secure solutions.

In this paper, we create Let's Hash, a programming resource to aid developers in creating secure code for password-based authentication. Our goal is to offer something as easy to use as Stack Overflow but as secure as official documentation or programming books. We created two versions of Let's Hash. The first is a simple website offering code snippets in a similar style to Stack Overflow. With this version, developers

are still required to select and assemble the code snippets themselves. The second adds a wizard on top, which lets developers specify the security requirements, and the wizard assembles all the necessary code, which is then ready to use.

Currently, Let's Hash can help developers create code for the following three tasks: hashing and salting passwords for storage, creating and enforcing password policies, and complementing password-based authentication with a second factor; two-factor authentication (2FA). To evaluate the usability and security of Let's Hash, we conducted a usability study with 179 freelance developers, who were asked to work on three short programming tasks in the context of password storage, password policies, and 2FA. Participants were split into three groups, one for each version of Let's Hash and one control group in which developers were allowed to use the resources they usually use during development.

The results show a vast improvement. Participants using Let's Hash were between 5 and 32 times as likely to achieve secure code than the control, depending on the task and version of Let's Hash. Post-hoc tests show that all improvements between Let's Hash and the control group are statistically significant. With these results, we believe that Let's Hash can offer a valuable contribution and help improve password security significantly.

2 Related Work

Authentication is a major part of security in IT, and it is susceptible to vulnerabilities in many ways. Attackers can gain unauthorized access to systems by manipulating or circumventing the authentication process, e.g., by guessing commonly chosen passwords [35, 53, 60], or through password leaks from databases [18, 29, 49]. End user focused research explored the difficulties that users have with security mechanisms in general [8, 57, 67], and specifically the authentication process like choosing and remembering passwords [30, 36, 63] or using alternative methods or second factors [17]. However, there is only limited knowledge of how to support software developers with secure programming [6, 33]. Recent work found that developers lack security expertise and often base their security decisions on misconceptions or outdated knowledge [6, 33, 42, 43]. But there already exist examples of APIs developed to support programmers with security, such as the Secure Socket API [45].

Let's Hash currently supports developers with password storage, password policies, and two-factor authentication, so we cover related work for each of these areas in the following.

2.1 Password Storage

For secure storage in a database, user passwords have to be salted and hashed [32]. Software developers, however, struggle with this task [7, 12, 29, 41–44, 68]. Previous studies showed that developers often search for programming

code on the Internet to copy and paste it to their applications [4, 27, 41, 42, 44]. While Fischer et al. [27] and Acar et al. [4] found that this behavior can lead to functional but insecure software, in a password-storage study with developers of Naiakshina et al. [44], all participants who submitted secure programming code had copied and pasted it from the Internet. The authors analyzed the used websites in detail and found that participants adopted code from blog posts, tutorials, and Stack Overflow. In [41], Naiakshina et al. conducted a further study on password storage with developers. If participants submitted insecure solutions, the authors provided links to websites of the Open Web Application Security Project (OWASP) and the National Institute of Standards and Technology (NIST), where advice and programming code for secure user password storage was available. The results showed that guiding developers to appropriate information sources suitable to their programming use-cases can improve software security. However, 47% of participants did not find the appropriate security information without help from the authors.

2.2 Password Policies

To ensure that users choose passwords that are hard to guess for potential attackers, there are certain guidelines that are often implemented as requirements, commonly referred to as password policies [32, 46]. There have been multiple studies to examine the effect of enforcing such password policies [23, 56, 59, 61, 61, 62]. Requirements that target the passwords' composition, while common, do little to encourage users to pick better quality passwords. A combination of minimum length and minimum strength is more effective [59]. Password strength meters have also been investigated [23, 61, 62]. While stricter password policies can help users create better passwords, they also increase user frustration and reduce password retention [61]. Strength meters offering constructive feedback performed better [61]. Segreti et al. [56] investigated adaptive password policies, which aimed to increase password diversity by comparing new passwords to the existing password database, resulting in policies that changed as new passwords were added to the database. They found that this improved security at little cost to usability and that the additional feedback they provided on how to improve password security did not make much of a difference concerning usability [56]. To our knowledge, the implementation of password policies has not yet been investigated from a developer's point of view.

2.3 Two-Factor Authentication

Authentication can be made significantly more secure by adding different factors [52]. Yubico Security keys (Yubikeys) are an example of a hardware authentication device, which supports two-factor authentication (2FA) standards,

like the Universal 2nd Factor (U2F) and Fast Identity Online 2 (FIDO2) protocols [20]. Alam et al. [9] investigated possible pitfalls in implementing the new open source authentication standard, FIDO2, by evaluating discussions about it on Stack Overflow and assessing existing libraries and documentation. They found that documentation is currently not very usable, libraries implementing the standard are often both insecure and incomplete, and that developers have wrong mental models of implementing the standard and threats that FIDO2 protects against. The authors call for better support for developers to mitigate these issues.

3 Let's Hash

Let's Hash was loosely inspired by Let's Encrypt and Certbot [2]. The mission of Let's encrypt is to enable all admins to easily acquire and set up TLS certificates to combat the many sites that did not offer TLS at all or suffered from one of many misconfigurations. Our goal with Let's Hash is very similar. We want to enable all developers to easily integrate secure password storage, password policy enforcement, and two-factor authentication into their applications without falling prey to the many mistakes that can be made. The code snippets contained in Let's Hash are presented in a way that makes it easy for developers to copy and paste them into their projects since it was shown in previous studies that this is a common use-case (see Section 2). We designed Let's Hash according to websites like Stack Overflow [21] and blog posts [13], by presenting the code snippets divided by topic, but not split apart into single functions as is often seen in documentation [28]. Unlike NIST and OWASP, where theory and guidelines are detailed, Let's Hash offers a code-centric view, combining these guidelines with easily adaptable code.

Let's Hash currently supports password storage, policy enforcement, and two-factor authentication. In the following, we will highlight the most relevant aspects in these three areas.

3.1 Password Storage

There are a lot of different password hashing schemes (PHSs), which can be used in the context of user password storage (e.g., MD5, SHA-1, SHA-2, PBKDF2, `bcrypt`, `scrypt`, `Argon2`) [50]. We evaluated them for security and usability and included `Argon2id` as the most secure choice according to recent academic results [10, 14, 48, 50] and `bcrypt` as a more usable solution in contrast to `Argon2id` since it does not require manual adaption to specific hardware and is recommended by OWASP [48]. Iterations are configured based on the hashing algorithm. Currently, Let's Hash offers programming code snippets on secure user password storage in two programming languages, Python3 and Java.

3.2 Password Policies

NIST and OWASP both recommend policies that do not restrict password composition (i.e., allowing all kinds of characters but not enforcing a specific combination) and enforce a length of at least eight characters [32, 47]. Additionally, they advocate for using a strength checker, as is implemented by the library `zxcvbn` [66]. The German Federal Office for Information Security (BSI) advises users to choose their passwords according to popular composition rules - using upper- and lowercase letters and special characters [16]. Let's Hash offers a JavaScript solution to enforce BSI recommendations, ensuring a certain length and composition of a password. Additionally, there is a code fragment for a password strength checker using the aforementioned `zxcvbn`, as recommended by NIST and OWASP.

3.3 Two-Factor Authentication

Let's Hash offers a code fragment that generates a time-based token that serves as a one-time password. This token can be used as verification in conjunction with an app such as the Google Authenticator, which is a popular way to use 2FA [31]. Currently, programming code is provided in Python3.

3.4 Let's Hash Wizard

We have a two-component system. The code repository contains the code snippets for all the above tasks and sub-tasks, and a wizard assists developers in selecting and configuring the right snippets. We wanted to explore a design of Let's Hash that included a wizard-like user interface (UI)-element which required the user to interact with it. The wizard should present developers with the code that best fits their specific use-case by first taking them through a series of questions. These questions let developers pick secure implementation options according to the different recommendations provided by NIST, OWASP, and BSI. After stepping through the wizard, the appropriate code snippets are selected, configured, and presented to the developer ready to use.

Screenshots of the base version (LH) and the wizard (LH-W) and details on the code snippets are available on Github.¹ Let's Hash will also be released as an open-source project.

4 Methodology

To evaluate if using Let's Hash would increase security, we designed and ran an online study with freelance developers recruited from Freelancer.com as recommended by Naiakshina et al. [41]. All participants were asked to complete three short programming tasks on password storage, password policies, and 2FA. After task completion, we asked participants to fill out a survey assessing their experiences with the tasks and

¹<https://github.com/BeSecResearch/LetsHash-Supplemental>

the information sources they used. The order of the tasks was randomized. We divided the participants into three groups:

- Group **LH**: Participants were asked to complete the programming tasks using version LH, the basic version of Let's Hash.
- Group **LH-W**: Participants were asked to complete the programming tasks using version LH-W of Let's Hash, the version with an added wizard for configuration.
- Group **C(ontrol)**: Participants were asked to complete the programming tasks by using any information source they normally use when programming. These participants did not have access to Let's Hash.

4.1 Study Setup

For the study setup, we used the open-source tool Developer Observatory [58], which has been used in previous security studies about code development by Acar et al. [3, 7]. The tool allowed us to let participants work on programming tasks remotely. The participants could run and test the output of their code in a sandbox-like environment accessed via their browser. We provided them with function signatures and examples of expected results. This offered several advantages: First, participants did not have to download anything. They could directly access the consent form, the task description, the programming code, and the link to the follow-up survey in a browser of their choice. Second, participants did not need to spend time setting up their IDE. They could write and test the programming code within the tool. Third, we were able to log study data without the participants having to submit any of this data to us manually. In the following, we describe the programming tasks in detail.

4.2 Task Design

The exact task descriptions can be found in Appendix A. Examples of the study interface are available on Github.²

Task 1 (T1) - Password hashing: To keep the task as simple as possible, we only asked participants to implement a function for hashing and verifying passwords in Python. The solution required outputting a hash value and the correct verification of a given password. The task description as it was presented to participants can be found in Appendix A.

Task 2 (T2) - Password policy: We asked participants to implement a short JavaScript program checking a string for adherence to a given password policy. Since it is still a widely used practice in real life, we asked them to implement a policy that would enforce composition rules in addition to a minimum length. The solution of the task required an output that correctly breaks down a given password's adherence to the

demanded policy. Appendix A contains the task description as it was presented to participants.

Task 3 (T3) - Two-factor authentication: We asked participants to implement a method that generates a time-based code that can be used as an authenticator. The solution to the task required an output containing the one-time code and its verification. The task description that was presented to study participants can be found in Appendix A.

4.3 Survey

In the survey, we asked participants of all groups about their experience with programming in general and the given topics in particular. They also indicated their perception of the difficulty of the programming tasks and were asked general demographic questions. Participants of groups LH and LH-W additionally had to answer the System Usability Scale (SUS) [37] for the version of Let's Hash that they used. Participants of the control group C were asked about the specific resources they used to solve the tasks and whether they were satisfied with them. The surveys for all the three groups can be found in Appendix B.

4.4 Usability Evaluation

In accord with ISO 9241, we define usability as encompassing effectiveness, efficiency, and user satisfaction [1].

Effectiveness: Every task submission was examined based on two criteria: functionality and security. To count as functional, the code submitted by the participants needed to run and produce an output that offered the information specified in the task description. Functionality was a prerequisite for security.

To determine whether the submitted code for Task 1 was secure, we adopted the security scale introduced by Naiakshina et al. in [43, 44] (see Appendix C). They used a scale of up to seven points for hashing and salting user passwords. We used the same scale, and only rated solutions as secure that reached at least 6 out of 7 points. We were strict in our evaluation because we were only interested in solutions that offered up-to-date security. This meant using a random salt and a key derivation function including an appropriate iteration count or a memory-hard function [32, 43]. We did not require 7 out of 7 points since the final point is for memory hardness which is not yet industry standard and we did not expect our participants to go beyond industry best practice. For Task 2, the code was rated as secure if the policy rules specified in the task description were correctly implemented without errors. The code for Task 3 was only rated as secure if the algorithm used to generate the second factor actually generated a time-based one-time code and used a salt that was randomly generated. The programming code was evaluated manually and independently by two computer science researchers. Differences were resolved through discussion.

²<https://github.com/BeSecResearch/LetsHash-Supplemental>

Efficiency: The number of clicks and time taken to solve a task are often used as efficiency variables in usability studies [26, 34, 38, 39, 55]. To evaluate how participants interacted with Let’s Hash, we tracked the time (in seconds) they actively spent on the website and the clicks they needed to find the correct code fragments to use. We assumed that a long time and more clicks might affect the usability of the website [38, 39].

User satisfaction and perceived usability: We calculated the SUS score [37] for the two versions of Let’s Hash. We used the SUS as one factor to compare the usability of the two versions of Let’s Hash.

Additionally, we evaluated the answers to several open survey questions to gain insight into the participants’ workflow as well as their general attitude towards IT security and Let’s Hash. Since all answers were relatively short, and we were interested in specific themes, such as positive or negative attitudes towards Let’s Hash, we used deductive thematic analysis to categorize and report on the participants’ answers [15]. One researcher coded the entirety of the answers given, and a second researcher recoded them using the same codebook. Afterward, intercoder agreement was calculated per document. The minimum agreement was $\kappa = 0.76$, and the maximum agreement was $\kappa = 1$ ($M=0.94$). For the groups LH and LH-W, the questions were about the user experience with Let’s Hash and how the website compares to other resources the developers would usually use to program. Participants of the control group C were asked to list the resources they used for the programming tasks. We categorized the answers into three types of resources: Stack Overflow, official documentation, and other, which included various blog posts and other resources.

Usability and security: To investigate the relationship between security and usability, we conducted Wilcoxon-Rank-Sum Tests for each of the different tasks, comparing submissions with errors (non-functional or insecure) with secure non-erroneous submissions, with respect to usability measures, such as SUS, time spent on Let’s Hash and clicks. We corrected p-values using the Bonferroni-Holm procedure.

4.5 Error Analysis

We analyzed the types of errors we found in the participants’ submissions to find out more about the kinds of errors that Let’s Hash helps prevent and which ones still occur. To do this, we conducted a qualitative analysis. Each submitted solution for a task that was not both functional and secure was manually reviewed with regard to the types of errors it included. During the coding process, we assigned multiple different error types to a single submission, and ended up with a maximum of three different error types per submission. We estimated a lower limit of $\kappa = 0.84$ by ordering the error types per submission alphabetically and only taking into account the first one and an upper limit by counting the raters as in agreement, when they agreed over at least one of the assigned

error types, which resulted in $\kappa = 0.91$.

We were also interested in errors occurring even though participants used Let’s Hash. There are 53 non-functional or non-secure task solutions from groups LH and LH-W. We manually investigated whether these included copied code from Let’s Hash. Both researchers judged 8 of them to be copied from Let’s Hash, agreeing on 5 of them, and disagreeing on 3. The secure submissions in groups LH and LH-W were also tested on whether they had copied their code from Let’s Hash or not. Due to the high amount of files for these cases this process was semi-automated. Details are in Appendix D.

All differences concerning the error types and the code-copying were resolved through further discussion, and full agreement was reached.

4.6 Hypotheses and Statistical Analysis

We were interested in the security of code developed with the help of Let’s Hash. Additionally, we wanted to study Let’s Hash’s usability, encompassing efficiency and effectiveness [1]. Therefore, we examined four main hypotheses in our study: one on the security score between the groups LH and LH-W and the control group C, denoted by S(ecurity), and three concerning the differences between groups LH and LH-W, denoted by D(ifference). While we hoped that LH-W would improve both security and usability over LH, there is not enough theoretical foundation to justify one-tailed hypotheses, so all hypotheses were tested two-tailed at the standard $p=.05$ level throughout.

- H-S: The groups LH and LH-W, that are working with Let’s Hash, produce code that is more Secure than that produced by the control group C, that had no access to Let’s Hash but could use any other source.
- H-D1: The System Usability Scale (SUS) Differs between the two versions of Let’s Hash.
- H-D2: There is a Difference in the number of clicks needed to reach the desired code fragments using the two different versions of Let’s Hash.
- H-D3: There is a Difference in time that participants need to reach the desired code fragments using the two different versions of Let’s Hash.

We used the freely available software Gnu R [51] for statistical analyses.

4.7 Pilot Study

We ran a pilot test before the main study to test the technical setup and ensure that we had correctly configured the Developer Observatory tool and the website Let’s Hash. We recruited three participants. These were students who worked

Table 1: Demographics of 179 participants

Gender	Male: 92%	Female: 8%	Prefer not to say: 0.6%
Ages	Mean: 28.6	Median: 27	SD: 7.5
Education and Occupation	University Degree: 80%	Employed at company: 28%	
Country of Origin	India: 22%	Pakistan: 9%	Other: 69%
Experience (Programming language)	Python3: Mean: 3.2 Median: 3 SD: 2.3	JavaScript: Mean: 4.3 Median: 4 SD: 3.3	Overall: Mean: 6.4 Median: 5 SD: 5.0

as research assistants in security-related fields within computer science. All participants were male and aged between 20 and 40 years. The pilot study indicated that the setup of the Developer Observatory and Let’s Hash worked as intended. The participants did not raise any serious issues, and we only made minor changes based on their feedback.

4.8 Power Analysis

We performed a power analysis based on our four main hypotheses to calculate the required sample size for this study. We used G*Power to perform two analyses [25], one for H-S, and one for H-D1, H-D2, and H-D3. We calculated a required sample size of at least 49 participants per group.

For H-S, we performed a Fisher’s exact test, comparing the groups LH and LH-W against group C. For this power calculation, we merged groups LH and LH-W, since both versions of Let’s Hash only differed in the existence of a wizard, but the code fragments were the same on both versions. We assumed that the code developed by LH and LH-W would be secure in 90% of the solved tasks, but the code for group C only in 60%. We based these percentages on the results of Acar et al., where a similar task on password storage was part of a user study with GitHub users [7]. Using these parameters, a desired error probability α of 0.05 and a desired power of 0.95 resulted in a sample size of 116 participants, 77 total for groups LH and LH-W, and 39 for group C.

With H-D1, H-D2, and H-D3, we aimed to figure out if the added wizard had a noticeable effect on the usability of Let’s Hash. To compute a required sample size, we used the two-tailed Wilcoxon-Mann-Whitney test with two groups of equal sizes. Since Klug stated that “the average SUS score is 68 with a standard deviation of 12.5” [37], we set our baseline value accordingly for H-D1. We used mean values of 68 and 78 for the two groups, which would for instance improve the usability of Let’s Hash from “Ok” to “Good” [11]. The standard deviation was set to 12.5 for both. This led to an effect size of 0.8. For H-D2 and H-D3, we used the same effect size for comparability since we were not aware of standardized measures for the number of clicks or the time spent on a website such as Let’s Hash. Using this effect size and the same values for α and power as in our first analysis we calculated a required sample size between 28 and 49 participants per group, depending on the distribution, so we selected 49 to be

on the safe side.

4.9 Participants

For our study, we needed developers with experience in Python and JavaScript. We used the support service of Freelancer.com for participant recruitment as suggested in [19,42]. All freelancers who finished the tasks and survey received 40€ for participation.

294 participants were invited to take part in our study to have enough participants to account for drop-outs and other issues. All 294 participants provided informed consent. Of these, 239 completed the tasks and the survey, 55 quit the study midway. The data of 60 participants was not considered for analysis for varying reasons. 31 of them were removed because we found inconsistencies in the recorded tracking, which showed that multiple participants had access to the wrong version of the website Let’s Hash or even to both versions. Some of these inconsistencies were caused by a bug in the Developer Observatory tool, which was reported by us upon discovery. 15 participants could not be tracked by Let’s Hash at all and had to be disregarded for that reason, 13 were excluded due to technical problems, 9 leading to data loss and 4 with incorrect condition assignment, and 1 participant was removed for speeding through the survey and giving nonsensical answers. Overall, 179 participants produced valid results, 58 in group LH, 57 in group LH-W, and 64 in group C exceeding our 49 target.

The majority of the participants were male (92%), while only 8% were female, and one person preferred not to disclose their gender, which is fairly typical for these platforms. They reported ages between 18 and 70 years ($M=28.6$, $SD=7.5$) and between 0 and 30 years of programming experience overall ($M=6.4$, $SD=5.0$), with slightly less experience in Python3 and JavaScript. Most participants were not from countries where English is the only primary language, with Indians (22%) and Pakistanis (9%) representing the largest groups of nationalities in our sample. The remaining 69% of participants were from a variety of different countries, none of which amounted to more than 4%. Further demographics information about the participants is in Table 1.

Table 2: Distribution of correct solutions by task and group

Task		LH	LH-W	C
1	Functionality	97%	95%	81%
	Security	93%	82%	33%
2	Functionality	95%	91%	70%
	Security	91%	75%	36%
3	Functionality	97%	88%	59%
	Security	86%	79%	16%

5 Limitations

As usual for usability studies, several limitations have to be considered in the context of this study. Firstly, we did not conduct the study in a lab setting. Due to requiring a high number of software developers and the ongoing COVID-19 pandemic, we conducted an online study. Consequently, we had less control and were not able to track participants' processes when working on the tasks.

Secondly, we used Freelancer.com for study recruitment, which limited the pool of possible participants to people registered on this platform. The vast majority of the recruited freelancers were not native English speakers, which might have lead to misunderstandings due to a language barrier. However, real-world projects are often outsourced to freelancers under similar conditions.

Thirdly, due to our study setting, our tasks were rather short, did not include finding the resource and we provided implementation stubs to prevent using developers' time unduly. Thus, participants might not have put as much effort into the task, or have applied different priorities than in a real task. However, results from a lab study have shown to be comparable to those in a field study in the context of password storage [41, 42]. Our work provides the basis for further investigation.

Finally, the tracking of time participants spent on Let's Hash and the number of clicks they needed to achieve their goal was not as accurate as we would have wished due to the remote nature of the study. We could not capture the actual screen and had to attempt to log the time on the server-side. In some cases during testing, clicks were either not registered or counted twice. Additionally, we only recorded the interactive time spent on Let's Hash. If participants stopped scrolling, clicking, or moving their mouse for more than a minute, the timer stopped increasing. We wanted to avoid recording time during which users had the website open in a tab, but they were not actually looking at it, for example, because it was minimized. Considering this, these values should be taken as a best-effort approximation.



Figure 1: Secure solutions, divided by task and group.

6 Ethics

This study was conducted in Germany and is compliant with the EU General Data Protection Regulation, a directive concerning the collection and storage of data. It also has Institutional Review Board (IRB) approval. All data was collected anonymously. Participants were asked to agree to a consent form that informed them about the study procedure and their rights, for example, to withdraw at any point during the study. Furthermore, we complied with the privacy policies of Freelancer.com, which meant we were not permitted to ask participants for their email addresses for further research or to inform them about study results.

7 Results

For evaluation, we compared the results of participants using Let's Hash and information sources of their own choice, as well as the two versions of Let's Hash.

7.1 Participants' Submissions

The participants who finished the study self-reported taking a median time of one hour to solve the programming tasks. Table 2 shows an overview of the evaluation of participants' submissions concerning functionality and security. We only rated security if the programming code was functional and included only functional solutions in our statistical tests comparing security. All participants combined produced functional solutions for Task 1 in 91%, Task 2 in 85% and Task 3 in 80% of the cases, and secure solutions for Task 1 in 68%, Task 2 in 66% and Task 3 in 59% of the cases.

7.1.1 Functionality

Of all the participants who produced valid data, 168 submitted programming code that we considered functional for at

Table 3: Overview over the results of the 3-way FET for H-S

Hypothesis	IV	DV	<i>p</i> -value	<i>cor</i> - <i>p</i> -value
H-S: T1	LH vs. LH-W vs. C	Achieved security	<0.001*	<0.001*
H-S: T2	LH vs. LH-W vs. C	Achieved security	<0.001*	<0.001*
H-S: T3	LH vs. LH-W vs. C	Achieved security	<0.001*	<0.001*

T1: Task on secure password storage, T2: Task on password policy, T3: Task on 2FA;
 IV: Independent Variable; DV: Dependent Variable; FET: Fisher’s Exact test
cor - *p*-value: *p*-value, Bonferroni-Holm corrected; tests marked with *: statistically significant.

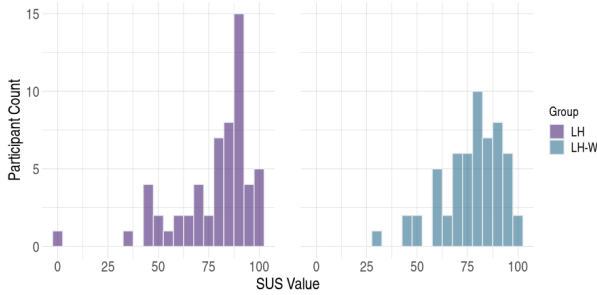


Figure 2: Distribution of SUS values.
 One bar covers a range of n+5 points.

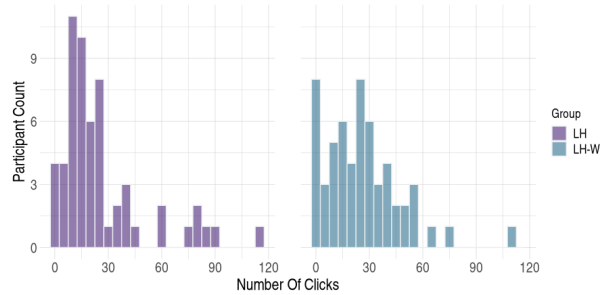


Figure 3: Distribution of counted clicks.
 One bar covers a range of n+5 clicks.

least one of the tasks. The task with the most functional solutions overall was Task 1. For this task, 162 participants submitted a functional solution. Of these, 157 participants indicated that they had previous experience with storing passwords in a database. For Task 2, 152 participants submitted a functional solution, and 143 participants indicated they had previous experience with implementing password policies in the context of a login form. The task that produced the least amount of functional solutions was Task 3. 144 participants submitted a functional solution for Task 3. Of these, 96 participants indicated that they had previous experience with implementing 2FA. Compared to the two other tasks, 2FA seems less frequently demanded among the population on Freelancer.com.

7.1.2 Security

The task with the most secure solutions overall was Task 1 with a total of 122 secure solutions, followed by Task 2 with 119 secure solutions. Task 3 had the least amount of secure solutions with only 105.

7.2 Hypotheses

7.2.1 Security

Figure 1 shows the proportions of secure solutions, divided by groups and tasks. We conducted separate 3-way Fisher’s Exact tests for each programming task to test H-S. We corrected *p*-values for multiple testing using the Bonferroni-

Holm-correction since we conducted three tests for this hypothesis. All *p*-values for the main analyses, including the corrected ones, are below 0.001, as can be seen in Table 3. Our analyses indicated that for all of the tasks, there was a significant difference with respect to security between the three groups. We then conducted post-hoc 2-way Fisher’s Exact tests to compare the groups individually (see Table 5) and included these in our correction. We found that for all of the tasks, the solutions achieved with the help of Let’s Hash were significantly more secure than those achieved with resources of the participants’ own choosing. This shows that the primary goal of our work was achieved. Let’s Hash significantly increases the odds of developers creating secure solutions by a large margin.

7.2.2 Usability

To compare the usability of the two Let’s Hash versions, we evaluated the SUS, the number of clicks, and the time spent actively on each version of the website. Since neither of the groups LH or LH-W had a normal distribution in SUS values, time spent, or amount of clicks used on the website Let’s Hash, we performed Wilcoxon-Rank-Sum tests. The results of our statistical analyses of H-D1, H-D2, and H-D3 are available in Table 4.

For H-D1, we did not find a significant difference in usability as measured by SUS between the version with (M=78.6, median=80, SD=15.1) and the version without a wizard (M=79.1, median=87.5, SD=19.3). In general, both versions of Let’s Hash achieved results that were close to being con-

Table 4: Overview over the results of the WRS tests for H-D1, H-D2 and H-D3

Hypothesis	IV	DV	\mathcal{W}	r	p -value
H-D1	LH vs. LH-W	Achieved SUS	1500	0.08	0.3926
H-D2	LH vs. LH-W	Amount of clicks	1804	0.08	0.3996
H-D3	LH vs. LH-W	Time spent on website	1329	0.17	0.0704

IV: Independent Variable; DV: Dependent Variable; WRS: Wilcoxon-Rank-Sum test;
 \mathcal{W} : Wilcoxon- \mathcal{W} ; r : Effect size (Pearson's r)

Table 5: Overview over the results of the post-hoc 2-way FETs for H-S

IV (Group)	Task	DV	OR	CI	p -value	cor - p -value
LH vs. LH-W	Password Storage	Achieved security	0.252	[0.024, 1.408]	0.09	0.19
LH vs. LH-W	Password Policy	Achieved security	0.183	[0.018, 0.95]	0.03*	0.08
LH vs. LH-W	2-Factor Authentication	Achieved security	1.079	[0.255, 4,798]	1	1
C vs. LH-W	Password Storage	Achieved security	9.668	[3.483, 30.399]	<0.001*	<0.001*
C vs. LH-W	Password Policy	Achieved security	4.115	[1.521, 11.958]	0.002*	0.009*
C vs. LH-W	2-Factor Authentication	Achieved security	23.886	[6.985, 100.032]	<0.001*	<0.001*
C vs. LH	Password Storage	Achieved security	38.372	[8.543, 359.027]	<0.001*	<0.001*
C vs. LH	Password Policy	Achieved security	22.456	[4.873, 212.115]	<0.001*	<0.001*
C vs. LH	2-Factor Authentication	Achieved security	22.208	[6.884, 84.436]	<0.001*	<0.001*

FET: Fisher's Exact test; IV: Independent Variable; DV: Dependent Variable; O.R.: Odds ratio; C.I.: Confidence interval
 cor - p -value: p -value, Bonferroni-Holm corrected, including nine post-hoc tests and three main 3-way analyses; tests marked with *: statistically significant

sidered excellent in usability [11, 54]. Figure 2 shows the distribution of SUS values for both versions of Let's Hash. The majority of ratings designate the usability of both versions as good (>71), although there was slightly more variance in ratings for group LH, which used the website without a wizard.

For hypothesis H-D2, we also did not find a significant difference in clicks needed between group LH (M=25.6, median=17.5, SD=25.1) and group LH-W (M=26.2, median=24, SD=21.5). Figure 3 shows the distribution of the counted clicks. Again, the distributions are roughly similar, but more participants issued very few clicks in group LH-W, and more participants issued high numbers of clicks (>60) in group LH.

To test H-D3, we compared the time actively spent on the website. Participants of group LH spent slightly more time on their version of the website (M=233.8, median=179, SD=200.5) than those in group LH-W (M=176.3, median=115, SD=156.4), but the difference was not significant. The time participants spent on Let's Hash is available in Figure 4. Participants spent a median of fewer than 3 minutes on version LH and fewer than 2 minutes on version LH-W.

7.2.3 Hypothesis Takeaways

The main takeaway of the hypotheses analysis can be summarized as follows: Using either version of Let's Hash as a resource during code development has a large and significant positive effect on the security of the developed code.

To our surprise, the usability of the two versions was rated almost identically, and thus, there were no statistically significant differences between the two, implying that the wizard did not improve usability as we had expected. Even more sur-

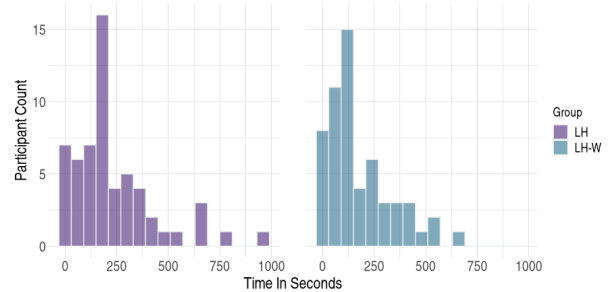


Figure 4: Distribution of time spent on Let's Hash.
 One bar covers a range of n+60 seconds.

prisingly, LH has higher security odds than the LH-W. While the difference was not statistically significant, we still find it interesting and discuss it further in the following section. For now, we conclude that the plain Let's Hash resource improves the odds for a secure solution between 17 and 32 times compared to the control and does not have any downsides compared to LH-W, so the extra effort for the wizard does not seem justified or necessary. Although further research into the reasons why is recommendable.

7.3 Error Analysis

Despite the excellent results, some participants created insecure code despite using Let's Hash. This section analyzes these errors and examines the usability judgments of participants making errors despite using Let's Hash. We found eight

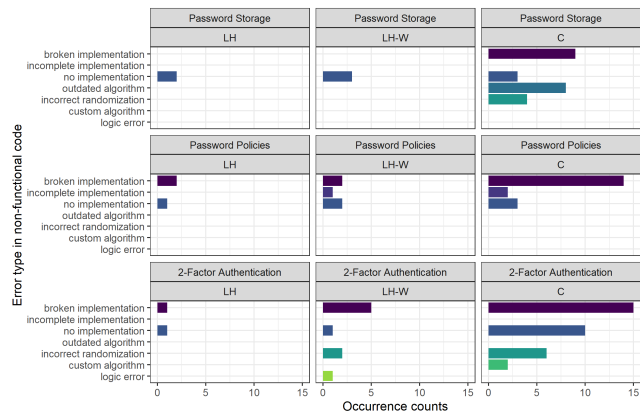


Figure 5: Types of errors in non-functional code, divided by task and group.

different types of errors in the participants' submissions, three related to functionality, and five types concerning security. Even for non-functional submissions, we also documented security errors.

Control group submissions proportionally more often had multiple error types, which led to them lacking in security and/or functionality. This is in addition to boasting more errors overall, as has already been established and shows that the insecurity and/or non-functionality of code was often due to multiple types of errors and not merely one cause.

The errors participants made that would lead to their code being non-functional were categorized as 'no implementation' when there was no attempt at a solution, 'incomplete implementation' when the participants did not finish the task and 'broken implementation' when the code contained severe programming errors. As can be seen in Figure 5, 'broken implementation' was the most common error, with 11 cases overall in LH and LH-W and 38 cases in group C. We included 'no implementation' as a category under the assumption that participants were likely not unwilling to solve the task, but instead overwhelmed. The security errors were 'plain text', 'outdated algorithms' and 'use of custom algorithms', which all refer to the cryptographic algorithms implemented by participants in tasks 1 and 3. For example, some participants used the outdated md5 to hash the password in Task 1, which we classified accordingly as 'outdated algorithms'. Another type of error concerning these two tasks is 'incorrect randomization', when the salt or shared secret was either too short, or not sufficiently randomized to be considered secure. Figure 6 shows that this was the most common security error type for the functional cryptographic tasks in all groups, with 18 overall occurrences in groups LH and LH-W and 55 in group C. One instance of this error was a participant who used their own first name as the shared secret in Task 3. None of these errors apply to Task 2 since there was no cryptogra-

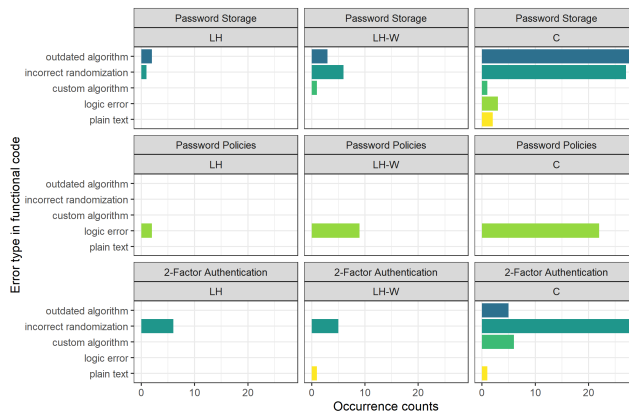


Figure 6: Types of security errors in functional code, divided by task and group.

phy involved in this task. Instead, the security error consisted of participants implementing a function that appeared to deliver the requested results but was not consistent. This leads to issues such as passwords being erroneously classified as adhering to policy. We refer to these errors as 'logic error'.

7.3.1 Non-functional Submissions

The frequency of different types of errors for those submissions which were classified as non-functional is depicted in Figure 5. It shows that non-functional solutions also suffer from security errors, in addition to purely functionality-related errors. This is especially common in the control group. In general, a wider range of different errors occurred in the control group compared to the two Let's Hash groups. The most striking example was in the Password Storage task, where LH and LH-W participants' submissions were non-functional because they did not implement anything, but group C submissions exhibited a wide range of functional and security errors.

7.3.2 Functional, but Non-secure Submissions

The frequency of different error types for submissions which were functional, but insecure is shown in Figure 6. Like for non-functional solutions, there were fewer different types of errors in the Let's Hash groups. For example, for the password storage task, there were five different types of security errors in the control group, but only two types of security in the LH and three types in the LH-W group. Some types of errors, the plain text and logic errors, were more common in functional solutions than non-functional solutions.

7.3.3 Errors Despite Using Let's Hash

In groups LH and LH-W, most errors occurred when participants did not use the code provided by Let's Hash. Of 53 tasks where participants submitted erroneous solutions, 45 did not use copied code from Let's Hash. In contrast, of 292 correct tasks in the groups LH and LH-W, only 19 were not copied. Overall, the most common task in which participants did not choose to use copied code was Task 2. This may be because the code had to be adapted more than in the other two tasks, since participants would have had to change the definition of a variable. Common mistakes that participants made when they did not copy code from Let's Hash were using outdated algorithms like md5 (3 cases) or hard-coding the salt (12 cases). One participant submitted a solution for Task 3 which did not contain a cryptographic algorithm at all. There were only eight submissions for tasks where participants copied code from Let's Hash but still created an insecure solution. Four of those had functionality errors which we classified as broken implementation. These errors mainly were caused by participants not adapting the code correctly from the website, or introducing faulty syntax, like indentation errors. The remaining four occurrences all fall into LH-W. Three of the errors were in the password storage task, and all included incorrect randomization. Two of those participants had removed the generation of the salt from the function `hash_password()` and instead used a static variable, and one of the participants lowered the amount of rounds for the salt's generation to 4 instead of the recommended minimum of 16. The remaining error occurred in the password policies task and was a logic error, specifically in the implementation of the function `composition()`, which would allow a password without any special characters to pass the check, violating the policy requested in the task description. This error was introduced because the participant either removed or did not copy a part of the function.

7.3.4 Usability and Errors

We found significant differences in usability between tasks with errors and those without for Task 1 concerning clicks, and for Task 3 concerning SUS, time (in seconds) and clicks. Participants with erroneous submissions issued significantly fewer clicks on Let's Hash both for Task 1 ($M=9.4$, $SD=8.5$) and Task 3 ($M=9.1$, $SD=10.4$) than participants with secure submissions (Task 1: $M=28.2$, $SD=23.8$; Task 3: $M=29.4$, $SD=23.7$), $W=1116.5$, corrected- $p=.004$ for Task 1, $W=1557.5$, corrected- $p<.001$ for Task 3. Participants who submitted wrong solutions also spent significantly less time on Let's Hash for Task 3 ($M=93.5$, $SD=76.9$) than those with secure solutions ($M=229$, $SD=189$), $W=1418$, corrected $p=.004$. Finally, participants with secure solutions rated Let's Hash with as significantly more usable (SUS score $M=81.3$, $SD=16.0$) than those with errors in their submissions ($M=67.3$, $SD=19.1$), $W=1393$, corrected $p=.006$. This

suggests that the participants who made mistakes despite having Let's Hash at hand may have abandoned this resource early on in their coding process before being able to solve their problem. The full results of this analysis are available on Github.³

7.4 Participants' Feedback on Resources

In general, feedback on Let's Hash as a resource for code development was positive. 57 participants gave detailed feedback, and of those, 40 participants reported that they found the website easy to use, and 28 participants said they found it pleasant, enjoyed the UI, and wished to use it again. Requested changes included the addition of more languages (both programming languages and spoken languages), tutorials on how to use the code, the ability to run the code directly on Let's Hash in a sandbox-like environment, and improvements to the UI. The most requested change was additional information on the presented code and security-related challenges, which was mentioned by 34 participants. This request suggests that in their usual workflow, most developers do look for at least some background information when incorporating code found online into their work. Furthermore, 56 of the 115 participants who worked with Let's Hash reported that they found Let's Hash to be generally easier to use than their usual resource, citing reasons such as easy navigation and a well-structured presentation of the code. One participant mentioned that he would trust Let's Hash more than his usual resource in terms of security since it is "not a forum post."

68 participants from groups LH and LH-W indicated that the main resource they would usually use is Stack Overflow. 41 said they would use the official documentation, but almost none of them cited other resources and those that did mostly indicated they would usually "search on google."

Participants of group C also most commonly mentioned Stack Overflow and official documentation. Both resources were mentioned by more than 20 of the 65 participants (>30%), and 18 of them (28%) indicated that Stack Overflow was their main resource. Other websites that were mentioned in group C were various blogs and some online schools like W3Schools [65] or Vitosh Academy [64].

8 Discussion

We found that Let's Hash significantly improved the security of our participants' code. Although all participants were asked to solve the three authentication tasks securely, most secure solutions were submitted by participants using Let's Hash – regardless of which Let's Hash version they were using. While for both groups LH and LH-W, the submitted solutions were secure in at least 82% of the cases for Task 1, 75% for Task 2 and 79% for Task 3, 72% of participants in the

³<https://github.com/BeSecResearch/LetsHash-Supplemental>

control group submitted insecure programming code, with 33% secure solutions for Task 1, 36% for Task 2, and 16% for Task 3, which is alarming. These results suggest a large positive effect of Let's Hash on software security.

We also compared the efficiency and perceived usability of the two Let's Hash versions. With the configuration wizard, we wanted to take a burden off developers and provide them with a better overview of current recommended security practices for the three security-sensitive authentication tasks. However, we did not find a significant difference between the two Let's Hash versions concerning participants' perceived usability, which we measured with a SUS score. The usability was fairly high for both versions. So it seems participants were satisfied with using either Let's Hash version. In the follow-up survey, 24% of participants also reported that they felt supported by Let's Hash and would use it again. Most importantly, participants indicated that they trusted Let's Hash more than their usual resources. The fact that trust can impact the chosen resources and thus indirectly affect software security was already reported in [42]. We believe that trust and a high measurement of perceived usability are key factors for the successful establishment of Let's Hash.

Furthermore, we did not find a significant difference in the number of clicks and the time participants needed to solve the tasks with either version of Let's Hash. With an average of fewer than 26 clicks and 3 minutes to solve the tasks with either Let's Hash version, participants completed the tasks in a short time with little effort. With a success rate in terms of functionality of at least 95% for Task 1, 91% for Task 2, and 88% for Task 3, almost none of the participants gave up.

That there was no significant difference in clicks or time between the Let's Hash versions is especially interesting since participants using version LH-W had to interact with a wizard and decide between different requirements for the tasks. This wizard requires some development and maintenance effort. The fact that we did not observe a significant difference in the effectiveness, efficiency, and perceived usability between the two Let's Hash versions suggests omitting the wizard might be prudent. Then, the secure code snippets will be directly presented to developers, as they were in version LH. In either case, users can simply copy and paste the presented code into their projects. Having a central resource that is known to contain up-to-date code may help to mitigate the difficulties that developers have when looking for and assessing resources [4, 27]. We hope that by expanding Let's Hash, its relevance will increase over time. One such addition could be an implementation of a Single-Sign On (SSO). We plan to publish Let's Hash and build an open-source community for researchers and developers.

9 Conclusion

Previous work showed that developers struggle to adhere to security best practices. Programming resources aimed at helping

developers work often are either complex, hard to understand and to use but secure, or easy to use but outdated and poorly maintained concerning security. To improve software security, we developed Let's Hash, a resource to support developers in implementing the security-critical authentication tasks: user password storage in a database, password policies, and 2FA.

The difference in security achieved with either version of Let's Hash compared to the developers' usual resources was highly significant. We further found that the two versions of Let's Hash did not differ significantly in either SUS score, time spent, or the number of clicks needed. The participants' perceived usability of both Let's Hash versions was excellent, and the participants' feedback was highly positive.

Our results indicated that Let's Hash has a great potential to improve the security of code that developers produce while also decreasing the effort needed. Consequently, we plan to deploy Let's Hash as a resource for developers and researchers. Future efforts could include incorporating additional topics, like SSO, or programming languages and more background information. Also, it might be helpful to explore how to best highlight security-critical parts of the code that should not be altered to mitigate some of the errors participants made while using Let's Hash. To keep Let's Hash well maintained, we will be releasing it as an open-source project on GitHub, and we hope to build a community.

Acknowledgments

This work was partially funded by the Werner Siemens Foundation, and the ERC Grant 678341: Frontiers of Usable Security. The authors would like to thank Martin Welsch, Manfred Paul and Ben Swierzy for their help in developing and evaluating Let's Hash. We thank our anonymous reviewers and shepherd for helping us improve our paper.

References

- [1] Ergonomics of human-system interaction - part 11: Usability: Definitions and concepts. Technical Report ISO 9241-11:2018, March 2018.
- [2] Josh Aas, Richard Barnes, Benton Case, Zakir Durumeric, Peter Eckersley, Alan Flores-López, J Alex Halderman, Jacob Hoffman-Andrews, James Kasten, Eric Rescorla, et al. Let's encrypt: an automated certificate authority to encrypt the entire web. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2473–2487, 2019.
- [3] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. Comparing the usability of cryptographic apis. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171, 2017.

- [4] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305. IEEE, 2016.
- [5] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. How internet resources might be helping you develop faster but less securely. *IEEE Security & Privacy*, 15(2):50–60, 2017.
- [6] Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8. IEEE, 2016.
- [7] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L Mazurek, and Sascha Fahl. Security developer studies with github users: Exploring a convenience sample. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 81–95, 2017.
- [8] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Commun. ACM*, 42(12):40–46, December 1999.
- [9] Aftab Alam, Katharina Krombholz, and Sven Bugiel. Poster: Let history not repeat itself (this time) – tackling webauthn developer issues early on. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19*, page 2669–2671, New York, NY, USA, 2019. Association for Computing Machinery.
- [10] Joël Alwen and Jeremiah Blocki. Towards practical attacks on argon2i and balloon hashing. In *2017 IEEE European Symposium on Security and Privacy (EuroSP)*, pages 142–157, 2017.
- [11] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.
- [12] Jason Bau, Frank Wang, Elie Bursztein, Patrick Mutchler, and John C Mitchell. Vulnerability factors in new web applications: Audit tools, developer selection & languages. *Stanford, Tech. Rep*, 2012.
- [13] Ruan Bekker. Salt and hash example using python with bcrypt on alpine, 2018. Last retrieved April 30, 2021 from <https://blog.ruanbekker.com/blog/2018/07/04/salt-and-hash-example-using-python-with-bcrypt-on-alpine/>.
- [14] Jeremiah Blocki, Benjamin Harsha, and Samson Zhou. On the economics of offline password cracking. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 853–871, 2018.
- [15] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- [16] BSI. Bundesamt für sicherheit in der informationstechnik, 2021. Last retrieved April 21, 2021 from https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Informationen-und-Empfehlungen/Cyber-Sicherheitsempfehlungen/Accountschutz/Sichere-Passwoerter-erstellen/sichere-passwoerter-erstellen_node.html.
- [17] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. “it’s not actually that horrible” exploring adoption of two-factor authentication at a university. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.
- [18] Nik Cubrilovic. Rockyou hack: From bad to worse, 2009. Last retrieved February 19, 2021 from <https://techcrunch.com/2009/12/14/rockyou-hack-security-myspace-facebook-passwords/>.
- [19] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. Replication: On the ecological validity of online security developer studies: Exploring deception in a password-storage study with freelancers. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 165–183. USENIX Association, August 2020.
- [20] Sanchari Das, Andrew Dingman, and L Jean Camp. Why johnny doesn't use two factor a two-phase usability study of the fido u2f security key. In *International Conference on Financial Cryptography and Data Security*, pages 160–179. Springer, 2018.
- [21] C. Dutrow and M. Amery. Salt and hash a password in python, 2019. Last retrieved April 30, 2021 from <https://stackoverflow.com/questions/9594125/salt-and-hash-a-password-in-python>.
- [22] Manuel Egele, David Brumley, Yanick Fratantonio, and Christopher Kruegel. An empirical study of cryptographic misuse in android applications. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 73–84. ACM, 2013.
- [23] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. Does

- my password go up to eleven? the impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2379–2388, New York, NY, USA, 2013. Association for Computing Machinery.
- [24] Sascha Fahl, Marian Harbach, Henning Perl, Markus Koetter, and Matthew Smith. Rethinking ssl development in an appified world. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 49–60, 2013.
- [25] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.
- [26] Juan M Ferreira, Silvia T Acuna, Oscar Dieste, Sira Vegas, Adrian Santos, Francy Rodriguez, and Natalia Juristo. Impact of usability mechanisms: An experiment on efficiency, effectiveness and user satisfaction. *Information and Software Technology*, 117:106195, 2020.
- [27] F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl. Stack overflow considered harmful? the impact of copy paste on android application security. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 121–136, 2017.
- [28] Flask. Flask-bcrypt, 2020. Last retrieved April 30, 2021 from <https://flask-bcrypt.readthedocs.io/en/latest/>.
- [29] Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. An administrator’s guide to internet password research. In *28th Large Installation System Administration Conference (LISA14)*, pages 44–61, 2014.
- [30] Alain Forget, Sonia Chiasson, and Robert Biddle. Helping users create better passwords: Is this the right approach? In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, pages 151–152, 2007.
- [31] Google. Google authenticator, 2020. Last retrieved March 23, 2020 from <https://github.com/google/google-authenticator>.
- [32] Paul A Grassi, James L Fenton, EM Newton, RA Perlner, AR Regenscheid, WE Burr, JP Richer, NB Lefkovitz, JM Danker, Yee-Yin Choong, et al. Nist special publication 800-63b: Digital identity guidelines. *Enrollment and Identity Proofing Requirements*. url: <https://pages.nist.gov/800-63-3/sp800-63a.html>, 2017.
- [33] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security & Privacy*, 14(5):40–46, 2016.
- [34] Aleksander Groth and Daniel Haslwanter. Efficiency, effectiveness, and satisfaction of responsive mobile tourism websites: a mobile usability study. *Information Technology & Tourism*, 16(2):201–228, 2016.
- [35] Troy Hunt. Have i been pwned. *Last retrieved*, 23, 2019.
- [36] Saranga Komanduri is a Phd. Helping users create better passwords. 2012.
- [37] Brandy Klug. An overview of the system usability scale in library website and system usability testing. *Weave: Journal of Library User Experience*, 1(6), 2017.
- [38] Philip Kortum and Claudia Ziegler Acemyan. The relationship between user mouse-based performance and subjective usability assessments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 1174–1178. SAGE Publications Sage CA: Los Angeles, CA, 2016.
- [39] Gitte Lindgaard and Cathy Dudek. *User Satisfaction, Aesthetics and Usability*, pages 231–246. Springer US, Boston, MA, 2002.
- [40] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. Jumping through hoops: Why do java developers struggle with cryptography apis? In *Proceedings of the 38th International Conference on Software Engineering, ICSE '16*, pages 935–946, New York, NY, USA, 2016. ACM.
- [41] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On conducting security developer studies with cs students: Examining a password-storage study with cs students, freelancers, and company developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
- [42] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "if you want, i can store the encrypted password" a password-storage field study with freelance developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.
- [43] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong? a qualitative usability study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 311–328, 2017.
- [44] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception task design in developer password studies: Exploring a student sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 297–313, 2018.

- [45] Mark O'Neill, Scott Heidbrink, Jordan Whitehead, Tanner Perdue, Luke Dickinson, Torstein Collett, Nick Bonner, Kent Seamons, and Daniel Zappala. The secure socket {API}:{TLS} as an operating system service. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 799–816, 2018.
- [46] OWASP. Open web application security project, 2021. Last retrieved April 26, 2021 from <https://owasp.org/>.
- [47] OWASP. Owasp authentication cheat sheet, 2021. Last retrieved February 20, 2021 from https://cheatsheetseries.owasp.org/cheatsheets/Authentication_Cheat_Sheet.html.
- [48] OWASP. Owasp password storage cheat sheet, 2021. Last retrieved February 20, 2021 from https://cheatsheetseries.owasp.org/cheatsheets/Password_Storage_Cheat_Sheet.html.
- [49] Sarah Perez. Recently confirmed myspace hack could be the largest yet, 2016. Last retrieved February 19, 2021 from <https://techcrunch.com/2016/05/31/recently-confirmed-myspace-hack-could-be-the-largest-yet/>.
- [50] PHC. Password hashing competition, 2019. Last retrieved February 20, 2021 from <https://www.password-hashing.net/>.
- [51] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. <https://www.R-project.org/>.
- [52] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A tale of two studies: The best and worst of yubikey usability. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 872–888. IEEE, 2018.
- [53] RockIt. The most common passwords 2020, 2020. Last retrieved February 19, 2021 from <https://rockit.cloud/2020/03/18/the-most-commonly-used-password-in-2020-is/>.
- [54] Jeff Sauro. 5 ways to interpret a sus score, 2018. Last retrieved April 25, 2021 from <https://measuringu.com/interpret-sus-score/>.
- [55] Jeff Sauro and James R. Lewis. Average task times in usability tests: What to report? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 2347–2350, New York, NY, USA, 2010. Association for Computing Machinery. <https://doi.org/10.1145/1753326.1753679>.
- [56] Sean M. Segreti, William Melicher, Saranga Komanduri, Darya Melicher, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. Diversify to survive: Making passwords stronger with adaptive policies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 1–12, Santa Clara, CA, July 2017. USENIX Association.
- [57] Steve Sheng, Levi Broderick, Colleen Alison Koranda, and Jeremy J Hyland. Why johnny still can't encrypt: evaluating the usability of email encryption software. In *Symposium On Usable Privacy and Security*, pages 3–4. ACM, 2006.
- [58] Christian Stransky, Yasemin Acar, Duc Cuong Nguyen, Dominik Wermke, Doowon Kim, Elissa M. Redmiles, Michael Backes, Simson Garfinkel, Michelle L. Mazurek, and Sascha Fahl. Lessons learned from using an online platform to conduct large-scale, online controlled security experiments with software developers. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)*, Vancouver, BC, August 2017. USENIX Association.
- [59] Joshua Tan, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Practical recommendations for stronger, more usable passwords combining minimum-strength, minimum-length, and blocklist requirements. 2020.
- [60] Dan U. Passwords, passwords everywhere, 2019. Last retrieved February 19, 2021 from <https://www.ncsc.gov.uk/blog-post/passwords-passwords-everywhere>.
- [61] Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, et al. Design and evaluation of a data-driven password meter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3775–3786, 2017.
- [62] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. How does your password measure up? the effect of strength meters on password creation. In *21st USENIX Security Symposium (USENIX Security 12)*, pages 65–80, Bellevue, WA, August 2012. USENIX Association.
- [63] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "I Added'!' at the End to Make It Secure": Observing Password Creation in the Lab. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 123–140, 2015.

- [64] Vitosh. Vitosh academy, 2021. Last retrieved April 25, 2021 from <https://www.vitoshacademy.com/>.
- [65] W3Schools. W3schools online web tutorials, 2021. Last retrieved April 25, 2021 from <https://www.w3schools.com/>.
- [66] Daniel Lowe Wheeler. zxcvbn: Low-budget password strength estimation. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 157–173, 2016.
- [67] Alma Whitten and J Doug Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *USENIX Security Symposium*, volume 348, pages 169–184, 1999.
- [68] Chamila Wijayarathna and Nalin A. G. Arachchilage. Why johnny can't store passwords securely? a usability evaluation of bouncycastle password hashing. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018, EASE'18*, page 205–210, New York, NY, USA, 2018. Association for Computing Machinery.

A Task descriptions

These are the task descriptions as they were presented to participants:

Create a method to hash and salt passwords for storage in a database.

You are asked to develop a method in a web-application backend that prepares a password for storage in a database. Assume that a user has chosen a password that gets handed to your function `hash_password()` as a string. Implement this function in such a way that it returns the password securely hashed and salted. Additionally, please implement a function `verify()`, which compares a password to a hash and returns True if they match, False if they do not.

The programming language for this task is Python3. Please only use the website LetsHash as a resource when solving this task.

When is the problem solved?

The problem is solved when you have successfully implemented the function to fulfill the required specifications, and the output printed by the main function reads:

“Your hash: <a hash value>

The correct password is s3cr3t: True

The correct password is s3cr4t: False”

Create a method to check if a password adheres to a given policy.

You are asked to develop a method in a web-application frontend that ensures that the password a user chooses meets the company policy. According to this policy, a password must

- be between 8 and 64 characters
 - have at least one upper- and one lowercase letter
 - have at least one special character - have at least one number
- Please implement the functions `length()` and `composition()` so they

return True if the password matches the criteria given, and False if it does not.

The programming language for this task is JavaScript. Please only use the website LetsHash as a resource when solving this task.

When is the problem solved?

The problem is solved when you have successfully implemented the functions to fulfill the required specifications, and upon clicking “Run and Test”, you receive an alert that reads:

“The password meets the requirement for length: true

The password meets the requirement for composition: false

The password is considered valid: false”

Create a method to set up a second factor for user authentication.

You are asked to develop a method in a web-application backend that offers users of a login system a second factor for authentication. Please implement the function `generate_second_factor()` so that it takes a shared secret as a parameter and returns a time-based one-time password (totp).

The programming language for this task is Python3. Please only use the website LetsHash as a resource when solving this task.

When is the problem solved?

The problem is solved when you have successfully implemented the function to fulfill the required specifications, and the output printed by the program reads:

“Your code: <a time-based one-time code>

Your code is verified: True”

B Surveys

Some questions of the surveys were specific to either the groups LH and LH-W, or group C. These questions are marked accordingly.

Survey

Thank you very much for working on the tasks assigned to you during this study! There are a few questions we would like to ask you to wrap things up.

- (Q1): Please answer the following questions by indicating a number on the scale from "1 - Not at all familiar" to "7 - Very familiar".
 - How familiar are you with Python? *I-7*
 - How familiar are you with Javascript? *I-7*
 - How familiar are you with password storage in a database? *I-7*
 - How familiar are you with the implementation of two factor authentication (2fa)? *I-7*
 - How familiar are you with the implementation of password policies? *I-7*

- (Q2): Have you ever looked up how to implement password policies as they are recommended by any of the following institutions before this study? You can choose more than one answer.
 - No, I have never looked up recommendations on password policies.
 - I have looked up NIST's recommendations on password policies.
 - I have looked up OWASP's recommendations on password policies.
 - I have looked up another institution's recommendations on password policies. Please specify: (*Free text*)
- (Q3): Please rate the correctness of the following statements by indicating a number on the scale from "1 - Does not describe me" to "7 - Describes me very well".
 - I am familiar with the implementation of login forms. *1-7*
 - I am familiar with the implementation of password strength checkers. *1-7*
 - I have a good understanding of security concepts. *1-7*
- (Q4): How long were you actively working on the task to solve it? Please indicate the time in full hours. (*Free text*)
- (Q5): Please answer the following question by indicating a number on the scale from "1 - Very easy" to "7 - Very hard". Overall, the task was... *1-7*
- (Q6): Please rate the correctness of the following statement by indicating a number on the scale from "1 - Not close at all" to "7 - Very close". How close was the task to reality compared to the projects that you develop in everyday life? *1-7*
- (Q7): Did you have any prior experience with storing passwords in a database? You can choose more than one answer.
 - No.
 - Yes, in university.
 - Yes, on a job.
 - Other - please specify: (*Free text*)
- **Only groups LH/ LH-W (Q8):** If yes: Please rate the correctness of the following statement by indicating a number on the scale from "1 - Not at all helpful" to "7 - Very helpful". Would the website you have used in this study have been helpful in solving problems you had then? *1-7*
- (Q9): Did you have any prior experience with implementing password policies? You can choose more than one answer.
 - No.
 - Yes, in university.
 - Yes, on a job.
 - Other - please specify: (*Free text*)
- **Only groups LH/ LH-W (Q10):** If yes: Please rate the correctness of the following statement by indicating a number on the scale from "1 - Not at all helpful" to "7 - Very helpful". Would the website you have used in this study have been helpful in solving problems you had then? *1-7*
- (Q11): Did you have any prior experience with implementing two-factor authentication? You can choose more than one answer.
 - No.
 - Yes, in university.
 - Yes, on a job.
 - Other - please specify: (*Free text*)
- **Only groups LH/ LH-W (Q12):** If yes: Please rate the correctness of the following statement by indicating a number on the scale from "1 - Not at all helpful" to "7 - Very helpful". Would the website you have used in this study have been helpful in solving problems you had then? *1-7*
- **Only groups LH/ LH-W (Q13):** What could be improved about the website? (*Free text*)
- (Q14): Please answer the following questions by indicating a number on the scale from "1 - Never" to "7 - Every time".
 - How often do you ask for help when faced with security problems? *1-7*
 - How often are you asked for help when others are faced with security problems? *1-7*
 - How often do you need to add security to the software you develop in general (apart from this study)? *1-7*
- (Q15): Please answer the following questions by indicating a number on the scale from "1 - Not knowledgeable at all" to "7 - Very knowledgeable".
 - How would you rate your background/knowledge with regard to secure password storage in a database? *1-7*

- How would you rate your background/knowledge with regard to the implementation of two factor authentication (2fa)? *I-7*
- How would you rate your background/knowledge with regard to the implementation of password policies? *I-7*
- (Q16): How often have you stored passwords in a database in the software you have developed (apart from this study)? *(Free text)*
- (Q17): How often have you implemented two factor authentication (apart from this study)? *(Free text)*
- (Q18): How often have you implemented a login form with a password strength checker (apart from this study)? *(Free text)*
- (Q19): What is your most-used resource for implementing security in your software development?
 - Stackoverflow
 - Official documentation
 - Other - please specify: *(Free text)*
- **Only groups LH/ LH-W:**
 - (Q20-LH): Please rate your agreement to the following questions on a scale from "1 - Strongly disagree" to "7 - Strongly agree".
 - * I needed a lot of background knowledge to complete the task. *I-7*
 - * The website provided well-structured information. *I-7*
 - * The website provided all necessary information to solve the task. *I-7*
 - * I spent a lot of time trying to navigate the website. *I-7*
 - * The assistance provided by the website to ease navigation was sufficient. *I-7*
 - * I would recommend this website to a colleague who needs assistance with the implementation of password storage. *I-7*
 - * I would recommend this website to a colleague who needs assistance with the implementation of two factor authentication. *I-7*
 - * I would recommend this website to a colleague with questions regarding the implementation of password policies. *I-7*
 - * I would use this website if I had to work on a similar task in a professional setting/ working on tasks within my job. *I-7*
 - (Q21-LH): Have you used only the website that was provided to you by this study? If not, which additional resources did you use to solve the tasks?
 - * I have only used the website that was provided to me
 - * I have used other resources as well: *(Free text)*
- (Q22-LH): Please answer the following question by indicating a number on the scale from "1 - Much better" to "7 - Much worse". Compared to your most used resource, how would you rate the ease of use of the website you worked with during this study when it comes to accomplishing your tasks **functionally**? *I-7*
- (Q23-LH): Please explain your decision: *(Free text)*
- (Q24-LH): Please answer the following question by indicating a number on the scale from "1 - Much better" to "7 - Much worse". Compared to your most used resource, how would you rate the ease of use of the website you worked with during this study when it comes to accomplishing your tasks **securely**? *I-7*
- (Q25-LH): Please explain your decision: *(Free text)*
- (Q26-LH): Please rate your agreement to the following statements about the website that was provided for you during this study on a scale from "1 - Strongly disagree" to "5 - Strongly agree".
 - * I think that I would like to use this website frequently. *I-5*
 - * I found the website unnecessarily complex. *I-5*
 - * I thought the website was easy to use. *I-5*
 - * I think that I would need the support of a technical person to be able to use this website. *I-5*
 - * I found the various functions in this website were well integrated. *I-5*
 - * I thought there was too much inconsistency in this website. *I-5*
 - * I would imagine that most people would learn to use this website very quickly. *I-5*
 - * I found the website very cumbersome to use. *I-5*
 - * I felt very confident using the website. *I-5*
 - * I needed to learn a lot of things before I could get going with this website. *I-5*
- **Only group C:**
 - (Q20-C): Which resources did you use to solve the tasks? Please be as specific as possible (for example, provide links to any websites you used). *(Free text)*
 - (Q21-C): Which of the resources you listed in the last question was your main resource? *(Free text)*

- (Q22-C): Please answer the following question by indicating a number on the scale from "1 - Very good" to "7 - Very bad". How would you rate the ease of use of the website(s) you worked with during this study when it comes to accomplishing your tasks **functionally**? 1-7
- (Q23-C): Please explain your decision: (*Free text*)
- (Q24-C): Please answer the following question by indicating a number on the scale from "1 - Very good" to "7 - Very bad". How would you rate the ease of use of the website(s) you worked with during this study when it comes to accomplishing your tasks **securely**? 1-7
- (Q25-C): Please explain your decision: (*Free text*)
- (Q26-C): Please rate your agreement to the following questions on a scale from "1 - Strongly disagree" to "7 - Strongly agree".
 - * I needed a lot of background knowledge to complete the task. 1-7
 - * The website(s) I used provided well-structured information. 1-7
 - * The website(s) I used provided all necessary information to solve the task. 1-7
 - * I spent a lot of time trying to navigate the website(s) I used. 1-7
 - * I would use the same website(s) if I had to work on a similar task in a professional setting/ was working on tasks within my job. 1-7
- (Q27-C): Please rate your agreement to the following statements about the website that was your main resource during this study on a scale from "1 - Strongly disagree" to "5 - Strongly agree".
 - * I think that I would like to use this website frequently. 1-5
 - * I found the website unnecessarily complex. 1-5
 - * I thought the website was easy to use. 1-5
 - * I think that I would need the support of a technical person to be able to use this website. 1-5
 - * I found the various functions in this website were well integrated. 1-5
 - * I thought there was too much inconsistency in this website. 1-5
 - * I would imagine that most people would learn to use this website very quickly. 1-5
 - * I found the website very cumbersome to use. 1-5
 - * I felt very confident using the website. 1-5
 - * I needed to learn a lot of things before I could get going with this website. 1-5
- (Q28): Please select your gender.
 - Male
 - Female
 - Prefer not to say
 - Other: (*Free text*)
- (Q29): Please state your age. (*Free text*)
- (Q30): What is your current main occupation?
 - Freelance developer
 - Industrial developer
 - Industrial researcher
 - Academic researcher
 - Undergraduate part-time student
 - Undergraduate full-time student
 - Graduate part-time student
 - Graduate full-time student
 - Other: (*Free text*)
- (Q31): What is your nationality? (*Free text*)
- (Q32): How did you gain your IT skills? (*Free text*)
- (Q33): What was your main source of learning about IT-security? (*Free text*)
- (Q34): Do you have a university degree? (*Yes/ No*)
- If yes:
 - (Q35): What was/is your subject? (*Free text*)
 - (Q36): Were/Are you taught about IT-security at university? (*Free text*)
 - (Q37): Were/Are you taught about IT-security in addition to your regular studies? (*Yes/ No*)
 - (Q38): If yes: Where were/are you taught about IT-security in addition to your regular studies? (*Free text*)
- (Q39): Are you working at a company? (*Yes/ No*)
- If yes:
 - (Q40): How old is your organization? Please specify in years. (*Free text*)
 - (Q41): What is the total number of employees in your organization?
 - * 1-9
 - * 10-249
 - * 250-499
 - * 500-999

- * 1000 or more
- (Q42): How many members are there in your team? (*Free text*)
- (Q43): Which field of activity does your company belong to? You can choose more than one answer.
 - * Game development
 - * Development of network and communication software
 - * Web development
 - * Development of middleware, system components, libraries and frameworks
 - * Development of other tools for developers, such as IDEs and compilers
 - * Other: (*Free text*)
- (Q44): Does your company have a security focus? (*Yes/ No*)
- (Q45): Does your team have a security focus in its current field of activity?
 - * Yes
 - * No
 - * I work alone and my field of activity has a security focus
 - * I work alone and my field of activity has no security focus
- (Q46): Do you also have to work on security-relevant tasks in your field of activity? (*Yes/ No*)
- (Q47): Were/Are you taught about IT-security in addition to your regular work? (*Yes/ No*)
- (Q48): If yes: Where were/are you taught about IT-security in addition to your regular work? (*Free text*)
- (Q49): What type(s) of software do you develop? You can choose more than one answer.
 - Web applications
 - Mobile/App applications
 - Desktop applications
 - Embedded Software Engineering
 - Enterprise applications
 - Other - please specify: (*Free text*)
- (Q50): How many years of experience do you have with software development in general? (*Free text*)

C Security Score

For the security evaluation of the code, we used an adapted version of this security score from Naiakshina et al. [43]:




- (Q51): How many years of experience do you have with Python development? (*Free text*)
- (Q52): How many years of experience do you have with Javascript development? (*Free text*)
- (Q53): If you have any comments or suggestions, please leave them here: (*Free text*)
 1. The end-user password is salted (+1) and hashed (+1).
 2. The derived length of the hash is at least 160 bits long (+1).
 3. The iteration count for key stretching is at least 1000 (+0.5) or 10000 (+1) for PBKDF2 and at least $2^{10} = 1024$ for bcrypt (+1).
 4. A memory-hard hashing function is used (+1).
 5. The salt value is generated randomly (+1).
 6. The salt is at least 32 bits in length (+1).

D Automated Detection Of Copied Code

This section describes the process which was used to semi-automatically determine whether participants from the groups LH and LH-W submitted code which they had copied off of Let's Hash.

Since participants would sometimes copy everything, including comments, while others only copied the exact lines that were needed, exact string matching would have been too strict for our purposes. We used approximate string matching, coded in python, to calculate a matching ratio. If the resulting ratio dropped below a threshold of 80% for the cryptographic tasks, or 50% for the task on password policies, the files were examined manually. The thresholds were chosen on the basis of manual spot sampling. The threshold for the password policy task was much lower than for the other two tasks because this task involved some changes to the code that participants had to make to be able to use it, while the code for the other two tasks could be used as is.

Exploring User Authentication with Windows Hello in a Small Business Environment

Florian M. Farke , Leona Lassak , Jannis Pinter, and Markus Dürmuth[‡] 
Ruhr University Bochum, [‡]Leibniz University Hannover

Abstract

Windows Hello for Business is Microsoft’s latest attempt to replace passwords in Windows enterprise environments introduced with Windows 10. It addresses some of the common password problems like password leaks or phishing attacks, comes with built-in support for biometric authentication methods like fingerprint or facial recognition, and a new user interface. We conducted a qualitative study with 13 employees accompanying the introduction of Windows Hello in a small business studying its usability and deployability. Over five weeks, we measured authentication times, let participants rate their user experience, and conducted interviews at the end. In general, participants liked Windows Hello and found it more usable than the traditional Windows sign-in scheme. Windows Hello was faster and perceived as more responsive than the traditional Windows login. However, participants tended to use PINs as a replacement for their (longer) passwords instead of using biometrics. Lack of hardware support (no biometric hardware available), the form factor of device or setup of their workplace (e.g., biometric hardware on the other side of the table) were some reasons to not use biometrics but stick with a well-known authentication method like a PIN.

1 Introduction

Replacing the omnipresent username and password scheme for authentication has become an ongoing quest in the usable security research community and parts of the software industry. Still, passwords are the most common approach to authenticate humans on digital devices, even though they have substantial drawbacks in terms of both usability and security:

Passwords can be phished or leaked, are often reused or hard to remember, easily guessable for password-crackers, and hard to use on devices without a physical keyboard [9, 31].

Despite the weaknesses of password-based authentication only a few of the proposed alternatives found broader adoption: Graphical passwords suffer from similar drawbacks as text passwords, but are better in terms of memorability and input behavior on small touchscreens, and are used for smartphone unlocking [4, 37]. Security tokens, typically in the form of two-factor authentication (2FA), found some use in corporate contexts where setup and management of the tokens are done by an IT department or for online services with high security requirements (e.g., online banking). Regular online services usually did not offer support for these tokens because they need extra care in case of loss or theft, and in the past were often incompatible when coming from different vendors (this problem is probably solved through FIDO2¹), and perhaps not wanted by the users [10]. Also, biometrics are not well suited for authentication at online services, as allowing the service provider to store biometric data poses a privacy and security risk to the user because biometric factors can not be changed after a data leak or when using another service. Furthermore, the provider requires access to the biometric hardware to perform the authentication.

In contrast, biometric authentication is well suited for local authentication such as smartphone unlock or sign-in on a desktop computer and can be implemented without the biometric data leaving the device. This local authentication can then be used to unlock cryptographic secrets stored in a secure enclave.

In recent years, this approach of unlocking stored login credentials was adopted in several authentication protocols (e.g., in FIDO2) and products, like for example Microsoft’s Windows Hello for Business which was first introduced with Windows 10. Windows Hello for Business replaces the traditional Windows login (i.e., with username and password) with certificate-based authentication in which the private key

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

¹<https://fidoalliance.org/fido2/>, as of June 9, 2022

of the user is stored locally and unlocked via facial or fingerprint recognition, security token, or PIN. Since Microsoft Windows still dominated the desktop operating systems market in 2021 (approx. 75% market share² of which roughly 82% was Windows 10³) and Microsoft's strategy of encouraging their customers to use the latest version of their operating system, we assume that Windows Hello for Business will replace the traditional Windows login in companies in the long run.

In this work, we accompanied the introduction of Windows Hello for Business in a small company, investigating the usability and perceived security of the new sign-in method. We were particularly interested in which benefits and challenges participants see when using Windows Hello and which authentication options (facial recognition, fingerprint, or PIN) they prefer to use and why. For our case study, we recruited 13 employees to voluntarily participate during their working hours. Over the course of five weeks, we followed the transition from traditional Windows authentication (with username and password or smart card) to Windows Hello for Business.

In detail, we explored the following question:

- RQ1** What are the usability differences between Windows Hello for Business and the traditional Windows sign-in? (*Usability Comparison*)
- RQ2** What is the perceived security of Windows Hello for Business? (*Perceived Security*)
- RQ3** Which authentication options of Windows Hello for Business are people willing to adopt and why? (*Use of Biometrics*)

Our case study sheds some light on usability aspects that are important when deploying a new sign-in method and provides insights into why people adopt (or not adopt) biometric authentication in the corporate context which may also apply to contexts other than Windows Hello.

2 Background: Windows Hello for Business

With the release of Windows 10 in 2015, Microsoft introduced *Windows Hello* and *Windows Hello for Business* as new options to authenticate on a Windows computer. Both variants allow authentication using hardware tokens, biometrics (e.g., facial or fingerprint recognition), or picture passwords⁴. Text-based passwords or six-digit PINs are still available for cases in which biometric hardware is not available, is not accessible for certain user groups (e.g., people

²<https://gs.statcounter.com/os-market-share/desktop/worldwide/2021>, as of June 9, 2022

³<https://gs.statcounter.com/windows-version-market-share/desktop/worldwide/2021>, as of June 9, 2022

⁴<https://docs.microsoft.com/en-us/windows/security/identity-protection/hello-for-business/hello-overview>, as of June 9, 2022

with impairments), or can not be used due to other restrictions (e.g., company policy).

As the name suggests, Windows Hello for Business integrates into Microsoft's enterprise authentication solutions while Windows Hello was designed for the consumer versions of Windows 10. Even though both variants provide the very same user interface, the inner workings and security features are quite different to meet the different use cases (i.e., authenticating via an authentication server in an enterprise network versus local authentication on a personal device). Windows Hello unlocks a password, which is stored encrypted, that is then used for authentication. In contrast to that, Windows Hello for Business is built upon public-key cryptography and uses certificates to authenticate against a remote authentication server. A Trusted Platform Module (TPM), when available, is used to securely store the login credentials and to perform cryptographic operations.

Since the login credentials are stored locally on the device and unlocked via Windows Hello, every device must be registered with a Windows account before Windows Hello for Business can be used. This binding to a device marks a paradigm shift from a knowledge-based authentication scheme (i.e., username and password) to a combination of several factors such as knowledge/biometrics and possession (i.e., the device). Such a possession-based authentication scheme is more secure because potential attackers have to gain physical control over the device instead of simply stealing the password remotely. However, requiring the use of a specific device may not be feasible for every use case, especially, for people who frequently sign in to different devices, Windows Hello for Business is not a usable solution.

3 Related Work

Windows Hello (for Business) has not been studied extensively, especially there is little research with a focus on usability. For completeness, we report the research on Windows Hello and Windows picture passwords here. However, the research is not comparable or closely related to our study. Kim et al. [22] analyzed the security of Windows Hello and propose a migration attack to compromise Windows Hello's authentication data. This attack is only applicable on devices without hardware protection. In our study, all participants used devices with TPMs where the attack is not applicable.

Issues with the Windows 8 graphical password scheme were identified by Gao et al. [17]. They studied user choices of graphical passwords in the lab and the field finding that significant hotspots exist which can be exploited in an attack.

Studying the influences of human cognition on password strength in picture passwords, Katsini et al. [21] conducted an eye-tracking study using Windows Picture Passwords. However, their goal was not to study usability or security aspects of the authentication mechanism but just used it as a working example of picture passwords for their research.

Passwordless Authentication Since Windows Hello for Business is an alternative to passwords and thus a form of passwordless authentication, in the following, we discuss recent literature in this field. The FIDO2 protocols are the latest proposal for passwordless authentication. Most related in terms of methodology is a study by Farke et al. [16] in which they compared the use of security keys and passwords in a small software company. The participants found the security key to be slower than using their password manager and due to further usability issues, several employees stopped using the key despite its security benefits.

Lyastani et al. [26] compared user perceptions of passwordless FIDO2 security key logins to signing in with a site-specific password. While participants preferred the security key over passwords, the hardware-related shortcomings like account access on devices without USB ports questioned the keys' real-world suitability for passwordless authentication. FIDO2-related issues like key recovery and account revocation in case the key is lost or stolen were also mentioned.

Passwordless *biometric* FIDO2 was first studied by Oogami et al. [29] who documented the WebAuthn registration process with 10 participants on their existing Yahoo! Japan accounts. Issues with the user interface design, like a fingerprint icon being mistaken for the fingerprint reader, were identified. These results, however, are mainly relevant for the specific design of the Yahoo! Japan website.

Misconceptions about biometric FIDO2 and how to mitigate them was studied by Lassak et al. [24]. First, 42 crowdworkers used biometric WebAuthn to log in to a website and answered a questionnaire about misconceptions surrounding the login. 67% of the participants assumed that the biometric information would be transmitted to the website. In focus groups, the researchers then designed several notifications and with 345 crowdworkers investigated how they could be used to counteract users' misconceptions. The researchers found that some of the notifications partially addressed misconceptions, but misconceptions about where the biometric is stored partially persisted.

Less related but also focused on FIDO2 passwordless authentication is research by Owens et al. [30]. 97 participants logged in to a fictitious bank website over the course of two weeks with either a password or a smartphone as a FIDO2 roaming authenticator (via a prototype protocol called Neo). While Neo's security benefits were recognized by participants, login times with Neo were substantially higher than for passwords. Participants also recognized availability concerns regarding account recovery and availability of the phone.

4 Method

We designed our study to observe, in the context of a small business, the benefits employees see in using Windows Hello for Business but also what challenges they face. The em-

ployees voluntarily participated in the study using Windows Hello for Business on their work computers during their regular work time. For five weeks, we gathered sign-in data from each participant using a custom survey software application installed on the participants' work devices. Week one captured their interaction and satisfaction with their previous login method (password or smartcard). Week 2 to 5 did the same for Windows Hello. To get in-depth feedback on their experience with Windows Hello for Business, we conducted interviews at the end of the study. In the following, we outline our study protocol and explain relevant aspects of our survey application.

4.1 Study Procedure

The study procedure consists of three parts, where the second part is subdivided into two phases: (i) An initial workshop, in which we introduced the study, its procedure, and the participants' task; (ii) A five weeks long data collection phase, in which we measured authentication timings and gathered quantitative and qualitative feedback about logins consisting of: (a) One week of using a traditional password or smartcard-based login and (b) four weeks of using Windows Hello for Business; (iii) Final interviews to learn about the participants' experience using Windows Hello. An overview of the study procedure can be seen in Figure 1.

Part of the quantitative feedback was a User Experience Questionnaire (UEQ) [25]. We used the German version consisting of 26 pairs of contrasting items describing aspects of usability and user experience. These items belong to the six different categories *Attractiveness*, *Perspicuity*, *Efficiency*, *Dependability*, *Stimulation*, and *Novelty*. The only modification we performed was replacing the term *product* by *sign-in method* to fit the context of our study.

Initial Workshop To introduce our study, explain Windows Hello, and recruit participants, we invited all eligible employees to a 15-minute workshop. We began by communicating our study's purpose, procedure, and the associated risks. Since Windows Hello for Business requires to set up a PIN, we proceeded by briefing the potential participants on rules for choosing secure and memorable PINs.

By the end of the workshop, we handed out consent forms for participants to read and sign if they agreed to participate. Participation was voluntary and the entire time participants had to invest in the study took place during their working hours. In-person workshops are a typical format in which new technology is introduced in this specific company.

Using the Established Mechanism (week one) In the first week (Phase 1) of our study, we measured our baseline, with participants continuing to use their established authentication scheme (passwords or smartcards). We collected data

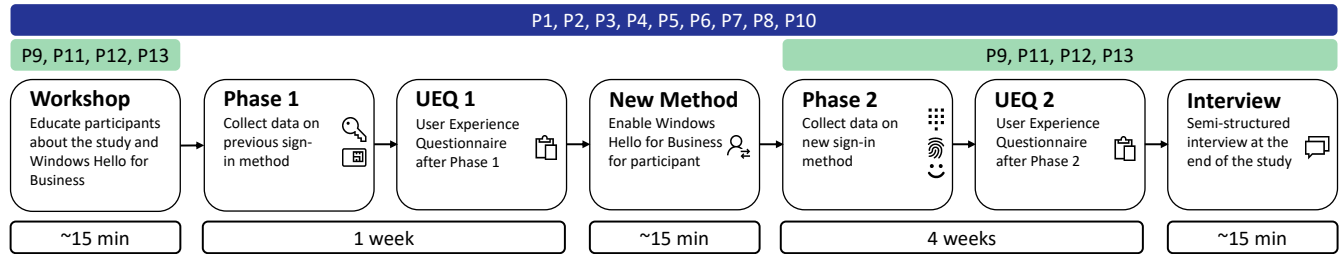


Figure 1: The study was divided into two phases: (1) The participants used the traditional Windows login for one week and filled out an UEQ afterwards; (2) They used Windows Hello for Business for four weeks and again filled out an UEQ. We conducted interviews with each participant after they finished Phase 2.

on the usage and experience with participants’ previous sign-in methods via a self-developed survey pop-up we call *SurveyApp*. It automates the process of a diary study which was previously used in a paper-pencil format in similar studies [16, 23]. After every login, the SurveyApp appeared and asked the participant to rate their satisfaction with the sign-in on a five-point Likert scale. Additionally, participants had the option to add a comment. We describe the *SurveyApp* in more detail in Section 4.3. At the end of this week, participants filled in a UEQ as described above.

Using Windows Hello (week two to five) Following Phase 1, Windows Hello for Business was enabled for the participants’ user accounts. To keep the changeover time as short as possible, all participants received individual assistance from the IT department during the setup. This approach ensured the best possible onboarding process and thus the satisfaction in the early study stages might reflect an upper bound. However, we think that it had little influence on the longterm usage which was our main study focus.

Once Windows Hello was successfully configured, the IT department disabled the ability to sign in with a password for the participants, leaving Windows Hello for Business as the only sign-in method available. Phase 2 was used to collect data on the usage and experience with Windows Hello for Business, in a rather normal usage context. Those participants whose devices had biometric capabilities could choose freely between authenticating via PIN or with their biometrics. Again, we collected the participants’ satisfaction with the logins via our *SurveyApp* (cf. Section 4.3). This phase lasted four weeks, subsequently, participants completed a UEQ. Since users were free to choose between the different Windows Hello sign-in options, the UEQ just represents the overall experience with Windows Hello.

Interviews To gather more fine-grained feedback from the participants, we followed up with 15-minute, semi-structured, one-on-one interviews. We aimed to explore participants’ impressions, feelings, and attitudes about and towards Windows Hello in more detail. The interviews were conducted and transcribed in German (see Appendix A).

We started the interviews by discussing the participants’ overall perception of the new sign-in method, as well as the differences to the old, password or smartcard-based approach. For those with biometric sign-in options, we asked the participants how of they have used biometric sign-in in comparison to other sign-in options. We also asked for participants’ opinions and potential general reservations towards biometrics and whether any issues occurred during the four weeks period. To compare participants’ impression of authentication speed with our time measurements, we asked the interviewees to gauge how much time they usually spend per login with Windows Hello and whether this time differs from the traditional Windows sign-in. This helped us to better understand potential reasons for preferences of one or the other sign-in option. Lastly, we were interested in the participants’ security perception of Windows Hello compared to their previous sign-in method. We concluded the interviews with questions about the participants’ satisfaction with Windows Hello, letting them specify pros and cons, whether they would be use Windows Hello on their personal devices and if they were willing to continue using Windows Hello.

Based on the interview questions, we used an a priori coding approach to analyze the interviews [12]. The researcher who conducted the interviews created an initial set of codes using the sections of the questionnaire (see Appendix A) as themes under which one or more codes were grouped. We discussed and refined this codebook as a group to specifically focus the interview analysis on our Research Questions **RQ2** and **RQ3**. One researcher coded all 11 interviews with the initial codebook. During this coding sessions a few codes in the *Reservation Against Biometrics* Section were added We discussed the changes in the codebook and removed unused codes (e.g., the code *Password is faster* or *Prefers Fingerprint*). The revised codebook is presented in Appendix B.

Using the revised codebook, another researcher coded all interviews again. To determine the inter-coder agreement, we used the coefficient Kappa of Brennan and Prediger [7] (an improved version of Cohens’ Kappa). The second coder reached a substantial agreement of $\kappa = 0.72$.

4.2 Study Environment, Recruitment, and Participants

The study was conducted at a small German company specialized in high-quality furniture, interior design, and store fitting. This environment was particularly interesting to study Windows Hello in because the needs in terms of authentication are very diverse. Employees included those in production where multiple users share a single machine, executives and sales personnel that often authenticate in public while traveling, and accountants and draftsmen who work in a regular office environment. Out of the 20 employees who regularly access computer workplaces we invited 15 to participate in our study. The remaining five were working part-time or leaving the company soon so we excluded them from our participant pool. Everyone who was invited agreed to participate. We had technical issues with the Windows Hello setup for one participant and another left the company during the study so we excluded their data from our analyses.

Due to vacation or sickness, the study start and end dates of the participants varied. The total time frame for the entire study ranged from August 2021 when we held our initial workshop to November 2021 when our last participant was interviewed. The study was conducted during the Covid-19 pandemic, however this did not specifically influence the study environment or work flows in the company at that time.

Before the study, all except four participants had only used passwords or smartcards to sign in at their workplace. The four other participants had been part of a pilot test of Windows Hello that had been carried out by the company before our study. This pilot test solely tested the migration from passwords to Windows Hello from a technical perspective and was independent from our study. Since they already had three to four months of experience with Windows Hello, they did not participate in Phase 1 of the study as it would have been confusing to switch them back to passwords just for the purpose of the study.

Demographics Out of the 15 eligible employees, 13 participated. Out of these, five were women and eight were men. 40% were aged between 18 and 29, 20% between 30 and 49, and 40% were 50+. None of the participants had a background in IT. Participants' job positions ranged from Engineering and Design over Production Planning to Administration and Executive.

Computer Hardware and Password Policy The computer hardware of our participants varied in terms of model and also authentication capabilities. Most computers were Lenovo machines including different models of the *ThinkStation*, the *ThinkCentre*, and the *ThinkPad*. Other participants used a Microsoft Surface Pro 7+. All 13 computers had a built-in TPM and could be unlocked via PIN. Additionally, four out of the 13 machines had a fingerprint reader,

3 offered fingerprint and face recognition, and one offered only face recognition. During the study, all computers ran on Windows 10 Professional with build number 19043 (21H1), which was the latest release of Windows 10 at that time. The machines received monthly security updates by Microsoft, but no feature updates were installed during the time of the study to obviate issues as much as possible. Per the company's policy, all computers are locked automatically after 10 minutes of inactivity.

The company's password policy only specified a length of at least 10 characters. Further complexity requirements such as the use of upper/lower case letters, numbers, or special characters were not specified. Also, the company did not enforce regular password changes.

4.3 Implementation


SurveyApp To collect satisfaction ratings from our participants continuously during the entire period of the study, we developed a GUI application (which we call *SurveyApp*). It was displayed immediately after every sign-in the participants performed even before the actual desktop screen was shown, and participants could rate their experience on a 5-point emoji scale (cf. Figure 2). We chose emojis for their intuitive meaning and quick interaction [1]. To unify the interface and prevent misinterpretation due to differing renderings on different machines [27], we displayed the emojis as images. To further prevent misinterpretation, each emoji was equipped with a tooltip showing the satisfaction level as text, they were displayed in ascending order, and explained during our initial workshop. The SurveyApp also provided the option to submit voluntary comments which was, however, barely used by our participants.

Time Measurements To our knowledge, no Windows API exists that provides data on the authentication of a user, specifically the exact time frame when the user first starts the sign-in to when the authentication is finished. The Microsoft Windows lock screen *LogonUI* changes memory consumption deterministically depending on the user interaction. For example, dismissing the lock screen and the submission of credentials can be seen as separate events in the memory traces. An example of such a memory trace can be seen in Appendix C. We decided to use the memory profile as a side-channel to measure the authentication timings. To capture all sign-ins, our background application periodically checks the process table for a running *LogonUI* process and when it is present, records the processes' memory usage every 250 milliseconds. Since it is unlikely that a sign-in takes longer than 60 seconds and to limit the memory footprint of the background service itself we only preserved the last 60 seconds of memory usage data.

We are aware that these timings are not entirely accurate since users, e.g., might dismiss the lock screen by accident

Studie zur Ermittlung der Nutzerzufriedenheit

Wie zufrieden waren Sie mit der PC-Anmeldung?
Bitte wählen Sie ein Emoji, welches Ihre Anmeldeerfahrung am besten beschreibt



Kommentar (optional)

Teilen Sie uns Probleme bei der Anmeldung über das Kommentarfeld mit

Absenden

[Weitere Informationen zur Studie](#)

Figure 2: After each sign-in, the SurveyApp showed a window in which the participants were asked how satisfied they were with the login procedure. To answer the question, participants had to click one of the five emojis and the submit button (in German: “Absenden”). Optionally, participants could give additional feedback via the text field.

when not actually attempting a sign-in. Or its possible that some time between dismissing the lock screen and sign-in is not dedicated to authentication but might be spent talking to a colleague coming by. Since in our study our main interest surrounds the question of which method is faster and the limitations apply to passwords and Windows Hello authentication equally, we consider this method valid to answer our question. In short: it is important to acknowledge that our time measurements are a valid means of comparing authentication timings *within the scope of our study* but they are *not* an accurate representation of *actual* login times. LogonUI also provides information on the used sign-in method which allowed us to identify whether PIN or a biometric sign-in was used during the Windows Hello phase.

Timing Data Analysis A challenge in the analysis of the timing data we captured via the memory consumption “sidechannel” is the fact that the individual memory profiles differ on each machine and depend on the specific environment. For example, on some machines LogonUI’s memory usage increases when the screensaver is disabled while on others the memory usage decreases. These slight differences make it difficult to automate the evaluation of sign-in timing data. Additionally, specific memory profiles, like the first sign-in after the machine has booted do not follow the typical memory profile pattern. For our timing data analysis we therefore apply manual post-processing instead of an auto-

mated analysis. From the entire data set, we randomly sampled time measurements and ensured that a similar amount of timing data was analyzed for each participant. In total, we selected and analyzed 66 time measurements of the 226 sign-ins from Phase 1 and 244 time measurements of 1,419 sign-ins from Phase 2.

4.4 Ethical Considerations

Our institution does not have an ethics board governing such types of studies. We made sure to follow ethical principles laid out in the Belmont report [28] and discussed the study’s ethical perspective with peers. We collected our participant’s informed consent emphasizing that not participating or withdrawing from the the study later on would not include any negative consequences for them; neither personally nor for their work. We made sure to reiterate this information during every new phase of the study. We acknowledge that the involvement of the employer in this study might pose more pressure on subjects to participate, fearing negative consequences for their regular work. To minimize this effect we repeatedly emphasized how no negative consequences would occur even if participants withdrew their consent. Moreover, we thoroughly ensured that participants were aware that their sign-ins are monitored during and only during the time frame of the study. Participants did not receive any explicit compensation because the study took part entirely during their normal working hours so they did not have any additional effort or workload. Participants’ data was pseudonymized before analyzing and publishing the results. Basic data protection measures such as encrypting the data in transit and access controls were applied to reduce the risk of a breach.

4.5 Limitations

We did an exploratory study of Windows Hello for Business in a small company in Germany and due to the qualitative nature of the study, real-world setting, and the size of the company, we could only recruit 13 participants (which, nevertheless, is representative for this company). As four of the participants already used Windows Hello before we conducted the study, we could not collect the same data of their use of the traditional Windows login as for the other participants. Another participant (P5) used a smart card instead of a password for sign-in, which is a very different authentication method that does not allow a direct comparison. All these factors led to a small and heterogeneous sample. Thus, study results are not generalizable to other authentication settings or companies. Especially, the data on use of biometrics is very limited and does not allow in-depth comparisons or to draw conclusions for other environments. The specific workplace setup also influenced the specific outcomes in preferences and authentication choices and is not representative for different types of companies. Nevertheless, we consider the

setup quite common for office and stationary desk focused workplace settings. The company culture appeared fairly trusting which might have had a positive influence on the participants' openness to changes so results are closely related to trust, especially biometrics usage, should be interpreted carefully and rather considered as an upper bound.

Moreover, as mentioned above, our time measurements and analysis methodology do not allow statements about real-world authentication timings but only comparisons within the scope of our study.

5 Results

We structured this section along our research questions. First, we show the results concerning the usability differences of Windows Hello for Business and the traditional Windows login. Secondly, we present our findings on the perceived security of Windows Hello. Finally, we illustrate why participants used or not used Windows Hello with biometrics. To provide context to the data that we gathered via our SurveyApp, we present a brief summary of the number of sign-ins we used in our analysis.

Frequency of Sign-ins Through the SurveyApp, we measure 226 sign-ins performed by the nine participants of Phase 1 and 1,419 uses of Windows Hello for Business in Phase 2 of the study. On average, each participant performed between 28 and 29 sign-ins via Windows Hello per week ($SD: 10.8$). However, due to different work routines, the number of sign-ins greatly varies among the participants and over the course of the study since some participants went on vacation or got sick. Participant P5 had much fewer sign-ins (30 in total) than all other participants, because they were a trainee and attended school two days per week. During the four-week period, P13 performed the most sign-ins with a total of 203 logins. Figure 7 in Appendix C shows the frequency of sign-ins of the participants in more detail. A more detailed description of the usage of different biometric methods can be found in Section 5.3.

5.1 RQ1: Usability Comparison

To explore our first Research Question **RQ1**, we used the results from the UEQs, the time measurements and ratings from the SurveyApp, and the responses from the interviews. As participants P9, P11, P12, and P13 already used Windows Hello before the study, we could not gather bottom-line data (i.e., login times, satisfaction ratings, and UEQ for the traditional Windows login) for these participants. We also excluded results of participant P5 from the UEQ comparison because they used a smartcard instead of a password. However, we explicitly asked all participants in the interviews to compare Windows Hello for Business with the authentication method they used before.

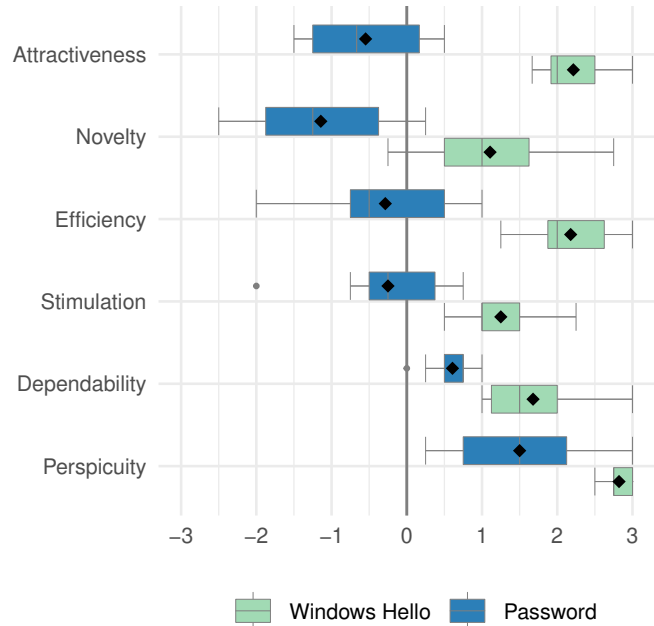


Figure 3: UEQ results for each of the six categories after the first phase (Password) and the second phase (Windows Hello). All boxes for Windows Hello are on the positive side of the scale, indicating an excellent usability experience. In contrast, ratings for password are more skewed to the negative side of the scale.

Usability Experience Questionnaires The evaluation of the UEQs showed that Windows Hello for Business scored better than the password-based Windows login across all six UEQ scales. As mentioned before, we only used the responses from the participants that originally used passwords (P1, P2, P3, P4, P6, P7, P8, and P9) for the comparison shown in Figure 3.

Windows Hello was rated particularly well on the *Perspicuity*, *Attractiveness*, and *Efficiency* scales with average scores higher than two. These high ratings indicate that Windows Hello is even easier to understand than passwords (which also receive fairly high ratings on the Perspicuity scale) while being much more attractive and efficient to use than passwords. Overall, the UEQ ratings for Windows Hello were all *above average*, most of them even *excellent*, compared to the UEQ benchmark data set [36]. In contrast, the ratings for the password-based login are considered as *bad*, except from Perspicuity which was rated *above average*. Comparing the UEQ results of password-based authentication and Windows Hello via t-tests (as described in the UEQ handbook [35]) reveals significant differences for *Attractiveness*, *Efficiency*, *Stimulation*, and *Novelty* scales ($p < 0.01$). The test results for *Perspicuity* and *Dependability* ($p < 0.02$) are almost significant. Table 1 gives an overview of the UEQ results of the two sign-in methods and puts them in relation to the benchmark data set.

Table 1: Comparison of UEQ results for password-based Windows login and Windows Hello for Business with the UEQ benchmark dataset [36] and paired t-test for each scale [35]. The results were corrected via Bonferroni–Holm method.

Scale	Password		Windows Hello		t-test	
	Mean	Benchmark	Mean	Benchmark	t	Pr(> z)
Attractiveness	-0.56	Bad	2.21	Excellent	-8.86	<0.001 ***
Novelty	-1.25	Bad	1.06	Good	-4.76	0.001 **
Efficiency	0.00	Bad	2.16	Excellent	-4.41	0.003 *
Stimulation	-0.41	Bad	1.19	Above Avrg.	-3.98	0.003 *
Dependability	0.75	Bad	1.66	Excellent	-2.89	0.015 .
Perspicuity	1.69	Above Avrg.	2.84	Excellent	-2.97	0.019 .

Signif. codes: *** $\hat{=}$ < 0.001 ; ** $\hat{=}$ < 0.01 ; * $\hat{=}$ < 0.05 ; . $\hat{=}$ < 0.1

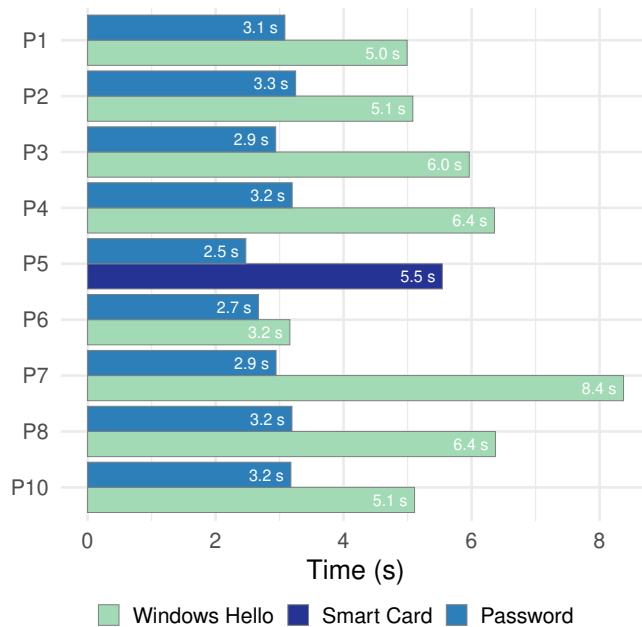


Figure 4: Average sign-in duration per participant and sign-in method. Windows Hello for Business was faster for each participant.

Authentication Speed Comparing the sign-in timings, Windows Hello for Business was faster for all participants in the study. The sign-in times for password ranged from 3.2s (P6) to 8.4s (P7), that is 5.7s on average. For Windows Hello, the sign-in took 2.5s (P5) to 3.3s (P2 and P10), on average 3.0s. These results indicate that authentication with Windows Hello was on average 47% faster than password-based authentication. Figure 4 shows the differences in the duration of the sign-in process between password authentication and Windows Hello for each participant.

When we asked the participants in the interviews which aspect of Windows Hello for Business they liked the most, all of them mentioned the faster authentication speed.

Definitely, happier than with the password, because it's simply faster. (P8)

This finding suggests that the usability of Windows Hello greatly benefits from the gain in authentication speed. Some participants attributed the speed gain to the shorter PINs.

Faster, because the PIN is obviously shorter than my old password. (P11)

Previously, participants used passwords of a length of at least 10 characters. The new PINs consist of 6 digits for all participants. Many participants reported that they used long and complex passwords which were hard to remember while the PINs are easier to remember.

With the password, well I've used, I think, a 16-character password consisting of upper/lower case letters, numbers, and special characters. Once you've learned it by heart, than it's fine but if you have to learn a new one then it takes some time to memorize it. A six-digit PIN is easier to learn. (P7)

Although, we did not measure authentication errors, the error-rate is another factor influencing the authentication speed and participants mentioned this aspect during the interviews.

I mistype less, as with upper/lower case letters and such, because it was just numbers. (P4)

Windows Hello also requires less interaction during the authentication procedure than the traditional login. Instead of waiting for confirmation after entering a PIN by clicking a button on the user interface or pressing Enter, Windows Hello tries to perform a sign-in after the correct number of key strokes. Participants noticed this subtle difference between PIN entry and the traditional password entry and saw it as a benefit.

I don't have to press enter, so it's a bit faster. (P7)

Another improvement of Windows Hello is that it ignores the state of Num-Lock and always allows to enter numbers on the number keypad.

I don't have to check the keyboard, I can type right away, even if that number light isn't on, it still worked, I liked that. (P4)

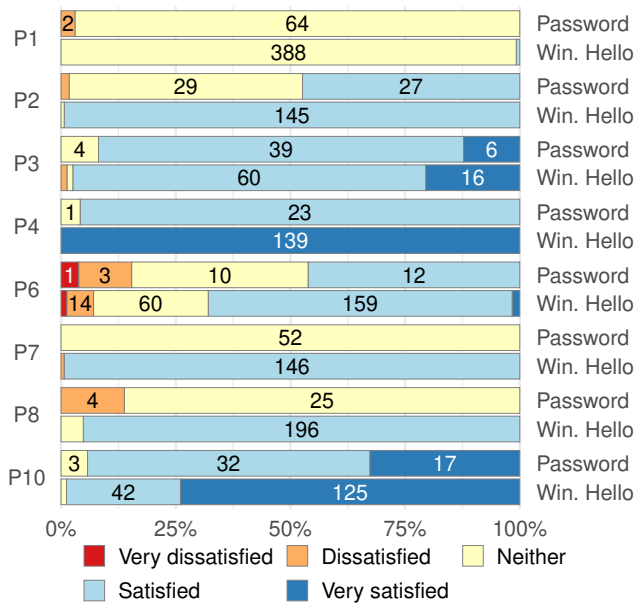


Figure 5: Satisfaction ratings submitted via the SurveyApp after each sign-in grouped by authentication method and participant. Only participants using passwords in Phase 1 are included in the bar chart. The numbers on the bars are the absolute number of ratings per level of satisfaction.

This improvement also helps to avoid errors and thus increases the authentication speed. One participant stated that fingerprint recognition annoys her and that she prefers to sign-in with her PIN because she feels it is faster and she already has her hands on the keyboard.

No, fingerprint annoys me. If I don't put my finger correctly on the reader, I can sign-in faster with the PIN, also because I can type quickly. In short, fingerprint makes me uncomfortable and I already have my fingers on the keyboard anyway. (P9)

Satisfaction Beside the time data that we gathered, we asked the participants after each successful sign-in how satisfied they were with it via the SurveyApp (see Section 4.3 and Figure 2). Figure 5 shows the satisfaction rating for the eight password-using participants.

Overall, most of the ratings were positive or neutral and participants ratings were mostly consistent over time (Participant 6 being an exception). However, the ratings for Windows Hello tend to be higher than for password-based authentication. which is also confirmed by a Wilcoxon signed-rank test ($p < 0.01, V = 0$). We used a Wilcoxon signed-rank test since we could not assume equidistance for our satisfaction items. The higher satisfaction with Windows Hello is in line with our findings from the UEQs and the feedback from the interviews.

5.2 RQ2: Perceived Security

In general, the participants were not aware of any of the security features that Windows Hello for Business offers. However, the perceived security of the different sign-in options provided by Windows Hello varied greatly among the participants.

Facial Recognition When asked which of the three sign-in options (i.e., facial recognition, fingerprint, PIN), eight participants found facial recognition to be the most secure. Participants trusted facial recognition more than fingerprint recognition, for example, they argued that the facial recognition software were more sophisticated, that it would involve several features of the face, and were resistant to simple forgery attacks.

I'm not an expert in this area but I imagine that facial recognition software can store several features, eye distance, face shape, etc., and that this may be even more secure than a PIN. (P2)

Perhaps facial recognition is even more secure, we tested it once, with a photo and via FaceTime, neither was accepted by the system. (P10)

Fingerprint Recognition One participant stated that fingerprint recognition is the most secure sign-in method but could not explain why.

I think fingerprint is probably even more secure than facial recognition. But it's just a feeling. (P6)

PIN Another participant said PIN is the most secure sign-in method, explaining that they did not trust the technology and that one can make up their own PIN which would be hard to guess.

I don't know. Sometimes I don't trust things, you come up with a PIN yourself, it's hard to steal but I don't know if there isn't a vulnerability, especially with the camera and the facial recognition, I'm not so sure. (P5)

Biometrics in General Participant P9 considered both biometric sign-in options more secure than PIN, but saw no difference between facial recognition and fingerprint in terms of security.

Fingerprints are not used without reason as a unique identifier in identity documents, so in this respect I think that such a fingerprint is really secure. But basically, facial recognition and fingerprints are equally secure. (P9)

Windows Hello for Business in General Participant P11 stated that all options of Windows Hello were equally secure. They explained in the interview that facial recognition and fingerprints are more secure than a password. When asked whether the PIN was more secure, less secure, or equivalent to the biometric methods, the participant said the PIN was equivalent to the biometric methods.

Passwords One participant considered facial recognition to be less secure than PINs and would prefer to not use it for more critical services like online banking.

I don't think I'd do my online banking with facial recognition now, because I just don't know enough about it, and if someone can fool it with photos or whatnot. (P7)

Instead, the participant considers a long, complex password with upper/lower case letters, numbers, and special characters to be the most secure.

A 16 character password with special characters, upper/lower case letters, and numbers seems more secure to me. (P7)

The participant has previously read up on secure passwords and learned that complex passwords are harder to guess than less complex ones.

So, I was on a password test site where I can enter a password, and the site tells me how long it would take a computer to hack it. When I use a longer, more complex password and more different character types, then it displays that it takes a million years to hack it. (P7)

However, guessing attacks are only one problem and phishing or password reuse are still problematic even if a complex password is used. This statement underlines, why security literacy is an important factor when deploying a new security feature since it helps people to better understand the change and may increase acceptance.

5.3 RQ3: Use of Biometrics

As described in Section 4.2, only eight participants had devices with biometric hardware compatible with Windows Hello (cf. Section 4.5). We encouraged these eight participants to try and use the different sign-in method available on their devices when Windows Hello was enabled in the beginning of Phase 2. Besides a PIN, which every participant could use, Participants P6, P7, P8, and P9 could use facial and fingerprint recognition. Participants P10, P11, and P12 had only fingerprint readers on their devices and P13 had a face recognition camera available.

However, our results show that only two participants frequently used the biometric sign-in option. Participant P13 used facial recognition most of the time (174 out of 203 logins; 86%), and P11, who used facial recognition sometimes

(21 out of 112; 19%). All other participants (almost) always used PINs for authentication via Windows Hello. Using PIN instead of biometric authentication is counter-intuitive as people usually find biometric authentication more secure than passwords or PINs [38].

5.3.1 Reasons Not to Use Biometrics

Influence of Workplace Setup While our participants did use laptops as work stations, these were mostly used with external screens and docking stations. The laptops were often placed on the far side of the desk with closed lid and hard to reach. This prevented easy access to the fingerprint scanner which was mostly located next to the laptop keyboard. Similarly, the built-in cameras were located at the top or bottom of the laptop display. Opening the laptop lid for each sign-in cancels out the convenience that biometrics may have because it requires additional steps and thus protracts the sign-in. We also found that for facial recognition, the angle and distance at which the laptop was placed to the user highly influenced the accuracy and success rate of the logins.

Maybe it's because the computer is too far away. About a meter on the right side and I have to face at my laptop, so to speak, and then sometimes it does not recognize me. (P12)

Participant P13, who used facial recognition most frequently, had a different setup than P12. They used a convertible device where the docking station is connected via a cable instead of a fixed mount. In order to use the device it also cannot be closed but must sit in an upright position. These factors address the exact issues mentioned by Participant P12; a closed laptop lid, a laptop positioned too far away, and at the wrong angle. Participant P13 had the flexibility to place the device closer and change the angle in which the device is directed towards them which could explain their more extensive facial recognition usage.

Fear of Being Observed Four participants (P9, P10, P11, and P12) reported a feeling of being watched by the camera and preferred, for this reason, the PIN sign-in option.

With facial recognition, you're always a bit more skeptical, the camera is always on, and you've heard stories that they can be hacked. (P11)

Participants who already used facial recognition sporadically were asked if they would use facial recognition more often if they had a camera not only on their laptop, but also on their main screen, which would be better oriented toward them and could eliminate the workplace setup issues discussed above. All three of these participants (P10, P11, and P12) expressed their discomfort with the idea of having a camera mounted on top of the main display.

But then I might feel like I'm being watched. (P10)

On the contrary, other participants suggested that mounting an external camera on top of the primary monitors could be a could solution to workplace setup issues.

Fear of Being Locked Out Three participants (P2, P4, P11) saw risks regarding availability if there were technical problems with facial or fingerprint recognition. While those participants did not consider their fear as a reason not to use biometric authentication, they did emphasize the importance of having the PIN at least as a fallback option. Specifically, they feared that the camera could break or fail to recognize them if features in their face changed.

Face recognition always requires a working camera. Does it work when I shave off my beard? I don't know if it still works then, no clue. (P11)

5.3.2 Reasons to Use Biometrics

Most participants did express to not have a general aversion against biometrics. Participant P3, for example, stated that they also use fingerprint recognition on their personal laptop. All five participants that used a PC without biometric capabilities mentioned that they would use biometrics if it were available on their PC, as well. Participant P13, who used facial recognition regularly, underlines this positive attitude towards biometrics, describing it as easier to use.

And with facial recognition, it's just a much easier recognition. (P13)

A situational preference for biometrics was described by participant P12. Overall, they preferred the PIN for the aforementioned reasons but used facial recognition in situations where other people are present in the same room or behind them when signing in. The participant deliberately used facial recognition to avoid other people being able to see their PIN when they entered it.

I have both options, if I wish that no one can see my PIN, then I just go to the camera, then I do not have to enter anything. I can decide freely. (P12)

Arguments for fingerprint and facial recognition, that are also commonly described in biometric authentication literature [3, 11] were also mentioned by a number of participants, for example, that facial recognition does not require anything to be remembered, since it does not require entering a secret. While this a valid statement, it should be noted that Windows Hello for Business always falls back to the six-digit PIN when facial or fingerprint recognition does not work (e.g., when the camera is disconnected or the lighting conditions have changed). Therefore, there is a risk that users will sign-in exclusively with biometrics over an extended period of time and then tend to forget their PIN over time, which would potentially lock them out of their device when the biometric recognition fails.

6 Discussion

Studying usability in the field instead of a lab or online setting helps to better understand what works well of, in our case, Windows Hello for Business and what obstacles people encounter when using it on a daily basis. In our study, Windows Hello for Business outperforms the traditional login in most usability aspects. We found that the Windows Hello login is on average 47% faster than the login with passwords. This is also reflected in the employees' satisfaction with the new authentication mechanism.

Knowledge-based Authentication vs. Biometry Usually, passwords are not very popular among users and suffer from various well-studied usability and security issues [9, 20, 31]. In recent years, a multitude of efforts have tried to counteract and overcome the password issues, trying to populate the use of password managers [15, 32], graphical authentication [4, 37], and Multi-factor authentication [19, 33, 34] are just a few to mention in this list. Biometric authentication has been one of the few approaches that have proven to be a viable, accepted, but still secure alternative to knowledge-based authentication – at least in certain authentication contexts [18]. Our results offer interesting insights into relevant factors for the real-world adoption of biometric authentication in the context of a corporate environment. Contrary to most findings in the literature, where users prefer biometric authentication over knowledge-based variants [3, 11], in our study, most participants resorted to using knowledge-based authentication (PINs) instead of biometrics. While participants in our company worked mostly stationary, this is different in other work settings especially for those traveling to customers and working on the go. One of our participants mentioned resorting to face recognition whenever someone was present nearby to prevent shoulder surfing. Compared to passwords, where shoulder surfing in public places is inevitable, the biometric option in Windows Hello allows for shoulder surfing prevention while on the go.

Recommendation: *Companies considering a switch to Windows Hello should take their employees work modalities into account. For example, by only providing biometric hardware for those in need of shoulder surfing protection.*

Next, we discuss factors that played a role in participants' decision to use PINs instead of available alternatives.

Both *hardware availability* and *hardware placement* played an important role in participants' decisions against biometric authentication. As described in Section 5.3, the participants mostly worked with external displays and docking stations for laptops. This is typical for office settings, where laptops are placed outside of direct reach, and lids are kept close. Consequently, built-in fingerprint readers are not (easily) accessible, and the built-in cameras are either not facing the user, or facing users from angles they typically are

not looking at. Furthermore, lock screens are typically displayed on the main display, which often is not the one with the camera, making face unlock awkward to use. This is in sharp contrast to authentication on mobile devices, where biometrics have a significant share (e.g., 80% consumer devices in the USA had biometrics enabled in 2020 [13]). Here, the sensors for fingerprint and face recognition are placed to be reachable easily when using the device.

Recommendation: *Biometric hardware can offer great usability benefits but only in the correct usage setting. If it is not essential for a company to offer biometric logins it can be sufficient to offer Windows Hello with PINs since high user satisfaction can be expected.*

Privacy Issues While in recent years, most of the laptops and convertibles come with built-in facial or fingerprint recognition capabilities, in office environments stationary computers are still broadly used which do not have these functionalities built-in (five participants had such a device). The aforementioned workplace setup can render some of the built-in authentication hardware useless. Consequently, additional hardware for biometric authentication needs to be purchased and set up which is an investment that some employees potentially disprove, either because they do not want to use their fingerprints for authentication at work or feel under surveillance when facial recognition cameras are mounted to their displays. Such *privacy-related concerns* with using biometrics are well-known in the literature [8] but have more severe implications in the corporate context than for private usage. The company we conducted our study at was a small, family-owned business with a trusting work climate so our participants did not strongly express any explicit privacy concerns with regard to their company. However, this might be different in larger corporations or in less positive work climates, especially in cases where employee surveillance has already been an issue [2, 5, 14]. In those cases, employees might feel like their privacy is invaded when being encouraged or even enforced to use biometrics. Facial recognition can even evoke a feeling of being monitored. Part of the solution could be the use of cameras with built-in shutters. However, since even PINs were highly accepted, more usable, and much faster than the traditional Windows sign-in, it might be sufficient to rely on non-biometric Windows Hello.

Recommendation: *When choosing (biometric) hardware for Windows Hello take your company culture and trust environment into account to obviate employees feeling uneasy or even monitored. For some companies it might be sufficient to rely on PINs and not introduce biometric hardware after all.*

Deployment of Windows Hello Even though Windows Hello offers some strong usability benefits (like *Quasi-Memorywise-Effortless*, *Quasi-Physically-Effortless*, *Easy-*

to-learn, *Efficient-to-Use*; cf. Bonneau et al. [6]) it introduces extra effort in the deployment phase (*Negligible-Cost-per-User* not fulfilled). As the login credentials of Windows Hello are tied to a specific device, i.e., every device, a person wants to use, has to be enrolled with the Windows account of that person (cf. Section 2). In organisations in which people share their devices, the default setup of Windows Hello for Business, i.e., using the built-in TPM, is not feasible.

Smartcards or other hardware tokens (e.g., FIDO2 tokens like YubiKeys⁵) have the advantage that they support roaming. The same goes for passwords which are stored with the user account on a server in the enterprise's network and not on the individual devices.

However, the authentication secret being bound to and never leaving the device is an intentional security and privacy-preserving feature of Windows Hello. This might not be an issue or even beneficial in work environments where every user has their own device and only uses that but in many work settings with several shared computers, Windows Hello will not provide the necessary flexibility a password or roaming token does.

Recommendation: *Windows Hello using the built-in TPM is less suited for shared devices, especially with many users. Relying on roaming authenticators or the traditional password is a better choice in these cases.*

7 Conclusion

We studied Windows Hello for Business, Microsoft's latest alternative to traditional password authentication. In a small business, we measured authentication times of 13 employees, collected their experience, and conducted interviews to understand their perceptions of and attitudes towards Windows Hello in the wild. Our five weeks long study revealed that, in general, participants like Windows Hello, finding it more usable than the traditional Windows sign-in methods. Windows Hello was measurably faster, perceived as more responsive, and convenient to use. Contrary to findings on biometrics usage in mobile devices, participants in our study tended to use PINs most of the time. This was partially due to a lack of availability of biometric hardware, the form factor of their device, and the setup of their workplace (e.g., biometric sensor not reachable).

Acknowledgment

This research was partially funded by the MKW-NRW research training group SecHuman and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2092 CASA – 390781972.

⁵<https://www.yubico.com/products/yubikey-5-overview/>, as of June 9, 2022

References

- [1] Sarah Alismail and Hengwei Zhang. The Use of Emoji in Electronic User Experience Questionnaire: An Exploratory Case Study. In *Hawaii International Conference on System Sciences*, pages 3366–3375. ScholarSpace, 2018.
- [2] Raluca Balica. Automated Data Analysis in Organizations: Sensory Algorithmic Devices, Intrusive Workplace Monitoring, and Employee Surveillance. *Psychosociological Issues in Human Resource Management*, 7(2):61–66, 2019.
- [3] Chandrasekhar Bhagavatula, Blase Ur, Kevin Iacovino, Su Mon Kywey, Lorrie Faith Cranor, and Marios Savvides. Biometric Authentication on iPhone and Android: Usability, Perceptions, and Influences on Adoption. In *Workshop on Usable Security*. ISOC, 2015.
- [4] Robert Biddle, Sonia Chiasson, and Paul C. Van Oorschot. Graphical Passwords: Learning from the First Twelve Years. *ACM Computing Surveys*, 44(4):19:1–19:41, 2012.
- [5] Stephen Blumenfeld, Gordon Anderson, and Val Hooper. COVID-19 and Employee Surveillance. *New Zealand Journal of Employment Relations*, 45(2):42–56, 2020.
- [6] Joseph Bonneau, Cormac Herley, Paul C. Van Oorschot, and Frank Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *IEEE Symposium on Security and Privacy*, pages 553–567. IEEE, 2012.
- [7] Robert L. Brennan and Dale J. Prediger. Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, 41(3):687–699, 1981.
- [8] Ivan Cherapau, Ildar Muslukhov, Nalin Asanka, and Konstantin Beznosov. On the Impact of Touch ID on iPhone Passcodes. In *Symposium on Usable Privacy and Security*, pages 257–276. USENIX, 2015.
- [9] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The Tangled Web of Password Reuse. In *Symposium on Network and Distributed System Security*. ISOC, 2014.
- [10] Sanchari Das, Andrew Dingman, and L. Jean Camp. Why Johnny Doesn’t Use Two Factor: A Two-Phase Usability Study of the FIDO U2F Security Key. In *Financial Cryptography and Data Security*, pages 160–179. Springer, 2018.
- [11] Alexander De Luca, Alina Hang, Emanuel von Zezschwitz, and Heinrich Hussmann. I Feel Like I’m Taking Selfies All Day!: Towards Understanding Biometric Authentication on Smartphones. In *ACM Conference on Human Factors in Computing Systems*, pages 1411–1414. ACM, 2015.
- [12] Jessica T. DeCuir-Gunby, Patricia L. Marshall, and Allison W. McCulloch. Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project. *Field Methods*, 23(2):136–155, 2011.
- [13] Duo Security. The 2020 Duo Trusted Access Report: A Remote Access Playbook, March 2021. <https://duo.com/resources/ebooks/the-2020-duo-trusted-access-report>, as of June 9, 2022.
- [14] Lilian Edwards, Laura Martin, and Tristan Henderson. Employee Surveillance: The Road to Surveillance is Paved with Good Intentions. In *Amsterdam Privacy Conference*, pages 1–30, 2018.
- [15] Michael Fagan, Yusuf Albayram, Mohammad Maifi Hasan Khan, and Ross Buck. An Investigation Into Users’ Considerations Towards Using Password Managers. *Human-Centric Computing and Information Sciences*, 7(1), 2017.
- [16] Florian M. Farke, Lennart Lorenz, Theodor Schnitzler, Philipp Markert, and Markus Dürmuth. “You still use the password after all” – Exploring FIDO2 Security Keys in a Small Company. In *Symposium on Usable Privacy and Security*, pages 19–35. USENIX, 2020.
- [17] Haichang Gao, Wei Jia, Ning Liu, and Kaisheng Li. The Hot-Spots Problem in Windows 8 Graphical Password Scheme. In *Symposium on Cyberspace Safety and Security*, pages 349–362. Springer, 2013.
- [18] Rachel L. German and K. Suzanne Barber. Consumer Attitudes About Biometric Authentication. Technical Report UT-CID-18-03, The University of Texas at Austin, May 2018.
- [19] Maximilian Golla, Grant Ho, Marika Lohmus, Monica Pulluri, and Elissa M. Redmiles. Driving 2FA Adoption at Scale: Optimizing Two-Factor Authentication Notification Design Patterns. In *USENIX Security Symposium*, pages 109–126. USENIX, 2021.
- [20] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. “What was that site doing with my Facebook password?” Designing Password-Reuse Notifications. In *ACM Conference on Computer and Communications Security*, pages 1549–1566. ACM, 2018.

- [21] Christina Katsini, Christos Fidas, George E. Raptis, Marios Belk, George Samaras, and Nikolaos Avouris. Influences of Human Cognition and Visual Behavior on Password Strength during Picture Password Composition. In *ACM Conference on Human Factors in Computing Systems*, pages 87:1–87:14. ACM, 2018.
- [22] Ejin Kim and Hyoung-Kee Choi. Security Analysis and Bypass User Authentication Bound to Device of Windows Hello in the Wild. *Security and Communication Networks*, 2021(1):6245306:1–6245306:13, 2021.
- [23] Kat Krol, Eleni Philippou, Emiliano De Cristofaro, and M. Angela Sasse. “They brought in the horrible key ring thing!” Analysing the Usability of Two-Factor Authentication in UK Online Banking. In *Workshop on Usable Security*. ISOC, 2015.
- [24] Leona Lassak, Annika Hildebrandt, Maximilian Golla, and Blase Ur. “It’s Stored, Hopefully, on an Encrypted Server”: Mitigating Users’ Misconceptions About FIDO2 Biometric WebAuthn. In *USENIX Security Symposium*, pages 91–108. USENIX, 2021.
- [25] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*, pages 63–76. Springer, 2008.
- [26] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *IEEE Symposium on Security and Privacy*, pages 268–285. IEEE, 2020.
- [27] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. “Blissfully Happy” or “Ready to Fight”: Varying Interpretations of Emoji. In *International Conference on Web and Social Media*, pages 259–268. AAAI, 2016.
- [28] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, September 1978. <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html>, as of June 9, 2022.
- [29] Wataru Oogami, Hidehito Gomi, Shuji Yamaguchi, Shota Yamanaka, and Tatsuru Higurashi. Poster: Observation Study on Usability Challenges for Fingerprint Authentication Using WebAuthn-enabled Android Smartphones. In *Symposium on Usable Privacy and Security*. USENIX, 2020.
- [30] Kentrell Owens, Blase Ur, and Olabode Anise. A Framework for Evaluating the Usability and Security of Smartphones as FIDO2 Roaming Authenticators. In *Who Are You?! Adventures in Authentication Workshop*, pages 1–5, 2020.
- [31] Bijeeta Pal, Tal Daniel, Rahul Chatterjee, and Thomas Ristenpart. Beyond Credential Stuffing: Password Similarity Models using Neural Networks. In *IEEE Symposium on Security and Privacy*, pages 866–883. IEEE, 2019.
- [32] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why People (Don’t) Use Password Managers Effectively. In *Symposium on Usable Privacy and Security*, pages 319–338. USENIX, 2019.
- [33] Ken Reese, Trevor Smith, Jonathan Dutton, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A Usability Study of Five Two-Factor Authentication Methods. In *Symposium on Usable Privacy and Security*, pages 357–370. USENIX, 2019.
- [34] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent E. Seamons. A Tale of Two Studies: The Best and Worst of YubiKey Usability. In *IEEE Symposium on Security and Privacy*, pages 872–888. IEEE, 2018.
- [35] Martin Schrepp. User Experience Questionnaire Handbook (Version 8), December 2019. <https://www.ueq-online.org/Material/Handbook.pdf>, as of June 9, 2022.
- [36] Martin Schrepp, Jörg Thomaschewski, and Andreas Hinderks. Construction of a Benchmark for the User Experience Questionnaire (UEQ). *Journal of Interactive Multimedia and Artificial Intelligence*, 4(4):40–44, 2017.
- [37] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *ACM Conference on Computer and Communications Security*, pages 161–172. ACM, 2013.
- [38] Verena Zimmermann and Nina Gerber. The Password Is Dead, Long Live the Password – A Laboratory Study on User Perceptions of Authentication Schemes. *International Journal of Human-Computer Studies*, 133(1):26–44, 2020.

Appendix

A Interview Questions

General Perception of Windows Hello

- You've been using Windows Hello for the past four weeks, how happy are you with it?
- Can you explain the differences between *password/smartcard* login and Windows Hello?
- Did you encounter any problems logging on to the PC over the past four weeks?
 - How have you solved them?

Use of Biometrics

Skip these questions if participant doesn't have biometric hardware available.

- Your device has a *facial recognition camera/finger print reader*, have you used this/these feature(s)?
 - How often have you used *facial recognition/finger print recognition* compared to the *PIN*?
- What stopped you from using *facial recognition/finger print recognition*?
- Which login method have you used most and why?

Perceived Authentication Speed

- How much time do you spend per login with Windows Hello compared to *password/smartcard*?

Perceived Security

- Please rate the security of Windows Hello.
 - Do you think there is a difference between *facial recognition, finger print recognition, and PIN* in terms of security?
 - Are there any security issues you see with the use of Windows Hello?

Satisfaction

- Windows Hello is available on many laptops and desktop computer, would you use it on your personal Windows computer?
- Is there anything you like in particular about Windows Hello?
- Is there anything you dislike about Windows Hello?
- Would you rather continue using Windows Hello or return to the traditional *password/smartcard*?

B Codebook

Table 2: The codebook used to code the interviews.

Code	IDs	Description	Example
<i>General Perception of Windows Hello</i>			
Hello is fast	P1, P2, P3, P4, P5, P6, P7, P8, P10, P13	Windows Hello is seen as a fast way of authentication.	<i>“The speed, I boot up and can start working right away, wouldn’t know what could make it better now.” (P13)</i>
Uses / Considers to use Hello personal devices	P1, P2, P3, P4, P7, P8, P9, P10, P13	Participant considers to use, or already uses, Windows Hello with his personal devices at home.	<i>“On my private laptop I’m using the fingerprint, that is also convenient.” (P3)</i>
Continue using Hello	<all>	Participant wants to continue using Windows Hello after the study.	<i>“No, no, the number combination was more appealing to me, I wouldn’t want to go back to the password now.” (P8)</i>
Hello is convenient	P2, P6, P7, P11, P12, P13	Windows Hello is seen as a convenient way of authentication (compared to the password).	<i>“It was definitely more convenient than before, I had a longer password before.” (P7)</i>
Hello is easy	P1, P2, P3, P5, P10, P13	Windows Hello is easy to learn and easy to use (compared to the password).	<i>“It was easier with the PIN. So, I found the handling better than before.” (P3)</i>
<i>Preferred Sign-in Method</i>			
Prefers Face	P13	Participant prefers facial recognition over all other methods.	<i>“And with facial recognition, it’s just a much easier recognition, you know, you are signed-in through then just the facial recognition.” (P13)</i>
Prefers PIN	P7, P8, P9, P10, P11, P12	Participant prefers PIN over all other methods.	<i>“The number combination. Yes, because as I said, it was a tad faster.” (P12)</i>
<i>Reservation Against Biometrics</i>			
Availability risks	P2, P4, P11	Participant describes risks regarding the availability of the authentication method (e.g. camera is not working, fingerprint is not detected).	<i>“The fingerprint recognition, I could imagine, if you have wet hands, so that’s not the case with us now, but if your hands are wet or have a possible injury, then you would have maybe problems with the sign-in.” (P2)</i>
Accustomed to PIN	P8, P8, P9, P10, P11	Participant is used to use the PIN and prefers it over other methods.	<i>“Yeah, I’m kind of used to typing.” (P10)</i>
Laptop lid closed	P6, P7, P10	The laptop is docked and the lid is usually closed which covers the biometric sensors of the device and makes them unavailable until the lid is opened manually.	<i>“I tried it once, but yes, by the fact that I actually always have the laptop closed, that would have been a circumstance for me to use it that way.” (P7)</i>
Laptop too far away	P11, P12	The laptop is docked and on the other side of the desk which brings the biometric sensors out of reach.	<i>“What is a bit stupid is that the notebook is not frontal to the monitor, if I now had a camera directly on the monitor it would certainly be a bit better.” (P11)</i>
Too secure / unnecessary	P6, P8, P11	Windows Hello is seen as too secure and unnecessary for this type of scenario it is used in.	<i>Personally, I don’t see this as a necessity. (P11)</i>

Table 2: Continued from previous page

Code	IDs	Description	Example
Fear of being observed	P9, P10, P11, P12	Participant feels observed through the web cam pointing at them.	<i>“Then I have the feeling that I am am being watched.”</i> (P12)
<i>Perceived Authentication Speed</i>			
Hello is faster	<all>	Authentication can occur faster using Windows Hello compared to the password.	<i>“Yeah, between half and 3/4 as long as the password, around the twist, I’d say.”</i> (P6)
<i>Perceived Usability</i>			
Hello nothing to carry	P5, P12, P13	Participants do not have to carry an additional token/device.	<i>“You don’t always have to carry the stick with you and you can’t forget it.”</i> (P12)
Hello fewer errors	P2, P3, P4, P6, P8, P9	Participant faces fewer authentication failures when signing-in (e.g. due to mistyped password).	<i>“I mistype less, with upper and lower case letters or something, because it was just numbers.”</i> (P4)
Hello memory-wise less effort	P4, P7, P8, P11, P13	Participant needs to remember less information to sign-in, a six-digit PIN is always shorter than a password with minimal length of 10 characters.	<i>“And I can remember it better, because sometimes, especially after a vacation, you kind of forget the password.”</i> (P4)
<i>Perceived Security</i>			
Biometrics most secure (no difference between face/fingerprint)	P9	Participant considers biometrics as most secure authentication method, fingerprint and facial recognition are seen as equally secure sign-in options.	<i>“But basically, the two sign-in options seem to be equally secure to me.”</i> (P9)
No difference between Hello methods	P11	Participant considers all sign-in options of Windows Hello as equally secure.	<i>“Yes, I would put them in the same category.”</i> (P11)
Facial recognition most secure	P1, P2, P3, P4, P8, P10, P12, P13	Participant considers facial recognition as most secure sign-in option.	<i>“If you now take facial recognition again and also the fingerprint, I also have the impression that facial recognition is easier, better and more secure.”</i> (P13)
Password most secure	P4, P7, P11	Participant considers traditional passwords as most secure sign-in option.	<i>“So I suppose it’s also secure, but yes, in theory it seems more insecure than a long password.”</i> (P7)
Fingerprint most secure	P6	Participant considers fingerprint as most secure sign-in option.	<i>“Yeah, I think fingerprint is probably even more secure than facial recognition, so purely emotionally, but yeah.”</i> (P6)
PIN most secure	P5	Participant considers PIN as most secure sign-in option.	<i>“No, I think the PIN is even more secure.”</i> (P5)

C Additional Plots

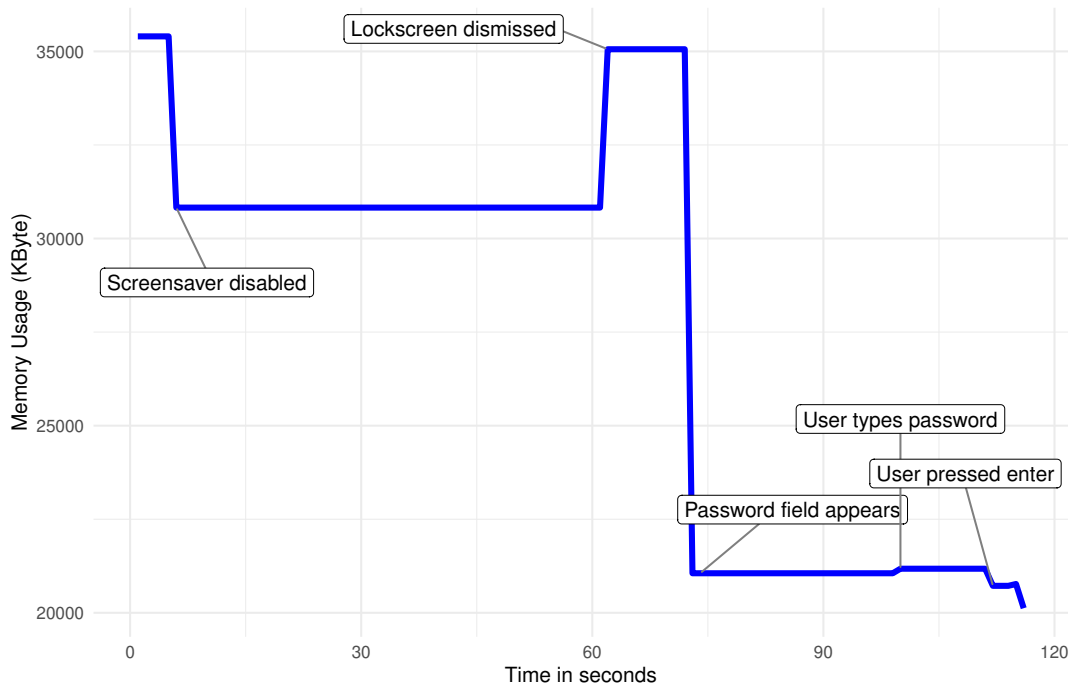


Figure 6: Example of a memory profile of Microsoft’s LogonUI used to determine the sign-in timings.

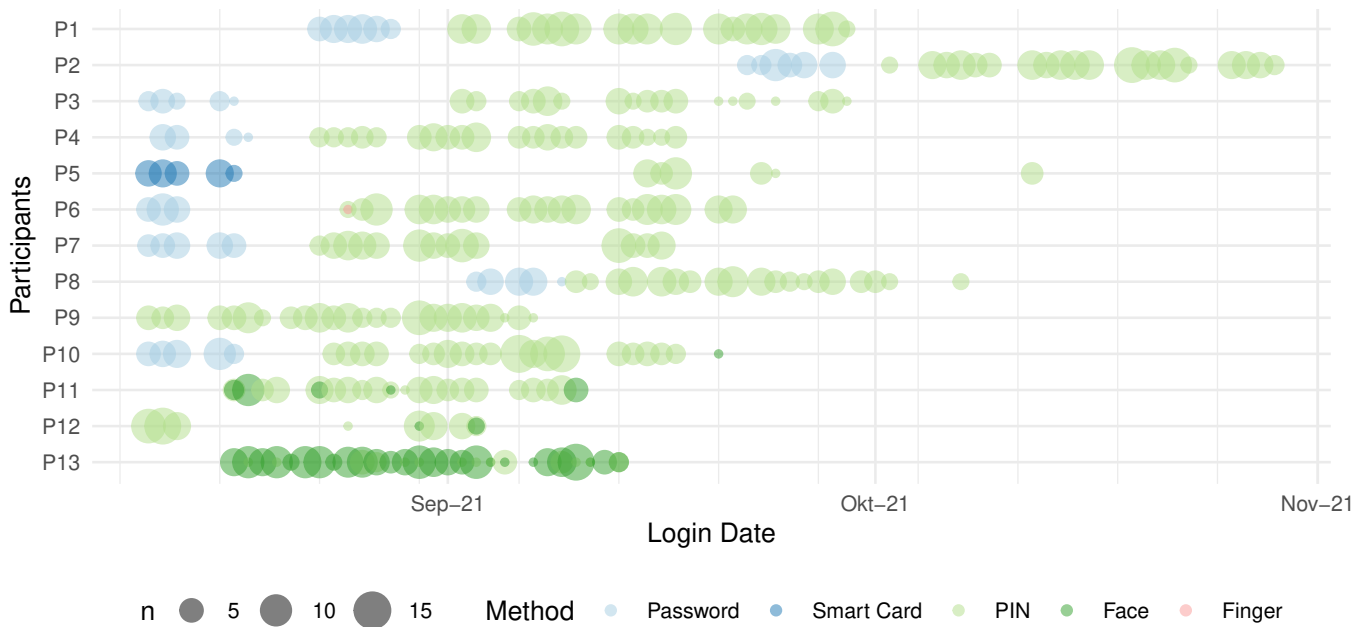


Figure 7: Sign-ins of the participants over the course of the study.

Improving Password Generation Through the Design of a Password Composition Policy Description Language

Anuj Gautam
The University of Tennessee

Shan Lalani
The University of Tennessee

Scott Ruoti
The University of Tennessee

Abstract

Password managers help users more effectively manage their passwords, yet the adoption of password generation is minimal. One explanation for this problem is that websites' password composition policies (PCPs) can reject generated passwords, creating a usability impediment. To address this issue, we design a PCP language that websites use to describe their PCP and that managers use to generate compliant passwords. We develop this language using an iterative process involving an extensive collection of PCPs scraped from the Web. We provide libraries for adopting our PCP language into websites and password managers and build proof-of-concept prototypes to verify the real-world feasibility of our PCP language. Using a 25-person user study, we demonstrate that our language and libraries are easy to pick up and correctly use for novice developers. Finally, we replicate and extend past research evaluating Web PCPs, showing that half of PCPs fail to require passwords that resist offline attacks when considering that users prefer certain character classes when selecting their passwords.

1 Introduction

Despite their problems [7–9, 27, 30, 34, 37], passwords remains the dominant form of authentication [5]. Password managers strengthen password-based authentication by helping users generate, store, and enter passwords, making it easier to adopt strong, unique passwords [19, 27]. Still, research has shown that password manager users underutilize password generation [19, 28]. One potential explanation for

this phenomenon is that websites' password composition policies (PCPs) can reject generated passwords, decreasing the usability and utility of the generator. [16, 25].

To address this issue, we design a PCP language that websites can use to encode and publish their PCP, with password managers downloading the PCP to ensure that they only generate compliant passwords. To inform the design of this PCP language, we extract 270 PCPs from a geographically diverse set of 626 popular websites. Using this dataset, we build an initial PCP language, then iteratively refine it as we encode the gathered PCPs, stopping once all PCPs in our data set can be efficiently and useably encoded. Our final PCP language is more feature-rich than previous efforts and is the first PCP language that can represent the full range of PCPs found in our dataset.

To demonstrate the feasibility of our proposed language, we (i) build proof-of-concept websites that publish their PCP using our language; (ii) modify BitWarden, a popular password manager, to download these PCPs and generate compliant passwords; and (iii) create Python and JavaScript libraries that make it easy to use our PCP language in server- and client-side code. Next, we conduct an online usability study with 25 participants, measuring their ability to author PCPs using our language and tools. Our results show that most participants can rapidly comprehend our language and author PCP descriptions, even for complex policies.

Finally, we replicate and extend prior work analyzing Web PCPs [10, 20]. In contrast to prior efforts that use a simple heuristic that only considers the minimum length and allowed characters for measuring PCP strength, our analysis takes into account all requirements of the PCP. Additionally, our analysis includes both upper- and lower-bound estimates for PCP strength that take into account how users select passwords [18, 36]. This improved analysis shows that most PCPs in our dataset fail to require passwords that resist offline attacks. Furthermore, for users that prefer passwords comprised primarily of digits [18], nearly half of the evaluated PCPs fail to require passwords that resist online attacks.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

Research Artifacts: Our data, scripts, and prototypes are available at <https://userlab.utk.edu/publications/gautam2022improving>.

2 PCP Dataset

To inform the design of our PCP language, we gathered an extensive corpus of PCPs deployed on the Web. Our sample is demographically diverse, including websites from highly-populated countries in each of the six inhabited continents: Africa—Nigeria, Asia—India, Europe—Germany and the United Kingdom (UK), Oceania—Australia, North America—United States (US), South American—Brazil. We also measured PCPs from China, Iran, and Russia to see if their high levels of Internet censorship [15] impacted PCP selection.

2.1 Sources

We used the Alexa and Quantcast lists of the most popular websites to select websites for each country. In January 2019, we downloaded the Alexa lists of the 250 most popular US websites and the top 50 lists for the remaining nine countries we examined. As we began to analyze these websites, we noticed a high overlap between the websites listed for each country. To obtain more unique websites for each country, in February 2019, we downloaded the Quantcast lists of the top 50 most popular websites for each country. We selected Quantcast as its country-specific lists had minimal overlap with global and US-specific websites from Alexa. We also analyzed the websites listed in the Quantcast top 50 global lists. In total, these lists identified 626 unique websites.

Next, we removed websites that do not support account creation, delegate all authentication to single sign-on (SSO) providers, or require resources we do not have to create an account (e.g., a bank account). For the remaining 320 websites, we identify websites that use the same authentication backend (e.g., google.com and youtube.com), keeping only a single representative website. We then extracted PCPs from the remaining 270 websites.

2.2 Analysis

To extract the PCP for each website, we took the following steps. First, we would look for PCP components described textually on the account creation web page or elsewhere on the domain. Second, we would examine the HTML form, looking for validation attributes that restricted what users could enter for their password. Third, we evaluated any JavaScript used to validate the password, identifying restrictions enforced therein. Fourth and finally, we manually tried to enter various passwords of different lengths and compositions.

2.3 Limitations

While our data collection resulted in a large and rich corpus, we recognize there are limitations to our methodology. First, while covering more features than past efforts [10, 14, 20], our data is not comprehensive. Still, we believe our dataset is sufficient for our purposes as we achieved saturation [2]—i.e., we stopped discovering new PCP features at the latter end of our analysis.

Second, it is likely that we missed some PCP edge cases. Only by investigating the server-side code would it be possible to identify the exact PCP definitively. Automating the process to check more password combinations would be problematic as this would involve flooding the website with passwords.

3 PCP Description Language

Using our PCP dataset, we design a language for describing PCPs. Our language has two key design goals: (1) describe the PCPs in our dataset and (2) be simple to read and write for administrators and machines. To achieve these goals, we followed an iterative design process:

First, we created a draft version of our PCP language based on prior research (§9) and PCP features in our dataset. Second, we encode the PCPs in our data set using this language. When we encountered a PCP that was onerous to encode, we modified our draft PCP language to address pain points. We would then re-encode all prior PCPs to ensure that our change did not cause a usability regression. Third, after encoding all PCPs, we reviewed our language with others from our research group, focusing on improving the language’s readability and identifying PCP features they had encountered in the wild but are absent in our PCP dataset. Based on their feedback, we updated our language and re-encoded the PCPs in our dataset (continuing to look for usability issues). After making a full pass encoding PCPs without changing our language, we considered it finished.

3.1 PCP Language

A PCP in our language is composed of two components: (a) a set of characters allowed in a password and (b) rules about password composition.

The allowed characters are grouped into named, disjoint sets of characters—a *charset*. By default, the PCP uses the following four default charsets: lowercase English letters (*lower*), uppercase English letters (*upper*), Arabic numerals (*digits*), and the OWASP password symbols [26] (*symbols*). Our language allows these default charsets to be modified, new charsets to be added, and default ones to be removed. Our language also provides an *alphabet* charset that, if used, merges and replaces the default *lower* and *upper* charsets.

A PCP composition rule is a set of *requirements* that passwords must comply with to be valid. If a PCP contains

multiple rules, a password need only satisfy the requirements for a single rule to be valid (the overwhelming majority of PCPs only have one rule). For example, if one rule specified that passwords must be eight characters long and contain lowercase letters and symbols and another rule specified that passwords must be fifteen characters long, fifteen character passwords of only digits would be valid, whereas fourteen character passwords of only digits would not.

The possible requirements in each rule are as follows:

- *min_length* is a positive integer specifying the password's minimum length (inclusive). All rules require that *min_length* is set, with all other requirements optional.
- *max_length* is a positive integer specifying the password's maximum length (inclusive).
- *max_consecutive* is a positive integer indicating the maximum number of times the same character can appear consecutively in a password. For example, to prevent passwords such as *AAA* or *ZZZ*, *max_consecutive* would be set to 2.
- *prohibited_substrings* is a set of strings that may not appear anywhere in the password. When used, this commonly includes the website name and other related words. For example, to prohibit the string "google", *prohibited_substrings* would be set to ["google"].
- *require* is a list of charsets that must appear in the password. For example, to require that a password must have letters and digits, *require* would be set to ["alphabet", "digits"].
- *require_subset* is an object containing a list of charsets (*options*) from which *count* of those options must appear in the password. For example, to require that a password must have digits and symbols, but not necessarily both, *require_subset* would be set to {"options": ["digits", "symbols"], "count": 1}. If not set, *options* defaults to using all the PCP's charsets; *count* defaults to one.
- *charset_requirements* is a map between charset names and requirements for the named charset. For example, to add additional requirements for digits, *charset_requirements* would be set as such: {"digits": {requirements}}. Possible requirements include:
 - *min_required* is a positive integer specifying the minimum number of times this charset must appear in the password.
 - *max_allowed* is a positive integer specifying the maximum number of times this charset may appear in the password. For example, if set to two for the digits charset, passwords containing *111* or *123* would be rejected.

```
{
  "charsets": {
    "name": "characters", ...
  },
  "rules": [{
    "min_length": Z+,
    "max_length": Z+,
    "max_consecutive": Z+,
    "prohibited_substrings": ["substring", ...],

    "required": ["charset_name", ...],
    "require_subset": {
      "options": ["charset_name", ...],
      "count": Z+
    },

    "charset_requirements": {
      "charset_name": {
        "min_required": Z+,
        "max_allowed": Z+,
        "max_consecutive": Z+,
        "required_locations": [Z+, ...],
        "prohibited_locations": [Z+, ...],
      }, ...
    }, ...
  ], ...
}
```

Listing 1: JSON schema for our PCP language

- *max_consecutive* is a positive integer indicating the maximum number of times this charset can appear consecutively in a password. For example, if set to two for the alphabet charset, passwords containing *abc* or *ddd* would be rejected.
- *required_locations* is a list of indices for the password at which this charset must appear. Passwords are zero-indexed and negative indices are supported (i.e., reverse string indexing). For example, to require a password that starts and ends with a symbol, *required_locations* for the symbols charset would be set to [0, -1].
- *prohibited_locations* is a list of indices for the password at which this charset must *not* appear. Passwords are zero-indexed and negative indices are supported (i.e., reverse string indexing). For example, to prevent a password from having the last two characters as digits, *prohibited_locations* for the digits charset would be set to [-1, -2].

A JSON schema for our final PCP language is given in Listing 1. Examples of real-world PCPs encoded using our language are given in Listing 2.

Examining the JSON-encoded PCPs in our dataset, we find that they are 17–205 characters long, with a median length of 36 characters. These small sizes are evidence that our PCP efficiently encodes passwords. Lastly, we note that while we used JSON to encode policies, they could also easily be encoded in a wide range of data-interchange formats (e.g., YAML, protobuf).

```

# Passwords of length six to twelve (walmart.com)
{"min_length": 6, "max_length": 12}

# Password must include at least one digit, symbol, and
  alphabetic character (facebook.com)
{
  "min_length": 6,
  "require": ["digits", "alphabet", "symbols"]
}

# Custom definition for symbols that are allowed
  (macys.com)
{
  "charsets": {"symbols": "!\\\"#$%&'()*+;,;<>@[\\]^_`{|}~"},
  "rules": [{"min_length": 7, "max_length": 16}]
}

# Password must have at least one alphabetic character
  and either a digit or a symbol (bbc.com)
{
  "min_length": 8,
  "max_length": 50,
  "require": ["alphabet"],
  "require_subset": {
    "count": 1,
    "options": ["digits", "symbols"]
  }
}

# Password can be eight characters if it contains a
  lowercase character and a digit. Otherwise, it must
  be fifteen characters long. (github.com)
{
  "rules": [
    {"min_length": 8, "require": ["lower", "digits"]},
    {"min_length": 15}
  ]
}

```

Listing 2: PCP examples encoded in our language

4 PCP-Compliant Password Generation

To demonstrate the feasibility of our proposed language, we (1) created libraries for using our PCP language, (2) built proof-of-concept websites that publish their PCP using our language, and (3) modified a password manager to generate PCP-compliant passwords.

4.1 Library Implementations

We constructed Python¹ and JavaScript² libraries to support our PCP language. These libraries enable the programmatic creation of PCPs, encoding PCPs to JSON, and parsing PCPs from JSON. They also automatically validate PCPs to ensure they are both semantically correct—e.g., that *min_length* is appropriately set and that character sets do not overlap—and logically consistent—e.g., that a policy does not simultaneously require and prohibit a character class.

These libraries also support checking passwords against a PCP. Finally, they can evaluate the strength PCPs, giving administrators an idea of how likely a PCP is to result in

¹<https://pypi.org/project/password-policy/>

²<https://www.npmjs.com/package/password-composition-policy>

passwords that resist online and offline guessing attacks (see Appendix B for more details).

4.2 Website Implementation

We built five proof-of-concept websites, each with a PCP of varying complexity. We implemented these websites using Flask (Python) on the backend and JavaScript on the frontend. Each website publishes its PCP and provides a form where passwords can be generated, submitted, and verified.

We identified three approaches for publishing PCPs:

1. **HTML:** A new attribute could be added to the password field, which would be set to the JSON-encoded PCP. Alternatively, the PCP could be encoded as XML within the HTML, adjacent to the password field.
2. **HTTP header:** An HTTP header (e.g., *X-PCP*) can specify the JSON-encoded PCP for relevant pages.
3. **File:** The JSON-encoded PCP could be available at a known URL (e.g., `domain.tld/pcp.json`). If there are multiple PCPs for a domain, this file could contain a mapping between URLs and PCPs.

Our websites use the third approach as it is the easiest to implement and the only approach which can work with non-browser-integrated managers. We checked the validity of submitted passwords on the client-side using our JavaScript library and on the server-side using our Python library. *A significant benefit of publishing PCP and using our tool to validate them is that if the PCP is ever updated, there is no need to separately update the validation code, simplifying developer workloads and preventing situations where the client- and sever-side validation may become out of sync.*

4.3 Password Manager Implementation

We modified BitWarden, a popular open-source password manager, to check if a domain hosts a `/pcp.json` file, and if so, to use it to generate PCP-compliant passwords. The actual generation is handled by our JavaScript library and occurs over three phases:

In the first phase, we set the password length to the smallest *min_length* (if there are multiple rules). Next, we use our JavaScript library to check if passwords of this length using this PCP will be offline-resistant password [11]. If not, we choose the smallest length that would result in an offline-resistant password.

In the second phase, we create an array of length equal to our calculated minimum length. Each position within the array contains an (initially empty) list of which charsets can appear at that position. To fill these lists, we first satisfy *required_locations* by setting the list at the specified index to its respective charset. Next, we set the remaining empty lists as necessary to satisfy *min_required* and *required*. Lastly, the

remaining empty lists are set to include all allowed character sets unless doing so would violate *max_allowed*.

In the third phase, we shuffle all indices not set due to *required_locations*. We then generate a password by randomly selecting a character at each index from the charsets in the list at that index. We then check the generated password against the other requirements in the PCP. If it is not, we repeat phase three until we generate a valid password. In addition to ensuring that generated passwords are PCP-compliant, we also follow recommendations by Oesch et al. [24] and ensure that generated passwords are not randomly weak. This is done by checking passwords using *zxcvbn* and ensuring that the generated passwords receive the highest strength rating (4).

5 Usability Study

To evaluate the usability of our developed language and libraries, we conducted an IRB-approved user study wherein participants authored five PCPs of varying complexity using our PCP language. This section gives an overview of the study and describes the tasks and study questionnaire. In addition, we discuss the development and limitations of the study. The study instrument is given in Appendix A.

5.1 Study setup

The study ran for three weeks starting Friday, January 28, 2022, and ending Tuesday, February 15, 2022. In total, 25 participants completed the study. The study was designed to take about thirty to forty minutes and participants were compensated with a \$25 Amazon gift card. Participants were required to have Python 3.6.1 or higher installed on their system. The study was administered online using Qualtrics.

Participants were recruited from the EECS department at our local university using posters, email invitations, and class announcements. We also asked researchers at other universities to share the study with their students. We chose to use EECS students as we felt they were a good representation of novice developers, and we hypothesized that our language and libraries would be sufficiently usable to support novice developers.

5.2 Study tasks

Participants started by reading and accepting an informed consent statement. Next, participants installed our Python library and executed a Python instruction that allowed us to confirm that the library was correctly installed. They then entered basic demographic information (class standing, major, gender).

Participants were told that in the study they would be authoring five PCPs. They were given a link to documentation for the Python library and informed that this

link would also be provided with each task. The documentation included a description of our language, source code examples, and JSON-encoded PCPs.

Participants encoded five PCPs:

1. The password must be at least 8 characters.
2. The password must be at least 8 characters and contain at least two of the following: uppercase, lowercase, digits, symbols.
3. The password must be at least 12 characters, contain a letter and a number, and not contain whitespace.
4. The password must be at at least 8 characters long and contain a letter and a number. Alternatively, the password must be at least 15 characters.
5. The password must be at least 8 characters, contain at least two symbols, contain either an upper or lowercase letter, not contain the string "mywebsite", and none of the following characters: `^'";/\`

Upon submitting a PCP, the survey checked whether the submitted PCP was parsed correctly. It also verified that the PCP was correct by checking two valid and two invalid passwords. Participants were allowed to continue when they submitted a correct PCP description or once two minutes had passed (to prevent participants from becoming stuck). After submitting their policy, participants completed an After-Scenario Questionnaire [31] (ASQ) about their experience.

Upon completing all five policies, participants were asked to fill out the System Usability Scale [6] (SUS) regarding their overall experience. They were also asked what they liked most and least about the system and library. Finally, they were asked to provide any other feedback they had.

5.3 Demographics

Participants were largely male: male (19; 76%), female (6, 24%). All students studied computer science (23; 92%) or electrical engineering (2; 8%). Participants were all more senior students: juniors (2; 8%), seniors (10, 40%), graduate students(13, 52%).

5.4 Study Design

Initially, we structured study compensation as a raffle, where five participants would receive a \$50 Amazon gift card. Under this incentive scheme, only two participants completed our study. This led us to revise our study to compensate every participant (including the two who had already completed it). After making this revision, re-obtaining IRB approval, and re-launching the study, we quickly gathered our remaining 23 participants.

We also changed our documentation between the two iterations of our study. Initially, the survey provided a link to the documentation explaining how to author policies in JSON, with that documentation providing a link the Python

Policy	Correct	JSON mistakes	Minor errors	Major errors	Mean time in minutes	Mean ASQ
1	92%	0	0	2	1.5	7.0
2	92%	1	0	1	1.4	6.7
3	88%	1	2	1	4.6	5.7
4	96%	1	1	0	0.6	6.3
5	64%	3	7	0	4.0	6.0

Table 1: Quantitative results by policy

library’s documentation. However, after looking at the first two participants’ results, it became clear that they lacked proficiency in JSON. To encourage participants to use the Python library, we changed the survey’s documentation link to point to the Python library’s documentation, with that documentation providing a link to the JSON documentation. Participants could still directly author JSON, and eight (40%) did for at least one task.

5.5 Limitations

Our students do not have the same experience as the administrators responsible for authoring PCPs. Similarly, participants had less incentive to learn and correctly enter policies than administrators trying to use these tools. As such, our results may not fully represent the usability of our tooling for the target audience. However, past research has shown that students can serve as a reasonable approximation for developers [22, 23]. Lastly, our study only measured the ability of participants to author policies, not to read them.

6 Study Results

In this section, we report the significant findings of our user study. Quantitative results for each policy are given in Table 1. Mean completion times use the geometric mean [31].

6.1 Success Rates

Overall, participants did very well at encoding policies. Two participants struggled at nearly all tasks, only correctly encoding a single PCP. Excluding them from our data, completion rates move to 100%, 100%, 96%, 100%, and 68%, respectively.

In policies, we detected three types of errors. First, incorrectly formatted JSON (6 total), likely stemming from unfamiliarity with JSON. Second, minor errors (10 total), such as forgetting to include a prohibited character or including a rule from a previous policy. We only classify errors as minor if users showed comprehension of the tested language and library features but made an error with the values used. Third, major errors (4 total) resulting in an entirely incorrect submission. These errors indicate that

participants failed to understand how to use the language and library.

Looking at Policy 5’s results more closely, we see that three errors (12%) arose due to incorrectly encoded JSON, with the remaining seven (28%) arising due to participants forgetting to include one or more of the prohibited characters. This happened even though these same participants had properly excluded characters in Policy 3.

6.2 Completion Times

Participants generally completed tasks quickly, with (geometric) mean times ranging between 36 seconds and 4 minutes. However, we note that these times are lower bounds as they do not include time participants may have spent reading documentation between tasks and before they started interacting with the task. Still, these times suggest that it is easy to pick up and use our language and library with no prior experience.

Using a two-way ANOVA, we find that while there is a statistically significant difference between how long each policy took to create ($F(4, 170) = 8.731, p < 0.001$), though this is not surprising given the difference in difficulty between policies. We do not find a statistically significant difference between time taken to author PCPs using JSON or our library ($F(1, 170) = 0.109, p = 0.74$), nor for the interaction effect ($F(4, 170) = 0.027, p = 1.00$). This is a surprising result as, based on our first two respondents, we expected participants to struggle authoring JSON.

6.3 Perceived Usability

Overall, policies received good ASQ scores (see Table 1), indicating that it was easy and relatively quick to author policies. The mean SUS score was 65, which can be interpreted as “Good” usability [3], receives a C grade [31], and is just above the 40th percentile of systems studied with SUS. While this is an acceptable score for our language and library to be used in the wild [3], it still fell short of our initial expectations.

Looking into the qualitative feedback, we discovered three primary critiques of our tooling. First, many participants felt that JSON was confusing. Second, participants wanted additional documentation. While we provided one example for every PCP feature, they wanted even more. Third, participants were confused by our library providing two ways to create PCPs: (a) a class exactly matching the JSON schema and (b) a simplified class that could be used to encode simple PCPs more directly. While we created this second method to reduce the amount of code participants needed to write for simple PCPs, it ended up causing unneeded confusion and is a prime candidate to remove from our library.

6.4 Takeaways

Overall, our results show that our proposed language is promising, though it has room for improvement. Other than the two participants who failed all but one task, every other participant correctly encoded Policies 1–4, except for one mistake in Policy 3.

However, of these 23 participants, nine (39%) submitted incorrect solutions for Policy 5. One-third of these errors (3) arose from improperly encoded JSON. This suggests that in line with participant feedback, it might be worthwhile to consider other more developer-friendly encodings (e.g., YAML) or supporting multiple encodings, allowing developers to choose which they will use. Alternatively, pushing for programmatic specification of PCPs could be used to avoid encoding issues entirely.

Two-thirds of the errors (6) for Policy 5 arose from minor issues with the PCP. Half of these issues (3) involved the participants removing some but not all of the prohibited characters from the symbols list. This may have arisen as the textual policy described a denylist for restricted characters, whereas participants chose to create an allowlist of symbols. To address this, the library could allow users to specify a denylist for characters and then have the library generate the appropriate character set, though further research would be needed to measure the efficacy of this approach.

The other issues with Policy 5 (3) arose from participants failing to include the list of restricted characters, even though the other requirements for this policy were included. This happened even though these same participants had properly excluded characters in Policy 3. It is unclear whether this issue stems from something in the design of our language, the general challenge of remembering all the requirements in a complex policy, or study fatigue.

7 Website Analysis

Using the PCP dataset we collected to build our language, we replicated and extended prior work analyzing website PCPs [10, 20]. Our analysis covers (1) the strength of PCPs, (2) the requirements used in PCPs, and (3) additional non-PCP authentication-related details.

To estimate PCP strength, we calculate the average number of guesses an adversary would need to discover a password that (a) complies with the PCP and (b) is of the smallest allowed length. In contrast to previous work [10, 20] which calculates strength based only on the smallest allowed length and count of allowed characters (i.e., $\#characters^{length}$), our estimates take into account all PCP features. First, we create a canonical representation of the PCP. Second, we enumerate all unique password compositions—a password composition specifies the number of characters from each character class that makes up a password. Third, for each password composition, we

Country	Count	Popularity	Count	Use case	Count
Global	65	Top 10	8	E-commerce	58
Australia	13	Top 50	24	Finance	10
Brazil	14	Top 100	25	News	72
Germany	17	Top 500	59	Social media	55
India	9	Top 1000	25	Software	13
Nigeria	13	Top 5000	79	Streaming	28
UK	8	5000+	50	Other	34
US	72				
China	28				
Iran	12				
Russia	19				

Ad Provider	Count	Public username	Count	Past breach	Count
Yes	158	Yes	43	Yes	51
No	112	No	227	No	219

Table 2: Number of PCPs in each category

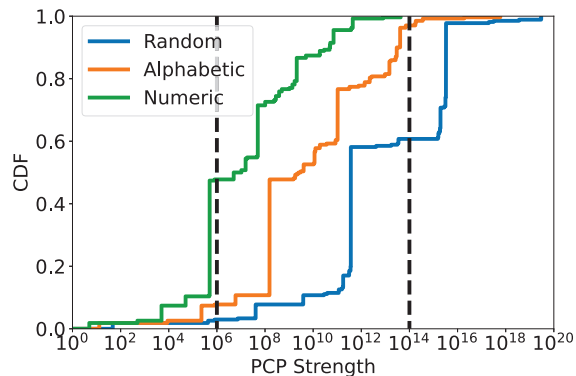
calculate the number of unique passwords that exist for that composition, reducing this number to account for passwords that fail to meet the various *charset_requirements*. Finally, we sum these counts. A more detailed description of this algorithm is given in Appendix B.1.

In addition to estimating PCP strength based on password chosen entirely at random (as is done in previous research [10, 20]), we also consider PCP strength under conditions where users prefer characters from certain character sets: (a) preferring alphabetic (particularly lowercase) characters over non-alphabetic characters (as commonly seen in the US [18]) and (b) preferring numeric characters (as commonly seen in China [18, 36]). These changes help our analysis to more accurately measure the strength of PCPs under a range of usage scenarios. These calculations are performed by modifying our enumeration of password compositions only to include compositions that use the most preferred character classes unless the PCP specifically requires another character class. A more detailed description is given in Appendix B.2.

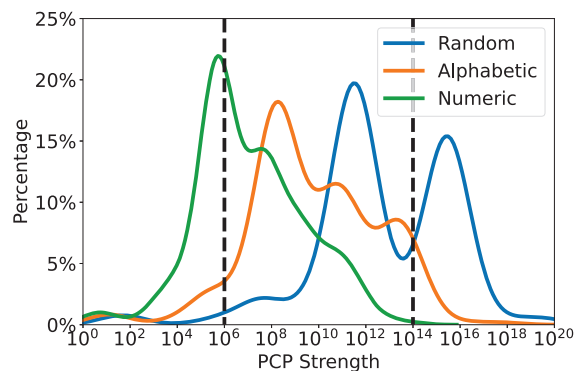
Throughout our analysis, we categorize PCPs by (i) the country where they are popular, (ii) their Alexa global rank, (iii) their use case, (iv) whether they generate revenue by displaying ads, (v) whether usernames on the website were publicly available or easily guessed, and (vi) whether a data breach had been reported for the website. All categorizations are mutually exclusive, with PCPs popular in multiple countries categorized as “Global”. Table 2 lists these categories and the number of PCPs in each.

7.1 PCP Strength

Figure 1 gives the distribution of password strengths. If passwords are generated entirely at random, nearly all PCPs are strong enough to resist online attacks (10^6 guesses [11]), though only about 40% are strong enough to resist offline



(a) CDF of PCP strengths



(b) Distribution of PCP strengths

10^6 and 10^{14} are estimates of the number of guesses a password should resist to survive online and offline attacks, respectively [11].

Figure 1: PCP Strengths

attacks. For passwords where alphabetic characters are preferred, nearly all PCPs fall into the online-offline chasm [9]—strong enough to resist online attacks but not offline attacks (surviving 10^{14} guesses [11]). This chasm is problematic because PCPs in it impose a usability burden to pick more complex passwords than necessary to resist online attacks, but which are still too weak to resist offline attacks. For passwords where numeric characters are preferred, half of the analyzed PCPs are insufficient to prevent online attacks, and none are strong enough to resist offline attacks.

Comparing mean PCP strength under different password generation strategies, we find that passwords generated at random (3.5×10^{17}) are roughly two orders of magnitude stronger than alphabetic-preferred passwords (2.3×10^{15}) and six orders of magnitude stronger than numeric-preferred passwords (2.1×10^{11}). This highlights the benefits of using a password generator to create passwords. It also demonstrates why it is crucial to consider generation strategy when estimating PCP strength, as assuming passwords are selected

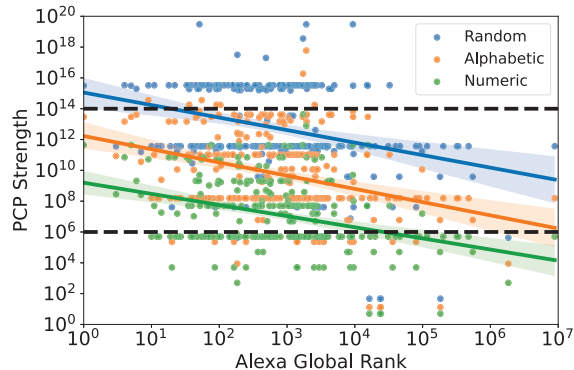


Figure 2: PCP strength by Alexa global rank

entirely at random can significantly overestimate the protectiveness of PCPs.

7.1.1 Strength by Category

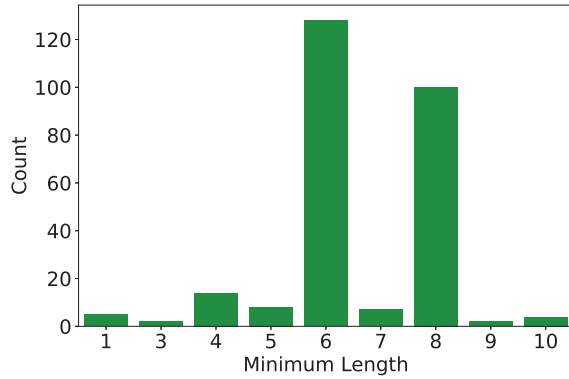
Figure 2 shows the correlation between PCP strength and a website’s Alexa global ranking. In general, we find that higher-ranked websites have stronger PCPs. Using Pearson’s r and log scales for both rank and PCP strength, we find a medium effect size for entirely random ($r = -0.30$, $p < 0.001$), alphabetic-first ($r = -0.34$, $p < 0.001$), and numeric-first ($r = -0.34$, $p < 0.001$) strengths.

We found a statistically significant difference between strengths based on country for generation at random and alphabetic first generation, but not for numeric-first generation (one-way ANOVA—entirely random— $F(10, 259) = 1.87$, $p < 0.05$; alphabetic-first— $F(10, 259) = 2.05$, $p < 0.05$; numeric-first— $F(10, 259) = 0.29$, $p = 0.98$). We did not find any meaningful pairwise differences for the statistically significant results using Tukey’s test. There was no significant difference based on use case (entirely random— $F(5, 263) = 1.04$, $p = 0.40$; alphabetic-first— $F(5, 263) = 0.59$, $p = 0.74$; numeric-first— $F(5, 263) = 0.40$, $p = 0.88$).

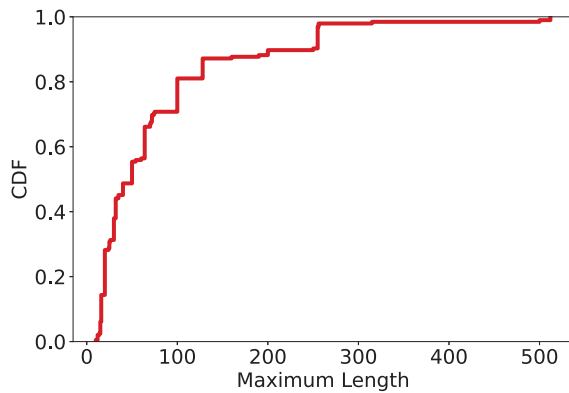
Figures showing strength differences based on country, global rank, and use case can be found in Appendix D. We also tested whether (i) ads, (ii) public usernames, (iii) or data breach history impacted PCP strength, finding no statistically significant differences.

7.2 PCP Features

The most common minimum lengths for PCPs are 6 (128; 47%) and 8 (100; 37%) (see Figure 3a). Just over a tenth of PCPs (29; 11%) allowed passwords with fewer than 6 characters, with five (5; 2%) allowing passwords with a single character. These low length requirements are not only problematic for user-generated passwords but also for



(a) Histogram of minimum lengths



(b) CDF of maximum lengths

Figure 3: PCP lengths

password generators, which are known to occasionally generate random but weak passwords at shorter password lengths [24].

Most PCP rules (195; 72%) set a maximum length for passwords, with a wide range of values (see Figure 3b). Just over a tenth (28; 10%) limit passwords to 16 or fewer characters, with four (4; 1%) limited to 12 or fewer characters.

The next most common requirement was having required character classes (51; 19%): digits (42/51; 82%), alphabet (37/51; 73%), lower (12/51; 24%), upper (10/51; 20%), and symbols (4/51; 8%). This was followed by requiring a subset of character classes (43; 16%): at least one (5/43; 12%), two (14/43; 33%), or three (13/43; 30%) characters from all character classes; at least one symbol or digit character (9/43; 21%); at least one upper or symbol character (1/43; 2%); or at least one upper, digit, or symbol character (1/43; 2%).

The remaining requirements only appeared rarely. For prohibited substrings (11; 4%), websites primarily restriction personal information (10/11; 91%): name (6/11; 55%), email (5/11; 45%), username311, birthday211, website

name (1/11; 9%). Rules also included max consecutive characters (9; 3%) with values of one (1/9; 11%), two (2/9; 22%), three (4/9; 44%), and seven (1/9; 11%). Finally, one PCP (1; 0%) required two lower case letters and two digits.

7.2.1 Multi-Rule PCPs

Of particular interest, we discovered three PCPs (3; 1%) that had more than one rule.

gumtree.com.au Required twenty-character passwords unless the password included an alphabetic character and either a digit or symbol, in which case ten-character passwords were allowed.

github.com Required fifteen-character passwords unless the password included both a lowercase character and a digit, in which case eight-character passwords were allowed.

yy.com Required nine-character passwords unless the password included an alphabetic character, in which case an eight-character password could be used. This could be to encourage Chinese users to pick non-digit-only passwords, which is common in that culture [18, 36].

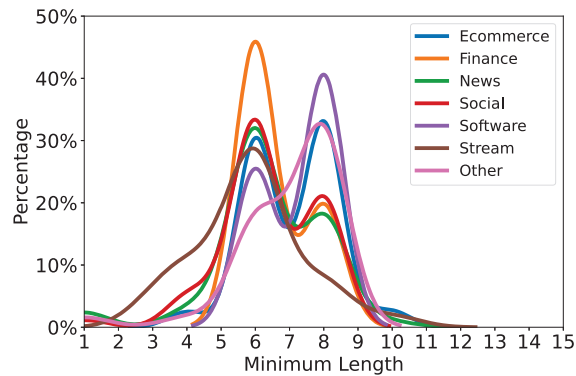
Ignoring specific requirements, these PCPs all share a common goal: allow users to choose between short but complex or long but simple passwords.

7.2.2 Features by Category

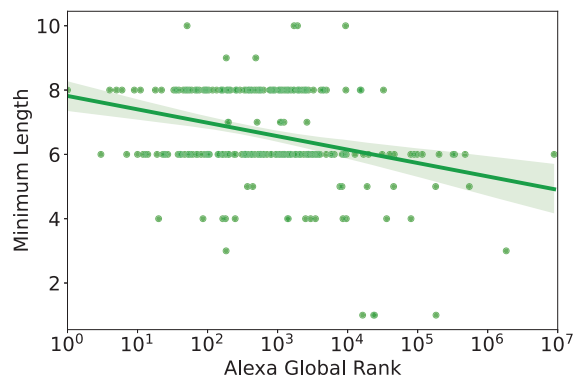
We find statistically significant difference for minimum length by country (one-way ANOVA— $F(10, 259) = 2.74$, $p < 0.01$), global rank (Pearson’s $r = -0.30$, $p < 0.001$), and use case (one-way ANOVA— $F(6, 263) = 3.57$, $p < 0.01$). Within these categories, high-ranked websites are much more likely to allow passwords shorter than six characters (see Figure 4b). Similarly, “streaming” websites have lower minimum length requirements (see Figure 4a), with the difference being statistically significant for “Ecommerce” ($p < 0.01$) and “Other” ($p < 0.05$).

We did not find statistically significant differences in maximum length by country (one-way ANOVA— $F(10, 259) = 1.05$, $p = 0.40$), global rank (Pearson’s $r = 0.01$, $p = 0.88$), or use case (one-way ANOVA— $F(6, 263) = 1.40$, $p = 0.21$). We did not see any meaningful difference for other restrictions, though we did not test for statistical significance.

Figures showing differences for minimum and maximum length based on country, global rank, and use case can be found in Appendix E. We also tested whether (i) ads, (ii) public usernames, (iii) or data breach history impacted PCP minimum and maximum length, finding no statistically significant differences.



(a) By use case



(b) By Alexa global rank

Figure 4: PCP minimum lengths

7.3 Website Analysis

We also examined the following items for each website: (a) whether account creation and login required HTTPS, (b) which SSO providers, if any, were supported, and (c) whether a password strength meter is shown to users.

For most websites (255; 94%) HTTPS was required to view the account creation and login pages. Still, there were fifteen (15; 6%) websites where we could access the account creation or login interface over HTTP.³

A third of websites (92; 34%) support at least one single sign-on (SSO) provider for account creation and authentication. The most popular SSO providers are Facebook (82/92; 89%), Google (65/92; 71%), Twitter (21/92; 23%), VK (10/92; 11%), and mail.ru (6/92; 7%), with the remaining 20 SSO providers being represented on fewer than five websites.

We find that just over a tenth (35; 13%) of websites show users a strength meter when they are creating passwords. We also find that just under a tenth (22; 8%) use a strength checker as part of their password policy—i.e., passwords must be a certain strength to be accepted.

³The list of websites is given in Appendix C.

7.3.1 Websites by Category

For websites whose account creation or login pages can be accessed over HTTP, the majority were in China: China (8/15; 53%), Russia (2/15; 13%), and one each (1/15; 7%) for India, Iran, Nigeria, Brazil, and the US. It is unclear why China is so different, but we find this correlation troubling. These types of websites are most likely to occur in less popular websites.³

Within certain countries we see much higher rates of adoption of SSO: Russian (11/19; 58%), Nigeria (6/13; 46%), Brazil (6/14; 43%), Australia (5/13; 38%), UK (3/8; 38%), India (3/9; 33%), Global (21/65; 32%), US (21/72; 29%), China (6/28; 21%), Iran (2/12; 17%). We also see a trend that the less popular sites are more likely to adopt 2FA: Top 10 (1/8; 13%), Top 50 (7/24; 29%), Top 100 (4/25; 16%), Top 500 (17/59; 29%), Top 1000 (6/25; 24%), Top 5000 (37/79; 47%), 5000+ (20/50; 40%). For categories, SSO is more evenly dispersed, though news (35/72; 49%) sites have higher support for SSO.

We do not find any meaningful effect from the categories on strength meters or internal strength checks for passwords.

8 Discussion

In this section, we discuss observations from our research.

8.1 PCP Recommendations

Of all the PCPs encountered in our analysis, we were most interested in the multi-rule PCPs, which allowed users to choose between short but complex or long but simple passwords. This ensures that passwords will resist offline attacks without causing unnecessary usability burdens. Moreover, this approach returns the locus of control to users—i.e., while PCPs are often viewed as restrictive, and therefore less usable [17, 32, 33], multi-rule PCPs give users a choice of which PCP is most appropriate for them. We hypothesize that by giving this control back to users, not only will they be more satisfied with the PCP, but they will also create stronger passwords. Future work could validate this hypothesis and try to determine what the ideal multi-rule construction is. For example, would more rules be even better, providing even more fine-grained control of the types of passwords users can select?

Another observation from our analysis is the importance of PCP design for ensuring the security of passwords not generated entirely at random. Whereas PCP requirements reduce the strength of passwords generated entirely at random (by shrinking the search space), they increase the strength of passwords generated with preferences to a given character class. Thus there is an interesting interplay between PCPs and passwords based on how they are generated. More specifically, we note that increasing length is the easiest way

to improve strength, regardless of generation strategy. Similarly, we find that it is likely advantageous to limit users from having too much of their password be composed of digits (or symbols), as this significantly weakens those passwords and may lead to passwords vulnerable to online guessing attacks. As such, we recommend that administrators use a multi-rule approach that allows users to choose between long but simple passwords or short but complex passwords. This allows machine-generated passwords to be short but ensures that human-generated passwords are strong enough to resist attack.

8.2 NIST Guidelines

NIST provides PCP guidelines (i.e., non-compulsory recommendations) for US companies and organization [12]. While our dataset includes a wealth of PCPs for global and non-US websites, we still think it is interesting to see which of these PCPs conform to the NIST guidelines.

We find that less than half of PCPs (106; 39%) meet NIST’s recommended minimum length of eight characters. Similarly, we find that most (195; 72%) implement unnecessary maximum length requirements.

In line with NIST recommendations, most PCPs (177; 66%) do not have any composition requirements (this would be more positive if they met the minimum length requirements). Similarly, only a small fraction (8; 3%) reject specific symbols, which can be an indication of improper password hashing.

9 Related Work

This section discusses related work on password generation, PCP languages, analysis of Web PCPs, and PCP usability.

9.1 PCP Languages

There have been previous proposals for building PCP languages, with each providing a different subset of the features used in our PCP language (see Table 3). Two proposals involve adding additional HTML attributes to input fields to specify PCP requirements [4, 21], though they only cover a small subset of the most common PCP features.

Horsch et al. developed an XML-based PCP language by automatically scanning and extracting PCPs for 72,125 services. Based on a sample of 200 manually verified PCPs, they estimated that their algorithm correctly extracted PCPs in just over four out of five cases, with the remaining cases evenly split between mostly correct and incorrect. Their resulting PCP language has most of the features found in our language. However, it is missing support for multiple rules, requiring a subset of character classes, limiting maximum consecutive characters from the same character class, and set required and prohibited locations based on distance from the

PCP Features	This paper	Daniel Bates [4]	Isiah Meadows [21]	Horsch et al. [14]
Define character sets	✓	✓	✓	✓
Multiple rule sets	✓			
<i>min_length</i>	✓	✓	✓	✓
<i>max_length</i>	✓	✓	✓	✓
<i>max_consecutive</i>	✓	✓	✓	✓
<i>prohibited_substrings</i>	✓		✓	
<i>required</i>	✓	✓	✓	✓
<i>require_subset</i>	✓		✓	
<i>charset_requirements</i>				
<i>.min_required</i>	✓			✓
<i>.max_allowed</i>	✓			✓
<i>.max_consecutive</i>	✓			
<i>.required_locations</i>	✓			✓
<i>.prohibited_locations</i>	✓			✓
reverse indexing	✓			

Table 3: Comparison between PCP languages

end of the password. This demonstrates the limitation of this type of automated PCP extraction—i.e., it can only find PCP features that the automated tool expects to find.

Examining our data, none of these PCP languages can encode all the PCPs in our dataset. However, these proposals could be extended to support the features identified in our research. During our PCP language generation process (see §3), our team built and tested several versions of our PCP language that were HTML- and XML-based. Ultimately, we rejected these approaches because our team felt that encoding policies in these languages was cumbersome and that the resulting policies were difficult to read. Still, the results of our user study show that there is significant room for improving our proposed language, and future work could explore integrating paradigms from these prior proposals with our language or testing whether, contrary to our team’s perceptions, HTML- or XML-based would be better received than our JSON-based approach by developers. In this regard, the main contribution of our paper is the identification of features that must be included in such PCP languages.

9.2 Web PCP Analysis

In 2010, Florêncio and Herley [10] retrieved PCPs for 75 websites in the US. They found that contrary to their intuition, the importance of a website had little correlation to the PCP used on that website. In many cases, the largest, most important websites had the weakest PCPs. They suggested that the reason for this was that due to market economics, these larger websites needed to be more concerned with usability than security, being able to absorb

the security cost of weak PCPs more readily than smaller sites.

In 2016, Mayer et al. [20] replicated and extended the work of Florêncio and Herley. In addition to re-examining 70 of the websites used in the original study (five did not work), they also analyzed 67 German websites. They find that overall, PCP strength has been increasing, though German PCPs, on average, are weaker than US PCPs.

In this paper, we replicate and further extend this work. We collect a dataset that is roughly twice as large and five times more geographically diverse than Mayer et al.'s dataset. Compared to this prior work, we gather more features of the PCPs used on these websites and develop a more fine-grained estimation of PCP strength.

For the most part, our results are similar to past findings. Overall, PCP strength (for random generation) is similar in all studies. However, as our improved strength calculation results in lower estimates of PCP strength, the similarity of our results suggests that PCPs have continued to get stronger over time, though that progress is slow and the delta is not that meaningful. When using PCP strength estimates based on random generation (as the prior work does), we find that PCP strength has become more bimodal, with a clear contrast between websites that require passwords to be offline-resilient and those that only require online-resilience. While this may only be an artifact of our increased precision in plotting PCP strength (the prior worked binned strength into large ranges), we do not believe so and think it is an area that could be explored more in future research. Like the prior work, we find that most PCPs reside within the online-offline chasm identified by Florêncio and Herley [11].

Like prior work, we find no statistically significant correlations when comparing PCP strength based on country, use case, public usernames, and past breaches. However, unlike the prior work, we find a correlation between a website's popularity and the strength of its PCPs. This difference is most likely explained by (a) our larger data set, (b) the increased fidelity of our strength estimates, and (c) the use of log adjusted strength and global ranks. Also, whereas prior work found a negative correlation between whether a website served ads and its PCP strength, we find no statistically significant correlation.

9.3 PCP Usability

Several studies have examined the effect of password policies on user behavior. These studies have shown that while strong PCPs make passwords harder to crack, they also make passwords harder for users to remember [29]. Furthermore, as the number of passwords a user needs increases, their ability to remember them decreases [1, 35]. This helps explain why when Florêncio and Herley [9] studied password behavior of half a million users, they found that

users had on average 25 passwords and reused any given password on an average of 6.5 different websites.

Other research explores what PCP features make passwords harder to remember, with most research finding that it is complex character class requirements that cause the most difficulty [17, 32, 33]. In contrast, minimum length is not nearly as significant of an impediment, leading researchers to suggest favoring longer but less complicated passwords. More recently, we have seen these suggestions reflected in NIST guidelines [12].

Our research finds that length has the greatest impact on PCP strength for both passwords generated at random and using an alphabetic-first approach. As such, we echo prior recommendations for PCPs to focus on length as opposed to complexity. For those that want the best of both worlds, multi-rule PCPs can be used that allow short but complex or long but simple passwords, giving users the locus of control for this decision and thereby increasing usability. Similarly, due to the weaknesses of digit-first generated passwords, PCPs should likely restrict the usage of too many digits in a password.

10 Conclusion and Future Work

In this work, we developed a PCP language that websites and password managers can use to support the generation of compliant passwords. We hope that our work will signal to both communities that adopting a PCP language has tangible benefits. For websites, it allows them to unify their PCP specification and checking, allowing changes to the PCP file to automatically update how checking happens on both the client and server. For password managers, it not only improves the usability and utility of password management but also supports opinionated generation algorithms (e.g., mobile-aware generation [13], security-focused generation [24]), which would otherwise frequently generate non-compliant passwords.

While we are encouraged by the positive results of our user study, they also indicated that there is room for improvements. Future work could expand our PCP language by identifying and adding support for rarely used PCP features, such as restricting sequences of characters (e.g., "abcde") or keyboard patterns (e.g., "qwerty"). Similarly, our language could be enhanced to allow Unicode characters. Future research could also examine how to allow our PCP language to handle dynamic strings (e.g., usernames). One potential solution is to use placeholders in the *prohibited_substrings* requirement, providing appropriate values to the library at password validation. Finally, research could explore automatically identifying PCPs, both in whitebox scenarios, helping web developers identify their website's PCP, and blackbox scenarios, helping password managers identify PCPs for websites that do not publish it, with care taken to avoid flooding servers with password guesses (approximating a DoS attack).

References

- [1] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.
- [2] Khaldoun M Aldiabat and Carole-Lynne Le Navenec. Data saturation: The mysterious step in grounded theory methodology. *The Qualitative Report*, 23(1):245–261, 2018.
- [3] Aaron Bangor, Philip T Kortum, and James T Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594, 2008.
- [4] Daniel Bates. Proposal: Html passwordrules attribute. <https://github.com/whatwg/html/issues/3518>, 2021.
- [5] Joseph Bonneau, Cormac Herley, Paul C Van Oorschot, and Frank Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy*, pages 553–567. IEEE, 2012.
- [6] John Brooke. Sus: a “quick and dirty” usability. *Usability evaluation in industry*, 189(3), 1996.
- [7] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The Tangled Web of Password Reuse. In *Network and Distributed System Security (NDSS)*, volume 14, pages 23–26, 2014.
- [8] Matteo Dell’Amico, Pietro Michiardi, and Yves Roudier. Password strength: An empirical analysis. In *2010 Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010.
- [9] Dinei Florencio and Cormac Herley. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web*, pages 657–666, 2007.
- [10] Dinei Florêncio and Cormac Herley. Where do security policies come from? In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, pages 1–14, 2010.
- [11] Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. An administrator’s guide to internet password research. In *28th Large Installation System Administration Conference (LISA14)*, pages 44–61, 2014.
- [12] Paul A Grassi, James L Fenton, Elaine M Newton, Ray A Perlner, Andrew R Regenscheid, William E Burr, Justin P Richer, Naomi B Lefkowitz, Jamie M Danker, YeeYin Choong, et al. Nist special publication 800-63b: Digital identity guidelines. *National Institute of Standards and Technology (NIST)*, 27, 2016.
- [13] Kristen K Greene, John Michael Kelsey, Joshua M Franklin, et al. *Measuring the usability and security of permuted passwords on mobile platforms*. US Department of Commerce, National Institute of Standards and Technology, 2016.
- [14] Moritz Horsch, Mario Schlipf, Johannes Braun, and Johannes Buchmann. Password requirements markup language. In *Australasian Conference on Information Security and Privacy*, pages 426–439. Springer, 2016.
- [15] Freedom House. Freedom house (fh) freedom of the press report. <https://freedomhouse.org/reports/publication-archives>.
- [16] N. Huaman, S. Amft, M. Oltrogge, Y. Acar, and S. Fahl. They would do better if they worked together: The case of interaction problems between password managers and websites. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1626–1640, Los Alamitos, CA, USA, may 2021. IEEE Computer Society.
- [17] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. Of passwords and people: measuring the effect of password-composition policies. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 2595–2604, 2011.
- [18] Zhigong Li, Weili Han, and Wenyuan Xu. A large-scale empirical analysis of chinese web passwords. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 559–574, San Diego, CA, August 2014. USENIX Association.
- [19] Sanam Ghorbani Lyastani, Michael Schilling, Sascha Fahl, Michael Backes, and Sven Bugiel. Better managed than memorized? studying the impact of managers on password strength and reuse. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 203–220, 2018.
- [20] Peter Mayer, Jan Kirchner, and Melanie Volkamer. A second look at password composition policies in the wild: Comparing samples from 2010 and 2016. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 13–28, Santa Clara, CA, July 2017. USENIX Association.
- [21] Isiah Meadows. Add password restriction attributes. <https://discourse.wicg.io/t/add-password-restriction-attributes-to-input-type-password/4767>, Sep 2020.

- [22] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On conducting security developer studies with cs students: Examining a password-storage study with cs students, freelancers, and company developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
- [23] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel Von Zezschwitz, and Matthew Smith. "if you want, i can store the encrypted password" a password-storage field study with freelance developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.
- [24] Sean Oesch and Scott Ruoti. That was then, this is now: A security evaluation of password generation, storage, and autofill in browser-based password managers. In *USENIX Security Symposium*, 2020.
- [25] Timothy Oesch. *An Analysis of Modern Password Manager Security and Usage on Desktop and Mobile Devices*. PhD thesis, The University of Tennessee, 2021.
- [26] OWASP. Password special characters. <https://owasp.org/www-community/password-special-characters>, 2021.
- [27] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let’s go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 295–310. ACM, 2017.
- [28] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why people (don’t) use password managers effectively. In *Fifteenth Symposium On Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA, pages 319–338, 2019.
- [29] Robert W Proctor, Mei-Ching Lien, Kim-Phuong L Vu, E Eugene Schultz, and Gavriel Salvendy. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods, Instruments, & Computers*, 34(2):163–169, 2002.
- [30] Shannon Riley. Password security: What users know and what they actually do. *Usability News*, 8(1):2833–2836, 2006.
- [31] Jeff Sauro and James R Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.
- [32] Richard Shay, Saranga Komanduri, Adam L Durity, Phillip Huh, Michelle L Mazurek, Sean M Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Can long passwords be secure and usable? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2927–2936, 2014.
- [33] Richard Shay, Saranga Komanduri, Adam L Durity, Phillip Huh, Michelle L Mazurek, Sean M Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Designing password policies for strength and usability. *ACM Transactions on Information and System Security (TISSEC)*, 18(4):1–34, 2016.
- [34] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. ‘I added ‘!’ at the end to make it secure’: Observing Password Creation in the Lab. In *Proceedings of the Eleventh Symposium On Usable Privacy and Security*, 2015.
- [35] Kim-Phuong L Vu, Robert W Proctor, Abhilasha Bhargav-Spantzel, Bik-Lam Belin Tai, Joshua Cook, and E Eugene Schultz. Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies*, 65(8):744–757, 2007.
- [36] Ding Wang, Ping Wang, Debiao He, and Yuan Tian. Birthday, name and bifacial-security: understanding passwords of chinese web users. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1537–1555, 2019.
- [37] Ke Coby Wang and Michael K Reiter. How to end password reuse on the web. *arXiv preprint arXiv:1805.00566*, 2018.

A Study Instrument

Setup For this study, you will be using a python library we developed. Please install this library using pip: `python3 -m pip install -user -upgrade password-policy`. If for some reason, you don’t have pip installed, you can install it using: `python3 -m ensurepip -user -upgrade`.

After installation is complete, check that everything is working correctly by copying and pasting the following command into your terminal. Enter the resulting output below: `python3 -c "import password_policy; print(password_policy.__version__)"`.

Q1. Enter version

Demographics

Q2.1. What is your class standing?

- Junior
- Senior
- MS student
- PhD student

Q2.2. What is your major?

- Computer Science
- Computer Engineering
- Electrical Engineering
- other [Enter here]

Q2.3. What is your sex?

- Male
- Female
- Non-binary
- Prefer not to answer

Tasks Different websites have different requirements for passwords. For example, some websites may require passwords to have a minimum length, include certain types of characters, and avoid using other characters. In our research group, we are studying a system for describing password policies using JSON. We are also studying libraries that can be used to construct these JSON policy descriptions and validate passwords based on these descriptions.

In this study, you will use this system and a python library to encode several password policies. Our goal is to understand how usable this system and library is.

To help you learn about this system and the python library you installed, please click [this link to view the relevant documentation]. You will be using the knowledge from this documentation for the rest of the study. Feel free to refer to it throughout the study. A link to this documentation will always be available on the pages describing your tasks for this study.

When you feel ready to use this system, click continue to be given your first task.

The following questions were the same for each policy, except for the policy requirements. We give the full text for Policy 1's questions, and only the policy requirements for Policy 2–5.

Q3.1.1. Using the python library, please write a policy description for the following password policy. When finished, encode it in JSON and enter it into the text field below. We will validate the entered policy description to make sure it is correct. You may also directly write the policy as JSON (not using the library) if desired.

Password policy:

- The password must be at least 8 characters long

[Documentation link]

Q3.1.2. Did you manually write the JSON policy description, or did you generate it using the python library?

- Generated it using Python library
- Manually entered the JSON policy

Q3.1.3. Based on your experience authoring the JSON policy description, indicate to what extent you agree with the following statements. Options: Strongly disagree-1..Strongly agree-7

- Overall, I am satisfied with the ease of completing this task.
- Overall, I am satisfied with the amount of time it took to complete this task.
- Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing this task.

Q3.2.1. Password policy:

- The password must be at least 8 characters long
- The password must contain characters from at least two of the following: uppercase letters, lowercase letters, numbers, symbols

Q3.3.1. Password policy:

- The password must be at least 12 characters long
- The password must contain at least one letter and one number
- The password must NOT contain space

Q3.4.1. Password policy:

- The password must satisfy one OR the other of the following policies:
 - The password must be at least 8 characters long
 - The password must contain at least one letter and one number
- OR
 - The password must be at least 15 characters long

Q3.5.1. Password policy:

- The password must be at least 8 characters long
- The password must contain at least two symbols
- The password must contain at least one uppercase letter and one lowercase letter
- The password must NOT contain space, the carrot symbol (^), quotes ('), double quotes ("), semicolons (;), slashes (/), or backslashes (\).
- The password must NOT contain the substring "mywebsite"

Post-Study Questionnaire That was the last policy you will need to write for this study. We will now ask you a few questions about your experience the password policy description system and python library.

Q4.1. Please answer the following question about your experience. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

Options: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strong Agree

1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system

Q4.2. What did you like the most about the system and library?

Q4.3. What did you like the least about the system and library?

Q4.4. Is there any other feedback you would like us to know about the system or library?

B PCP Strength Calculations

We measure the strength of password composition policies (PCPs) by estimating how many passwords exist that (a) satisfy the PCP and (b) are of the shortest possible length. We then divide this number by two to estimate the average number of guesses an adversary needs to find a user's passwords. This approach gives an exact estimate of strength when passwords are generated entirely at random. To estimate strength for human-generated passwords, we allow our strength estimates to be parameterized by what character classes are preferred [18].

B.1 Algorithm

Step 1—Preprocessing First, we filter the rules and only consider those that have the smallest *min_length* (there may be multiple). Next, we simplify *require_subset*, creating a new rule with *require* set for each possible combination of the listed *options* of length *count*. Lastly, we simplify the

shortcut rules *require*, setting *min_require* for each charset listed in the requirement.

Step 2—Enumerating Password Compositions In this step, we enumerate all possible password compositions for the rules identified in Step 1. A password composition is simply a list specifying how many characters from each character class are used to make up a password. For example, for a PCP that (a) only allows lowercase letters and digits and (b) has a rule that sets *min_length* to 2 (but no other requirements), there are three password compositions: (1) two lowercase letters, (2) two digits, (3) one lowercase letter and one digit. Note, we only consider compositions where the sum of character counts equals *min_length*.

We take the following steps to derive the password compositions for a rule. First, we create a password composition with values set based on *min_required* for each charset. We also calculate *required_chars*, which tracks the total number of required characters (sum of the calculated password composition). Second, we create a list of length *min_length - required_chars*. At each index *i* (one-indexed) of this new list, we include a list of which character classes could appear *i* more times in the password composition without violating *max_allowed* for each charset (if set). Third, we calculate the full factorial combination of items in this list of lists. For each such combination, we create a new password composition that takes the original password combination and adds the character classes in the combination. For each composition, we also store any restrictions related to that composition that may not yet have been handled (e.g., *max_consecutive*).

For example, consider a policy with *min_length* set to 3, which requires the *alphabet* character set to be used once and has at most one digit. Our initial password composition would be $[1, 0, 0]$ representing 1 alphabet character, 0 digits, and 0 symbols; *required_characters* would be 1. Our list of lists would be $[[alphabet, digit, symbol], [alphabet, symbol]]$. In total, there are six $(3 * 2)$ possible combinations of this list, which after added to initial password composition result give the following password compositions: $[[3, 0, 0], [2, 0, 1], [2, 1, 0], [1, 1, 1], [2, 0, 1], [1, 0, 2]]$.

This method will not result in overlapping compositions within a given rule but can between rules. If this occurs, duplicate compositions are trimmed.

Step 3—Calculating Combinations and Permutations

For each composition, we will calculate the number of passwords (i.e., size of the search space) represented by each composition that also satisfy the PCP. As a password only maps to a single composition, the sum of search space sizes for each composition is the size of the overall password search space. For each composition, we take the following steps to calculate its search space size:

We start by calculating the number of combinations of characters from the charsets that make up the composition:

$$\prod_i \text{charset_size}_i^{\text{composition}_i} \quad (1)$$

We then multiply this value by the number of unique permutations in the composition:

$$\frac{(\sum_i \text{composition}_i)!}{\prod_i (\text{composition}_i)!} \quad (2)$$

If there are no additional requirements to be considered, this value is used as the composition’s search space size. If there are additional requirements, we will reduce this calculated by value by the number of passwords removed by each requirement.

First, we consider the *required_locations* requirement. If used, we recalculate our baseline using the same calculations above, except that we reduce the permutation calculation to only consider character classes not at fixed positions due to *required_locations*.

For the remaining four requirements, we take an approach wherein we create one or more invalid compositions that violate the requirement, calculate the search space for the invalid composition, and subtract the invalid composition’s search space size from the overall composition’s search space size (calculated above). We continue doing so until there are no more requirements to handle. We generate these invalid compositions as follows:

- For *max_consecutive*, we identify all charsets which have enough occurrences in the composition to violate this rule. For each of these charsets, we create a new, invalid composition that removes (*max_consecutive* + 1) occurrences from matched charset and adds a single occurrence of a new charset of size equal to the matched charset (representing the repeated character).
- For *max_consecutive* in *charset_restrictions*, we do much the same as above, except that the size of the new charset in the invalid compositions will equal *matched_charset_size*^{*max_consecutive*+1}, representing all possible combinations of the charset.
- For each substring in *prohibited_substrings*, we create a new, invalid composition that removes the appropriate charset for each character in the substring. We then append a charset of size 1 to the composition, representing the prohibited string.
- For each location in *prohibited_location*, we do not modify the current composition but instead calculate its search space as if the prohibited location were required.

B.2 Estimating Human-Generation

Prior research has shown that when generating passwords, humans prefer characters from specific character classes,

though this preference can differ based on country [18, 36]. Our PCP strength estimation can be parameterized based on what character classes users prefer to represent this behavior. For example, American users’ preferences might be lowercase, uppercase, then digits [18]. For Chinese users, their preferences are more likely to be digits, lowercase, then uppercase [18, 36].

We handle these preferences in Step 2 of our calculations. We initially execute step two as described up through calculating the list of lists representing characters that can occur in the remaining spots of the initially calculated password composition. For each sublist of charsets, we check to see if any of those charsets appears in the list of preferred charsets. If one or more do, we replace the sublist with a new list with a single element matching the highest-ranked matching charsets. After this modification, calculations proceed as described.

Note, these preference-based calculations are Fermi approximations, underestimating character class diversity in user passwords and overestimating diversity of character selection within a character class, with the two errors hopefully canceling out. Even though these are not exact estimates for human-generated passwords, they are sufficient to help administrators and researchers estimate the overall strengths and weaknesses of a PCP.

B.3 Limitations

For PCPs that do not use any of the final four requirements discussed in Step 3, our method precisely calculates the PCP’s search space. Our calculation is also correct if only a single one of these requirements are used for a composition. Of the 270 PCPs in our dataset, 260 do not use any of the five requirements, and of the ten that do, each uses only a single requirements. This means that calculations used in our analysis are all precise, and it suggests that most PCPs will have their search space calculated precisely.

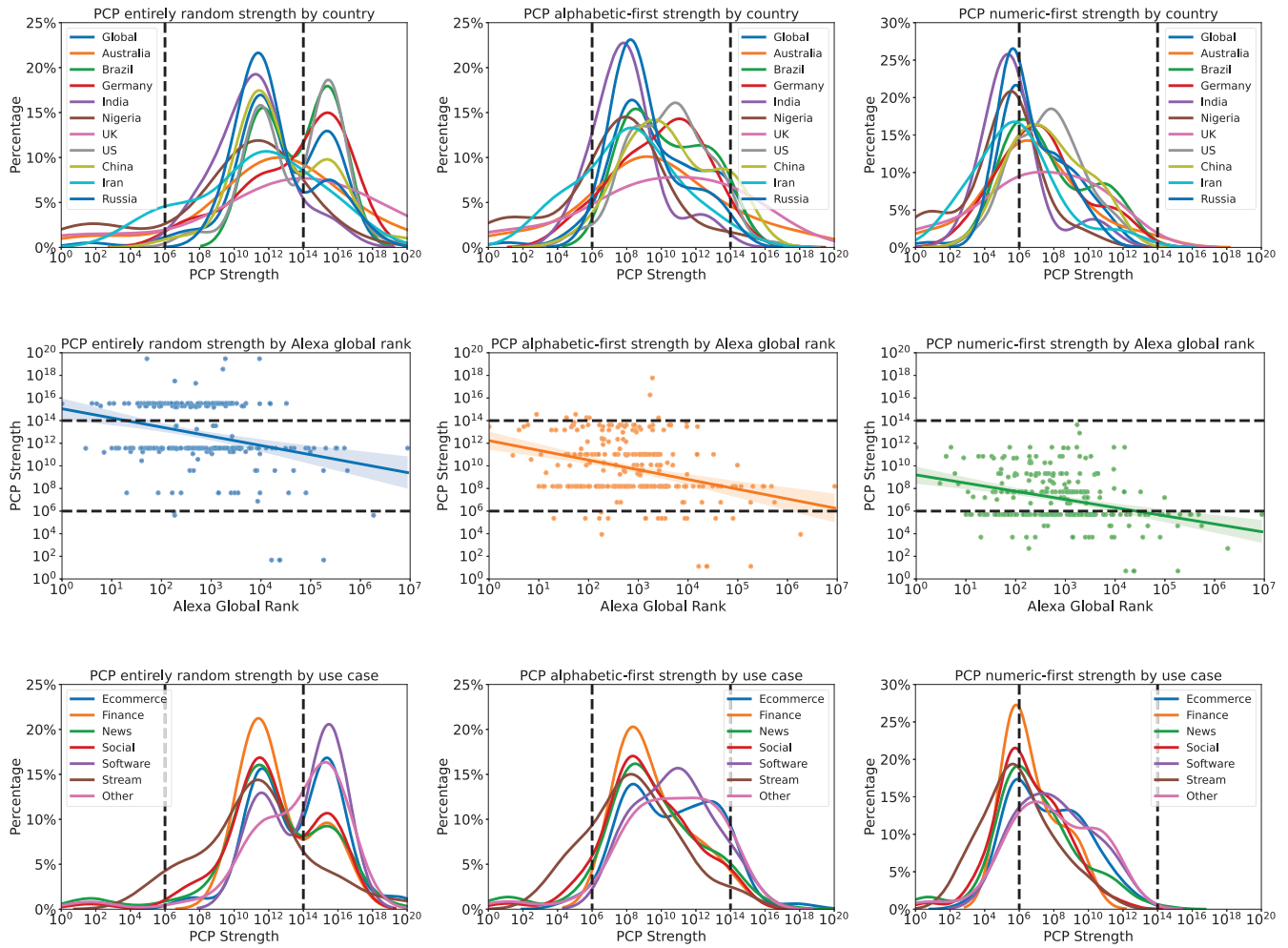
Still, more complicated PCPs that use multiple of the five requirements could have their search spaces underestimated. This occurs because these requirements have the possibility of removing the same passwords. To our knowledge, the only way to prevent this would be to enumerate the password combinations and permutations—as we did with compositions—but this is not tractable for any meaningful length of passwords. However, as the reduction to the search space for each of these requirements will generally be small compared to the overall size of the composition’s search space, we believe that the underestimates should be minimal. Additionally, in terms of strength estimates, underestimates are safer than overestimates.

C Webpages Accessible with HTTP

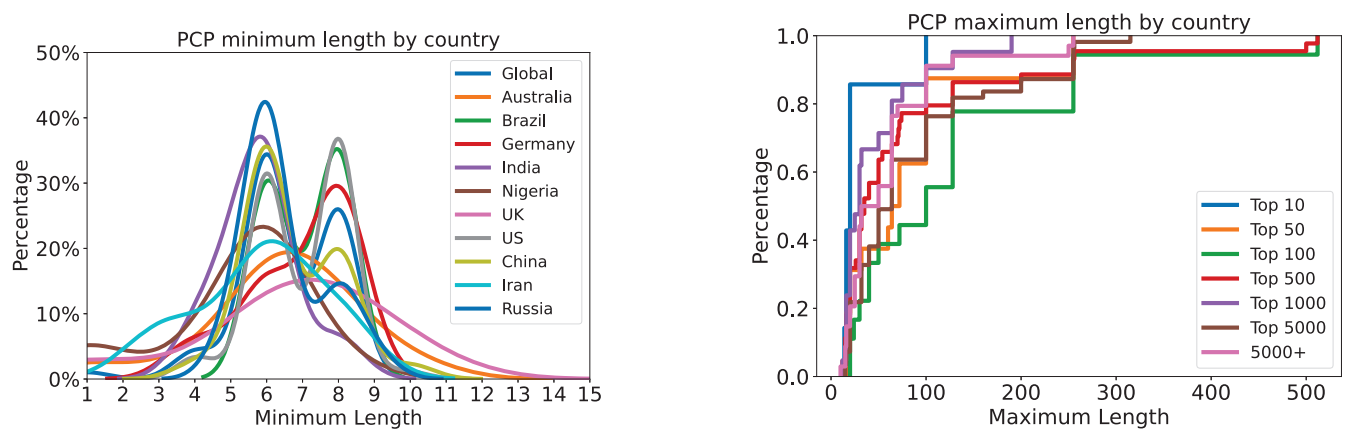
The following is the list of website for which we were able to access the account creation or login page using HTTP:

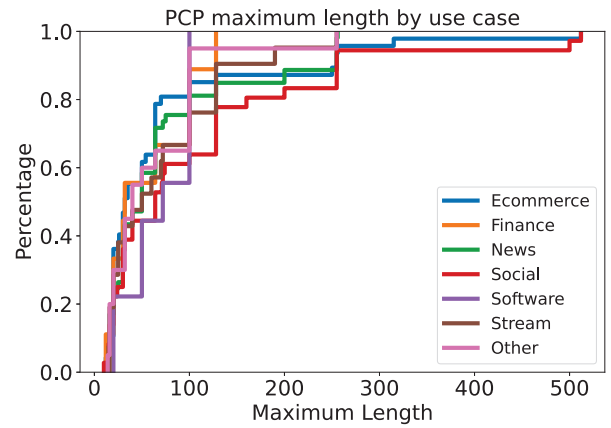
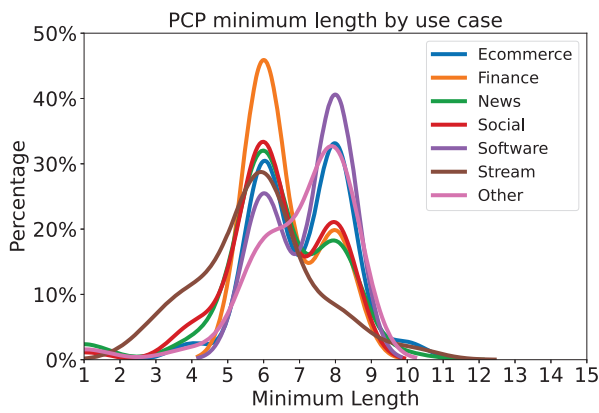
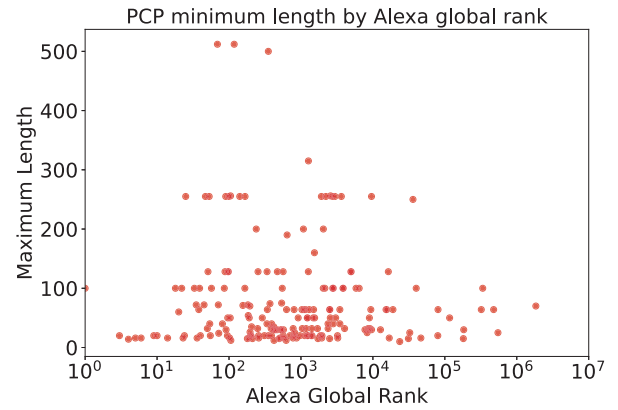
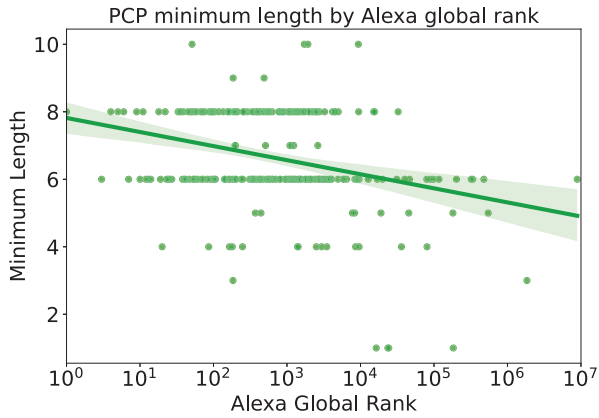
Website	Country	Popularity	Category
weibo.com	China	Top 50	Social
babytree.com	China	Top 100	Social
usatoday.com	US	Top 500	News
yaplakal.com	Russia	Top 5000	Social
ig.com.br	Brazil	Top 5000	Social
wikidot.com	China	Top 5000	Other
fb.ru	Russia	Top 5000	News
javlibrary.com	China	5000+	Stream
dwnews.com	China	5000+	News
metacafe.com	India	5000+	Social
eskimi.com	Nigeria	5000+	Social
ci123.com	China	5000+	Stream
sinovision.net	China	5000+	News
sugardaddyforme.com	China	5000+	Social
mydiba.xyz	Iran	5000+	Stream

D PCP Strength By Category



E PCP Features by Category





Password policies of most top websites fail to follow best practices

Kevin Lee

Sten Sjöberg

Arvind Narayanan

*Department of Computer Science and Center for Information Technology Policy
Princeton University*

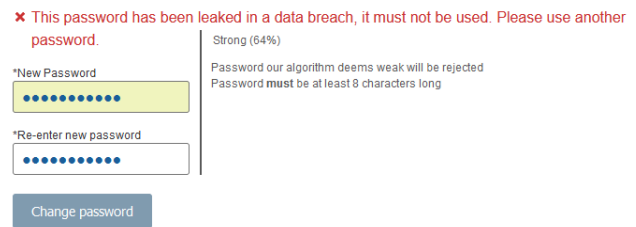
Abstract

We examined the policies of 120 of the most popular websites for when a user creates a new password for their account. Despite well-established advice that has emerged from the research community, we found that only 13% of websites followed all relevant best practices in their password policies. Specifically, 75% of websites do not stop users from choosing the most common passwords—like “abc123456” and “P@\$\$w0rd”, while 45% burden users by requiring specific character classes in their passwords for minimal security benefit. We found low adoption of password strength meters—a widely touted intervention to encourage stronger passwords, appearing on only 19% of websites. Even among those sites, we found nearly half misusing them to steer users to include certain character classes, and not for their intended purpose of encouraging freely-constructed strong passwords.

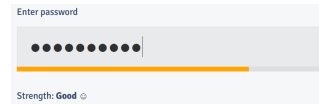
1 Introduction

Passwords remain the most common means of authentication on the web, despite their shortcomings. According to industry estimates, close to half of data breaches involved authentication failures [13, 14]. As such, the need to use strong passwords remains unchanged [15]. To encourage this, websites mainly use three types of interventions during password creation: blocklists, password composition rules / policies (PCPs), and strength meters (Fig. 1). All three interventions have been extensively researched in the information security community.

Prior research has generally concluded that blocklists and strength meters—when configured correctly—lead users to create stronger passwords without significantly burdening them [3, 6, 16]. However, PCPs that require specific character-



(a) A website preventing us from using a password (“passer2009”) that was leaked in a data breach.



(b) An example of a password strength meter. Its colored bar and text feedback changes in response to the entered password.



(c) A 3class8 character-class PCP, which requires passwords be at least 8 characters in length with at least 1 lowercase, 1 uppercase, and 1 number.

Figure 1: Examples of the three interventions we studied: blocklists, PCPs, and password strength meters.

classes (i.e., lowercase, uppercase, digits, and symbols) are not recommended. That’s because users fulfill requirements in predictable ways like capitalizing the first letter or placing a “!” at the end, negating the putative security benefits [17–19]. Additionally, character-class PCPs have consistently received poor usability ratings; in those same studies, users needed more attempts to create a compliant password and had difficulty recalling the password. Instead, websites should set only a minimum-length requirement while complementing it with a blocklist check or minimum-strength requirement [3].

The research is clear; what is less clear is whether these best practices are actually being followed. There has been no comprehensive study to understand how online services guide their users in setting up passwords (although previous studies have looked at narrow aspects of this question [5, 20]). We aimed to fill this gap by examining password policies of 120

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

	Best practices from prior research	Our key findings
Blocklists (§ 3)	<ul style="list-style-type: none"> Do check users' passwords against lists of leaked and easily-guessed passwords [1–4]. Do reject the password if it appears on a blocklist, prompt the user to select a different password [1, 4]. 	<ul style="list-style-type: none"> More than half (71 / 120) of websites do not check passwords at all, allowing all 40 of the most common passwords we tested (e.g., “12345678”, “rockyou”). 19 more websites block less than half of the most common passwords we tested.
Strength meters and min-strength reqs (§ 4)	<ul style="list-style-type: none"> Do provide real-time password strength estimates [5–7]. Do set minimum-strength requirements by estimating guessability (the number of guesses it would take for an adversary to crack the password) [3, 8–11]. 	<ul style="list-style-type: none"> Only 23 / 120 websites used password strength meters. Of those 23, 10 websites misuse meters as nudges toward character-class PCPs and do not incorporate any notion of guessability.
Composition policies (§ 5, § 6)	<ul style="list-style-type: none"> Do not require specific character-classes; let users freely construct passwords [2, 3, 7, 12]. NIST: Do set a minimum-length of at least 8 characters [1]. 	<ul style="list-style-type: none"> 54 / 120 sites still use character-class PCPs. We devised a new method to measure the security and usability of all 120 PCPs. Based on our method, we found that all PCPs performed poorly, none provided $\geq 60\%$ security and usability simultaneously.

Table 1: We contrast our key findings with established best practices for encouraging strong passwords.

of the most popular English-language websites in the world. By signing up for accounts and manually testing requirements for password creation, we discovered each website's blocklist strategy, PCP, and strength meter implementation (if any). We asked the following research questions:

1. Are websites preventing users from using the most common passwords? (§ 3)
2. Are websites using password strength meters to encourage strong passwords? (§ 4)
3. What PCPs are used by top websites? What are the security-usability tradeoffs of those PCPs? (§ 5, § 6)

We considered a website to be following best practices if it simultaneously satisfied the following security and usability criteria:

- **Security:**

- Allowed 5 or fewer of the 40 most common leaked passwords and easiest-to-guess passwords (e.g., “12345678”, “rockyou”) we tried.
- Required passwords be no shorter than 8 characters OR employed a password strength meter that accurately measured a password's resistance to being guessed by an adversary [7].

- **Usability:** Did not impose any character-class requirements.

We found that only 15 websites were following best practices. The remaining 105 / 120 either failed to adhere or explicitly flouted those recommendations in their policies,

leaving users at risk for password compromise or frustrated from being unable to use a sufficiently strong password. We compare our key findings with the best practices for all three interventions in Table 1.

We further devised a method to measure the security and usability of PCPs using a large corpus of breached passwords. Past studies have typically examined a small number of different PCPs due to constraints with hiring participants, which motivated us to design a method that could scale to the large number of PCPs we examined. These studies have also systematically neglected to investigate PCPs with short minimum-length requirements, which we frequently found during our study (the following paragraph suggests a reason why previous studies may have excluded these PCPs). While we were able to analyze the PCPs of all 120 websites we visited, we note that our strategy has limitations and should be used to complement findings from previous user studies. We found that no PCP had more than 60% security and usability simultaneously. These results further call into question some of the recommendations on PCPs that have been taken at face value, without any evidence.

While there is broad consensus on best practices in the prior literature, it is sometimes unclear exactly where to draw the line. For instance, the National Institute of Standards and Technology (NIST) recommends an 8-character minimum-length requirement in its current version of *Digital Identity Guidelines*—a widely relied-upon resource by both practitioners and researchers [1]. Yet, that recommendation does not cite any research. Even though we performed a thorough literature search and failed to find any research that had investigated the usefulness of setting an 8-character minimum-length, we

decided to count that recommendation as a best practice. Here, we have used our best judgment in defining what constitutes best practice, erring on the side of being lenient. While our exact number might change if we change our definition of “best practices”, our qualitative finding—that most websites are not following best practices—does not change.

Our findings reveal a disconnect between industry and the research community. Passwords have been heavily researched, yet few websites have implemented password policies that reflect the lessons learned. Researchers should make sure their findings have societal impact by engaging in outreach to website operators about their password practices.

2 Overview of password best practices

Websites mainly have three different ways to encourage users to create more secure passwords, as outlined by NIST [1]. Here we discuss previous research on the methods and their best practices.

2.1 Blocklists work, but need to be carefully configured

One simple way for websites to encourage more secure passwords is to keep a list of common insecure passwords (e.g., “123456”, “!QAZ!qaz”) and deny users from choosing passwords from that list (Fig. 1a). Prior research has found that password blocklists work. Kelley et. al (2012) created a blocklist of five billion passwords returned by a cracking algorithm created by Weir et. al (2009), and tested it in a subsequent user study [2, 21]. They found that users created passwords that were significantly harder-to-guess, compared to passwords created under four other widely-used but smaller blocklists. Shay et al. (2015) found that blocklists generally increase security without sacrificing password recall among users [7]. Habib et al. (2017) also supported using blocklists, and further recommended that websites also restrict users from submitting simple modifications to blocklisted passwords [16].

Blocklists may consist of common passwords gathered through different strategies, including commonly used passwords that have been exposed in data breaches and passwords that are likely to be guessed easily by password cracking tools. Websites may also have different approaches to checking passwords against the blocklist; for instance, some may perform exact matching while others strip out symbols before matching. While NIST recommends that websites block common passwords, it is neither prescriptive on which lists to use nor on the comparison method [1].

The National Cyber Security Centre (NCSC) provides more concrete guidance [4]. In collaboration with *Have I Been Pwned?* (HIBP)—an online service that allows users to check whether their credentials have been compromised in data breaches—the NCSC has released a list of the 100,000 most common passwords for websites to use as a blocklist (which we refer to as NCSC-HIBP-100k later on in the paper). NCSC guidance reasons that blocking the top 100,000 passwords

prevents users from “making poor password choices, whilst not making it too difficult for them to choose one.”

Tan et al. (2020) later investigated the security-usability tradeoffs of blocklist requirements and found that blocklists—while effective—can cause user frustration if not properly configured [3]. They recommend blocking passwords that appear in NCSC-HIBP-100k or blocking common passwords that appear in a corpus of 10 million leaked passwords, both of which we used in our experiments.

In this study, we empirically examine whether websites follow the best practices for blocklists established by prior work.

Compromised credential checking. In addition to blocking the most common passwords, some websites may employ compromised credential checking, which checks whether a username-password pair has been exposed in a previous breach [22, 23]. Websites can more accurately measure the risk of account compromise to the user because they additionally consider whether her full login credentials are already available to cybercriminals.

We did not test for compromised credential checking by websites in our study because it presents practical difficulties. Using other people’s compromised credentials raises ethical concerns, whereas using our own compromised credentials may be unreliable due to the small sample size. While not listed by NIST as a best practice, some websites in our study may check for compromised credentials due to its known effectiveness.

2.2 Min-strength requirements and strength meters are both effective and user-friendly

A newer approach to encourage strong passwords is to set minimum-strength requirements. When a user submits a candidate password, the website estimates the strength for the submission, and if it is greater than the minimum threshold, the candidate password is accepted. A strength meter that updates in real-time is often shown to nudge users as they craft their passwords (Fig. 1b).

To measure strength, researchers recommend and typically use adversarial guessing—the number of guesses needed to crack a password (i.e., the guess number or guessability). Previously, strength was often modeled using Shannon entropy—a function of the length and number of character-classes present in a password, or its complexity. However, complexity has since been deprecated since it is not a good proxy for guessability (see Appendix F for further background).

Estimating password strength is difficult, especially considering that users expect near-instantaneous feedback when setting a password. Previous research has found that among the password-strength meters in use on the web, most actually measured complexity instead of guessability, and were actually inconsistent with one another (de Carnavalet and Mannan, 2014) [5]. There was an open-source implementation that were found to be reliable, however: `zxcvbn` outputs

accurate strength estimates through 10^6 guesses, the threshold for online attacks [24].

In 2017, Ur et al. designed a data-driven strength meter that estimates password strength using a client-side neural-network created in a prior study (Melicher et al., 2016) [10]. Their meter received positive feedback from participants in the following user study, and was accurate when compared with results from password cracking tools, which were used as ground truth [6]. Tan et al. (2020) later updated the meter to enforce blocklists and minimum-strength requirements, while also making the meter freely available to use [3]. They concluded that minimum-strength requirements are the best way to encourage strong passwords, and recommend setting the minimum-strength threshold to at least 10^6 to prevent online guessing attacks [3]. Since their password-strength meter directly estimates guessability—as opposed to PCPs indirectly using complexity as a measure of strength—websites need only set a minimum-guesses threshold instead of character-class requirements, such as 10^6 for online attacks and 10^{14} for offline attacks. They further highlight that the meter’s underlying neural network can be “easily retrained to reflect changing patterns in passwords over time” and that its configurable integration with blocklists can penalize common passwords.

2.3 Character-class PCPs should not be used

To enforce the use of strong passwords, websites have employed password composition policies (PCPs). PCPs are rules which users must follow in creating their passwords. These rules most often include a minimum password length requirement along with character-class requirements (Fig. 1c). PCPs fall into two categories: ones with character-class requirements (which we’ll refer to as “character-class PCPs” throughout the paper) and ones without (PCPs that only have a minimum length requirement, which we’ll refer to as “minimum-length PCPs”).

As a vestige of when password strength was modeled by Shannon entropy, character-class PCPs force users to create complex passwords.¹ While prior empirical research has found that passwords containing multiple character classes were generally more resilient to password-guessing attacks, employing character-class PCPs is a hardly ideal solution (Komanhuri et al., 2011), (Kelley et al., 2012) [2, 12]. Character-class PCPs have poor usability; users have found it difficult to comply with the complex rules and to remember the password they have created. Moreover, character-class PCPs do not account for a significant subset of users, who respond predictably to comply with character-class requirements (e.g., adding numbers at the end, capitalizing the first letter). These behaviors reduce the benefits of adding complexity (Shay et

¹Modeling password strength with Shannon entropy is different from using guess numbers. Fig. 4 in § 6 illustrates this; character-class PCPs not only reject most weak passwords, but most strong passwords as well. See Appendix F for further discussion on the differences.

al., 2010), (Weir et al., 2010), (Ur et al., 2015) [18, 19, 25]

As studies that have found that increasing minimum-length requirements while reducing character-class requirements can lead to strong passwords without decreased memorability, NIST has also updated its guidance to recommend websites remove character-class requirements (Kelley et al., 2012), (Shay et al., 2014) [2, 17]. It further recommended that websites require passwords be at least 8 characters long [1]. Tan et al. (2020) actually found that character-class PCPs do not make it harder for attackers using modern-day cracking tools anymore, since users now tend to incorporate multiple character-classes of their own accord. Still, they recommend against using character-class PCPs because users still find them annoying and some users will still fulfill requirements in predictable ways [3].

Even with the updated recommendations, character-class PCPs may remain ubiquitous, though they are largely unmeasured; the only previous study that explored PCPs on the web was from 2010 [20]. In our study, we measured the state of security and usability of PCPs present on the web by extracting them from websites we visited.

3 Study 1: password blocking

We measured whether popular websites prevent users from choosing the most common insecure passwords and found most of them insufficiently block users’ choices. We selected common passwords to test based on two different strategies: blocking the 100,000 most frequently-used passwords found in password breaches (NCSC-HIBP-100k) and blocking passwords guessed early on by state-of-the-art cracking tools.

3.1 Method

3.1.1 We tested 120 of the top websites

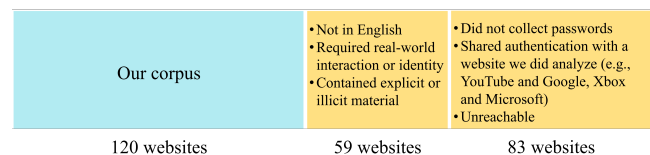


Figure 2: A breakdown of the 262 websites we attempted to study. We skipped 59 websites that did not fit our selection criteria. 83 websites either could not be analyzed or were already represented among our corpus of 120 websites.

In this study, we are concerned with password policies at the most popular English-language websites so our findings could be verified by all co-authors (who are all fluent in English). We focused on popular websites because previous research has shown that they generally have better security policies [20], which means that our results can be seen as an underestimate of conformity with best practices. Further, we wanted to hold these specific websites to account because they affect more users. Using an actively maintained ranked

list provided by other researchers [26],² we tested the top 120 websites that were accessible to us. We skipped some websites for the reasons shown in Fig. 2; we reached our total of 120 websites after trying the top 262 listed entries.

Before the tests, we extracted the PCP on each website and encoded them in a regular expression (detailed in § 5.1). This allowed us to select PCP-compliant passwords for testing.

3.1.2 Testing common passwords leaked in breaches

We sampled 20 passwords from the NCSC-HIBP-100k list, which was ordered from most common to least common. We started by removing passwords that did not fit the website’s PCP (with the aforementioned regular expressions) and sampling candidates to test at each website. In order to evenly represent the most frequently-leaked passwords along with the long tail of rest of the passwords in the list, we used a stratified sample based on powers-of-10 (1-10, 11-100, 101-1,000, 1,001-10,000, and 10,001-100,000). We randomly sampled candidates weighted by their position on the list ($\frac{1}{\text{position}}$), which gave us—in expectation—four passwords in each stratum. In order to ensure fair comparisons, websites with identical PCPs were tested with the same 20 passwords (e.g., all websites with a *lclass8* PCP were tested with “babygirl23”, “lifeisgood”, etc.) We refer to these tested passwords as *leaked* passwords hereinafter.

Using the accounts we had set up, we attempted to change our password to each of the *leaked* passwords. If the change was successful, we logged out and logged back in with the new password to confirm, then noted that the password was accepted.

3.1.3 Testing common easy-to-guess passwords

In addition to restricting *leaked* passwords, websites should discourage users from selecting common passwords that are easily guessed (e.g., block “Blink182”, which can be guessed in ~ 9 tries, or “Hello123”, which can be guessed in ~ 316 tries). Here we tested the first 20 passwords that were guessed by state-of-the-art cracking tools at each website. We refer to these tested passwords as *easiest-guessed* passwords hereinafter.

We used Password Guessability Service (PGS)—offered by the Passwords Research Team at Carnegie Mellon University—to find these passwords to test [27]. PGS simulates a real attacker guessing passwords; it leverages multiple (5, at the time of our study) cracking tools to arrive at the user-provided plaintext password, returning the guess number (i.e., the number of guesses needed to find the password) as the password’s strength rating. PGS also offers the `min_auto` configuration, which returns the minimum guess number for each password across all 5 tools. Previous research has found that the `min_auto` approach provides a conservative estimate for the performance of an unconstrained professional attacker [27]. Therefore, we referred to the `min_auto` guess

number for all of the passwords in this study.

Since PGS requires its users to provide passwords in plaintext in order to receive results, we selected passwords to use from the Xato 10-million password dataset [28]. To the best of our knowledge, the dataset—which we will refer to as the Xato passwords hereinafter—represents the largest and most recent corpus of real-world passwords accessible to academic researchers, and has been widely used in previous work [3, 6, 16, 24]. We did search for newer password dumps to complement the Xato passwords, but found they were either available only on the dark web or offered in hashes rather than plaintext (to prevent large-scale cracking) [29, 30].

With all of the Xato passwords rated, we used the 20 passwords with the lowest guess numbers as our *easiest-guessed* passwords, and tested whether websites allowed them to be used. As with our testing of *leaked* passwords, we only selected passwords that fit the website’s PCP. We excluded passwords that were already in the *leaked* passwords, and selected the password with the next-lowest guess number instead. Here, we also tested the same 20 passwords across websites with identical PCPs to ensure fairness (e.g., all websites with a *DigSym6* PCP—6+ characters with 1 digit or symbol—were tested with “jordan23”, “jessical”, etc.). Every selected password could be guessed within $10^{4.9}$ guesses, well within the threshold of online guessing attacks.

3.2 Results

1. **Most websites do not block *leaked* or *easiest-guessed* passwords at all.** 71 / 120 websites accepted all 40 passwords we tested. By allowing both *leaked* and *easiest-guessed* passwords, these websites put their users at risk of password compromise and subsequent account hijackings. Additionally, accounts at other websites may be at risk for compromise too; users often practice poor security hygiene by reusing their passwords across the web, so this misconception that their password was not blocked and therefore suitable can have widespread insecurity.

These 71 websites span different industries, including e-commerce (Amazon), social media (TikTok), entertainment (Netflix), and news (Wall Street Journal). Amazon, for instance, allowed us to change our password to “123456”, the most common password on the web. TikTok—despite requiring users to choose a *3class8* password—allowed us to use “p@ssw0rd” (guessed by PGS in 7 tries, the fourth-most common *3class8* password) on our account.

2. **Additionally, several websites had insufficient blocking.** In addition to the 71 websites which accepted all 40 passwords, 19 sites accepted more than half of the *leaked* or *easiest-guessed* passwords tested. In some of these cases, this was likely due to insufficient blocklists. For example, IBM seemed to only block choices containing the word “password”, which only blocked 1

²Available at <https://tranco-list.eu/list/VJ5N>. Generated on 29 July 2021.

(“Password1”) of the 40 passwords we tested on its site. Samsung only blocked number sequences (e.g., “123”), and Salesforce only blocked “password”. While this may prevent users from using the most guessable passwords, the majority of the most common passwords still get accepted.

3. **10 websites seemed to be using a shorter *leaked* passwords blacklist.** We found 10 websites that blocked most of the tested *leaked* passwords from the higher-rank strata (e.g., 1-10, 11-100) but then allowed a majority of leaked passwords from the lower-rank strata (e.g., 10001-100000). This could indicate that these websites are using a truncated version of the NCSC-HIBP-100k list to check passwords, sacrificing security for usability. Spotify, for instance, blocked all *leaked* passwords up to the 101-1000 stratum but allowed all passwords beyond that point, which suggests that it only checks for the top 1000 *leaked* passwords. Our finding here is tentative, however, since we assumed websites were using the NCSC-HIBP-100k list. Table 5a in Appendix D shows the breakdown by stratum for the 10 websites.
4. **7 websites blocked *easiest-guessed* passwords, but not *leaked* passwords.** 7 websites disproportionately accepted more *leaked* passwords than *easiest-guessed* ones (Microsoft: 14 *leaked* accepted / 0 *easiest-guessed* accepted, LinkedIn: 14 / 0, WeTransfer: 19 / 4, Roblox: 18 / 6, Reddit: 16 / 4, Twitter: 12 / 2, and Indeed: 9 / 1).

These 7 websites might have been using a minimum-strength requirement instead, since passwords they accepted generally had higher guess numbers than the passwords they rejected. If true, none of the websites set their minimum-strength requirement to prevent the threat of online guessing attacks (10^6 guesses). For example, Microsoft accepted one *leaked* password cracked with 251 guesses, and WeTransfer allowed one *leaked* password cracked with 32 guesses. Table 5b in Appendix D shows the minimum-strength cutoffs we found in our testing.

5. **Few websites prevented us from setting *leaked* and *easiest-guessed* passwords.** Only 15 websites—including Google, Adobe, Twitch, GitHub, and Grammarly—blocked all 40 passwords we tried. 7 more websites—including Apple, Canva, and VK—performed moderately well, allowing 5 or fewer tested passwords.
6. **Websites that allowed *leaked* and *easiest-guessed* passwords hold sensitive user information.** 38 of the 71 websites that allowed all 40 passwords store user payment information such as credit card or banking details, including Amazon, Netflix, GoDaddy, and Squarespace. 64 / 71 websites store PII about users, including Line, Intuit, Zoom, and MySpace. For each of the websites we analyzed, we checked whether it stored sensitive information using the test account we created earlier.

While some websites could be low-risk, the majority of websites (70 / 120) we studied collect payment information and thus are potentially high-risk.

Appendix G presents a risk categorization of these 120 websites based on PII and payment information collection, along with the number of accepted *leaked* and *easiest-guessed* passwords.

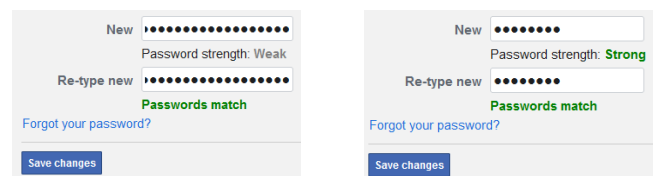
4 Study 2: strength meters

In 2014, de Carnavalet and Mannan investigated 11 password strength meters that were used in practice, and found most were estimating password complexity instead of guessability [5]. We wanted to know if there had been any changes; are meters at top sites now estimating guessability when a user chooses a password? To answer this, we reverse-engineered their patterns by testing different passwords.

4.1 Method

We considered all form elements on the password update page that updated in real-time to give feedback about the strength of the input on the password field. We then ran two tests on each of the strength meters to learn its patterns. First we investigated whether the meter was consistent in discouraging insecure choices; we tested the 100 *easiest-guessed* passwords from Xato that fit the website’s PCP and noted the feedback received on each password. Next we tried to reverse-engineer the mechanics of the meter through boundary testing. We tested passwords with different lengths and number of character-classes, as well as passwords that were not compliant with the website’s PCP. We selected passwords from Xato and also used passwords generated from password managers—Lastpass and 1Password—and noted movement patterns along the strength meter.

4.2 Results: most websites are not using strength meters to measure guessability



(a) “bkmmafwexucnvnsgppdk” (1 class, randomly generated) rated as Weak (1/3).
 (b) “Passw0rd” (3 classes) rated as Strong (3/3).

Figure 3: Despite having a *1class6* PCP, Facebook’s password strength meter is driven by adding more character-classes, and not password strength.

1. **Password strength meters are not widely used.** We found only 23 websites using password strength meters of any sort. Despite previous research touting the added

Password strength meters at websites with:	Our finding(s)	Implications on users	Prevalence
Minimum-length PCPs	<ul style="list-style-type: none"> • Encourages complex passwords over passwords that are harder-to-guess. Rates <i>easiest-guessed</i> passwords that have more character-classes higher than passwords with high guess numbers but fewer character-classes. • Discourages users' choices by nudging them toward fulfilling character-class PCPs. 	<ul style="list-style-type: none"> • Password strength feedback does not reflect password guessability. • Possible usability issues similar to when fulfilling character-class PCPs. 	6 / 18
Character-class PCPs	<ul style="list-style-type: none"> • Encourages more complexity than required. Meter tops out only if passwords include more character-classes than required by the PCP. 	<ul style="list-style-type: none"> • Password strength feedback does not reflect password guessability. • Potential usability issues when a candidate password meets all stated requirements but does not fill the meter. 	4 / 5

Table 2: Of the 23 websites that used password strength meters, 10 used those meters to encourage more complex passwords. 6 websites with minimum-length PCPs were actually using their meters as proxies for character-class PCPs.

security and usability benefits of using strength meters and robust open-source implementations like `zxcvbn`, most websites have not updated their password change procedures.

Regarding recommended strength meters, we found only two websites using `zxcvbn`; Dropbox (the organization behind the meter) and CPanel. The rest of the websites were using black box implementations that may not have been rigorously tested by the research community.

2. **10 / 23 websites misuse strength meters to measure complexity instead.** Rather than measuring password guessability, we found meters were actually being used as proxies for character-class PCPs. We break down our results by PCP here and in Table 2:

6 / 18 websites with minimum-length PCPs use strength meters as character-class PCP nudges. Their strength meters would only increase if a password had more character-classes than the one entered prior, and not if it had a higher guess number. Despite Facebook's *1class6* PCP, its 3-point strength meter—shown in Fig. 3—considered all-lowercase passwords weak; “zcdplgbtqldecfrzdqrw” (randomly generated) was considered Weak (1/3) while “Password1” was considered Strong (3/3). The strength meter at Yelp (*1class6* PCP) unconditionally considered 16-character passwords strong, while requiring shorter passwords contain all four character-classes to be considered as such; “123456789123456789” (guess number ~ 631) was considered Great (4/4) while “WzNGVE5uuWHd” (randomly generated) was considered only Good (3/4). Since these meters measured complexity instead of estimating guessability, their readings were not reflective of how diffi-

cult it would be for an adversary to crack the password. Furthermore, users are nudged into creating complex passwords at these sites. The other 12 websites with minimum-length PCPs—including Google, Yahoo, and Twitch—had strength meters that more closely corresponded with password guessability; they rated all 100 passwords we tested as weak (<50% on their respective meters), and we did not find any insecure patterns when testing passwords with different character-classes and length.

4 / 5 websites with character-class PCPs use their meters to encourage further complexity. They reserved the highest ratings on their meters for passwords that went beyond the required character-classes. Apple's strength meter, for instance, would only reach 100% if the password was 16-characters long and contained a symbol, despite its corresponding *3class8* PCP not requiring symbols. Aliexpress's 3-point meter only topped out if all 4 character-classes were included, despite a *2class6* PCP; “jmDy&!py\$Df&^tw*iBYy” (randomly generated 3-class) was rated Middle (2/3), yet “Abc123!@#” (guess number ~ 53) evaluated to High (3/3). Since users may already be led by these sites to believe that compliance with character-class requirements would automatically yield strong passwords, they may find it frustrating when their password does not top up the strength meter. We only found one website—ScienceDirect—which did not encourage further complexity, only because its meter already filled up completely upon PCP-compliance.

3. **12 / 23 websites were inconsistent between meter feedback and password acceptance.** We then raised

the question: is the feedback from the meter on the user-side consistent with the ultimate decision by the website to accept or reject a user’s chosen password on the server-side? Here, we used findings from our password blocking analysis—in which we had selected the 20 *easiest-guessed* Xato passwords that were compliant with a website’s PCP and tested whether the website would accept them (§ 3)—and compared them with feedback given by the website’s password strength meter. We now focused on feedback given by the website’s strength meter right before submitting each password to the server. For each password, if feedback from the strength meter was negative (i.e., <50% of the scale), we coded user-side feedback as “unacceptable,” otherwise, we coded the feedback as “acceptable.”³

12 / 23 websites had varying levels of inconsistency. 5 websites rated all 20 passwords as “unacceptable,” yet the server allowed all of them to be used; these websites rely solely on their strength meters, and do not perform additional checks before updating passwords. At CPanel, all 20 tested passwords were “unacceptable” (we found it was using `zxcvbn`), yet the server only rejected 13, which had all-letters or all-repeating-digits (e.g., rejected “66666666” but accepted “12345”).

Only 11 / 23 websites were consistent between their strength meter feedback on the user-side and acceptance on the server-side. 8 of those sites—including W3C, Tumblr, and TechCrunch—rated all 20 passwords as “unacceptable,” and all 20 were ultimately rejected. The other 3 sites were consistent in the opposite manner; they rated all 20 passwords “acceptable,” and all 20 were ultimately accepted. Overall, these inconsistencies can lead to insecurity stemming from users unknowingly setting *easiest-guessed* passwords, as well as frustration when a user is told a password is good enough but is rejected.

Our key finding is that despite the usefulness of password strength meters being established in the research literature, adoption has remained low, and 10 / 23 of the sites that have them—6 of which have minimum-length PCPs—actually misuse them as proxies for character-class PCPs.

5 Study 3: composition policies

5.1 Method

5.1.1 We extracted the PCPs on 120 of the top sites

Using the aforementioned Tranco list (§ 3.1), we visited the top 120 websites that were accessible to us. At each website, we created an account and subsequently navigated to the password change page to reverse-engineer the website’s PCP. We chose to use the password change page over the account

³Fortunately, we did not have to deal with any ambiguity between scale readings and labels on the meters we saw; all points below 50% had negative feedback, and all points 50% and above had neutral or positive feedback.

creation page in order to avoid the need to repeatedly create new accounts and enter sign-up information (e.g., usernames, email addresses, names).

We noted the static creation rules that loaded on the form, then extracted dynamic rules by varying the password input with sample strings we had prepared in advance. We varied our sample strings to include strings with 1 class only (all lowercase letters, all digits, etc.) and strings with multiple classes (uppercase, lowercase, digits, and symbols). For symbols (i.e., special characters), we limited our permutations to the 33 ASCII characters that could be typed on a standard U.S. keyboard (shown below; note presence of the space character):

```
!"#$%&'()*+,-./:;<=>?@[\\]^_`{|}~
```

We prioritized completeness in our method. For each PCP, we input and submitted non-compliant sample strings to make sure the website was enforcing its shown PCP. We also tested classes and characters that were not explicitly stated (e.g., for a hypothetical “include at least one number” character-class PCP we tested a sample string with only numbers to make sure there was no letter requirement, for a vague “include a special character such as `!#@()`” PCP we tested all 33 symbols and occasionally found websites that 1) counted other symbols towards the requirement, 2) allowed but did not count other symbols towards the requirement, or 3) disallowed other symbols entirely. The entire extraction task was done by hand and recorded in a spreadsheet (see Appendix C for a discussion of our attempts at an automated pipeline) by one of the co-authors and verified for correctness by a second co-author. After verification, we encoded the raw text of each website’s PCP into a regular expression (which was also verified by at least two co-authors). The regular expressions were later used for other analyses in our study. We ended up with 73 different regular expressions (hence, 73 distinct PCPs among the 120 websites).

5.2 Results: character-class PCPs are still widely used

Character-class requirement	Websites (N=54)
Lowercase letters	31 (57.4%)
Uppercase letters	30 (55.6%)
Letters (case-insensitive)	19 (35.2%)
Digits	53 (98.1%)
Symbols (special characters)	37 (68.5%)

Table 3: Character-class requirements on the 54 websites with character-class PCPs. Nearly all require that passwords include a digit.

Our findings are as follows:

1. **Character-class PCPs are still widely used.** 54 websites (45%) still require users to include specific character classes in their password, despite recommendations

against these requirements. As found in previous studies, character-class PCPs impose a huge usability cost for a minimal security benefit [17, 19]. Table 3 shows the breakdown of the required character-classes. Almost all character-class PCPs require a digit, and symbols were the second-most popular requirement.

2. **Websites with character-class PCPs are more likely to allow the most common insecure passwords.** Cross-referencing our findings from § 3.2, we found that 38 of the 54 websites (70.4%) with character-class PCPs accepted all 40 of the *leaked* and *easiest-guessed* passwords we tried, compared with 33 of the remaining 66 websites (50%) with minimum-length PCPs. This may suggest that websites believe that complexity requirements are sufficient in getting users to create strong passwords, so they do not need to check passwords on a case-by-case basis.
3. **The most common minimum-length requirement is now 8 characters.** In 2010, Bonneau and Preibusch found that 52% of websites studied were using 6-character minimum length—followed by 4 characters (14%) and 5 characters (10%)—and that less than 5% of websites studied had an 8-character minimum length [20]. In our results over a decade later, we found that 66 / 120 websites studied (55%) have an 8-character minimum length, followed by 6 characters (35 / 120) and 5 characters (7 / 120). Perhaps this is a result of updated guidance from NIST in 2017, which now recommends an 8-character minimum length for passwords, up from its previous recommendation of 6 characters [1, 31].
4. **9 websites had inconsistencies between the PCP and text shown.** 1 website mentioned only a minimum-length requirement, but we were unable to save our password unless it contained a digit or a symbol. 2 other websites similarly failed to mention an additional character-class requirement in their text, which we uncovered through our testing. On the flip side, 4 websites did not enforce all of the character-class requirements mentioned. For example, Canva seemed to require us to include “a mix of letters, numbers & symbols” in our password, but we found that there were actually no character-class requirements. 1 website mentioned that whitespace characters were not allowed, but still accepted our password containing it. Lastly, 1 website with a *2class8* PCP had no text at all. We were only able to reverse-engineer its PCP after opening up development tools on our browser to view the server responses and making multiple attempts with different character-class combinations. Overall, these inconsistencies can lead to a confusing user experience.

Our key finding is that character-class PCPs are still being used on 45% of popular websites, burdening users while

providing minimal security benefit. Even with the research against these complexity requirements, websites continue to force users to include extra characters like digits or symbols in their passwords, which some users may respond to in predictable ways. Furthermore, over 70% sites that continue to use character-class PCPs do not have any other password checks in place, allowing *leaked* and *easiest-guessed* passwords to be used.

We document several additional findings in Appendix E.

6 Study 4: PCP security and usability

In previous studies on PCPs, researchers typically conducted user studies by recruiting thousands of participants online (e.g., on Amazon Mechanical Turk) to perform password creation tasks on a testbed website. They would then analyze the passwords created for each PCP, such as measuring the complexity (entropy) of passwords created under each condition, the fraction of passwords guessed at a given guessing threshold, number of failed attempts, user sentiment, time taken to create a compliant password, and password recall rate [2, 12, 16]. These studies have influenced changes in password best practices over the past decade, particularly with the recommendation against character-class PCPs [1].

While it would certainly be useful to perform the same kind of password creation study for real-world PCPs, this was not feasible due to the large number of experimental conditions. As mentioned in § 5, we uncovered 73 unique PCPs among the 120 studied. For reference, in a previous user study, the authors recruited 5,099 participants who were assigned to 15 different PCPs; in order for us to replicate that power, we would have to recruit nearly 5 times as many participants [3]. These previous studies have also recognized the same limitations, and have kept the number of PCPs tested relatively small [2, 16, 17].

We therefore devised a different approach to measure the security and usability of PCPs studied. As we will show, our method has both advantages and limitations. Therefore our findings are tentative, and are ultimately intended to complement the findings from previous studies by providing insight into PCPs in practice.

6.1 Method

The fundamental insight of our method is to consider a PCP as a binary classifier, whose goal is to reject weak passwords and accept strong, hard-to-guess passwords. Here, we defined a password as weak if PGS could guess it in an online attack (within 10^6 guesses), and strong otherwise.

6.1.1 We assumed a corpus of passwords created without constraint

Users have different ways of generating passwords that are not influenced by a website’s PCP [32]. Some examples include:

- Using a fixed password for all websites (password reuse)

- Using a password manager to automatically generate passwords
- Using a fixed heuristic (e.g., dictionary word + digit)

For our analysis we needed a sample of these “unconstrained passwords” to make unbiased comparisons of security and usability across PCPs. Our sample used here is the Xato 10-million passwords set (56% / 44% split between strong and weak passwords) [28]. Even though it did not meet our requirement of passwords generated without constraint—because users were already subject to the PCPs of the breached websites in this set—we still used it for analysis. We discuss the implications of using the Xato passwords later on.

6.1.2 We used sensitivity as a proxy for security

We assumed that whenever a user sets a new password at each of the 120 websites we studied, they initially generate one using an unconstrained strategy. If allowed by the PCP, the user will then confirm and set the password. Any PCP will allow some fraction of weak passwords through, however, which is why we measured the sensitivity of the PCP. We consider sensitivity—the percentage of weak passwords rejected—as a proxy for security. A website that simply allows any password to be used (i.e., no PCP), for example, would have 0% sensitivity.

One advantage of using sensitivity is that it is unaffected by outliers. Some generated passwords may have extremely high guess numbers and thus skew the average strength of passwords accepted by a PCP, for example. Since we used PGS to obtain guess numbers for the Xato passwords, we also benefited from accurate password strength ratings, as opposed to using entropy [25]. Our method to measure security has one disadvantage, however. Unlike in an intervention study, we don’t know how users will react to any of the 120 PCPs, such as whether users go on to create strong passwords [12].

6.1.3 We used specificity as a proxy for usability

Some users—given their strategy for unconstrained password generation—will be frustrated by the PCP and be forced to pick a different strategy and password. While the usability cost would be justified if their password was actually weak, it would not be justified if it was strong. In the case of a password manager being incompatible with a PCP, they may be forced to pick a password manually, making it both weaker and less memorable (see § 2). Here, we used specificity to measure this usability cost. We used this measure as a proxy for usability of the PCP.

Specificity is an objective measure that complements other usability measures, like recall, user dropout, and time taken to enter a password. The main disadvantage to using specificity, however, is our inability to gauge user sentiment. That is, users may not necessarily feel frustrated by the PCP if their unconstrained password generation strategy is unsuccessful, especially if they have repeatedly encountered the same kind of PCP and have (predictable) adaptation strategies, or if their password manager accommodates them [33].

6.1.4 Limitations of using Xato passwords

Finally, we revisit the assumption about having a corpus of unconstrained passwords. Unfortunately, the Xato passwords set does not satisfy this requirement. While it is incredibly diverse—with weighted samples of over 1,000 password dumps collected over at least 5 years—most of the passwords were probably created by users reacting to some PCP [34]. One advantage to using the dataset, however, is that it doesn’t contain passwords that required cracking [34]. This means that there is no bias towards weaker passwords.

Ultimately, using the Xato passwords in our security / usability evaluation means that we will overestimate usability (e.g., the segment created under the same PCP as the one being tested will have 100% usability) and underestimate security (e.g., the segment created under the same PCP as the one being tested will have 0% security). We reiterate that our findings in this section should be regarded as tentative; yet the strong limitations of PCPs that they reveal call into question the usefulness of PCPs and call for further research using different corpora and/or methods. For instance, a future user study could ask users to create passwords under no constraint (i.e., “include at least 1 character”) and make that password set available for other researchers to use.

6.2 Results

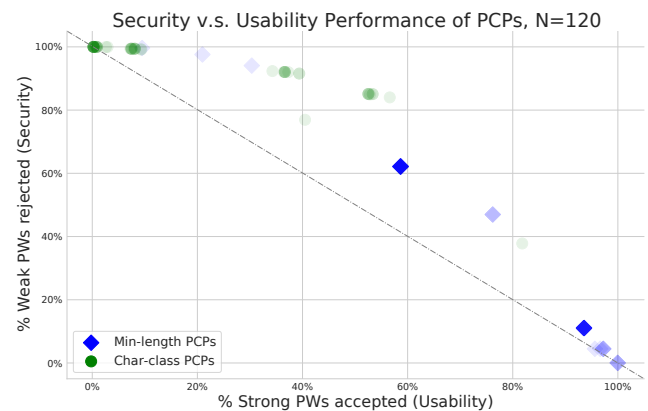


Figure 4: Scatter plot of security v.s. usability of PCPs for 120 websites. Each data point was plotted with 10% opacity, so more opaque areas reflect higher concentrations of PCPs with close scores.

Fig. 4 shows the scores of all websites we examined plotted along security and usability scores. Most of the 120 websites fall into one of three clusters: good security but poor usability (on the top left), good usability but poor security (bottom right), and average security and usability (in the middle of the graph). For comparison to a baseline, we also plotted a hypothetical PCP that rejects a random proportion α of passwords (and accepts $1 - \alpha$ of passwords), the diagonal line represents that PCP’s security and usability scores for $0 \leq \alpha \leq 1$. Our findings are as follows:

1. **No PCP simultaneously had more than 60% security and usability.** They either rejected too many strong passwords or accepted too many weak ones. Note that a hypothetical random PCP that blocks 50% of passwords has 50% security and usability simultaneously.
2. **PCPs fall on different parts of the security-usability spectrum.** Our results suggest a classic security-usability tension among PCPs. 69 / 120 websites we studied take opposite stances on the tradeoff; 33 have lenient policies (poor security cluster) and 36 have overly-stringent policies (poor usability cluster). Unsurprisingly, the PCPs within each of the 2 clusters are very similar to one another, with *1class6* being the majority PCP in the poor security cluster and *3class8* the majority in the poor usability cluster. We hypothesize that any PCP cannot be usable without allowing some weak passwords, and it cannot be secure without rejecting some strong password candidates.

1class8 policies make up most of the PCPs with middling acceptance rates, rejecting only 62% of weak passwords and accepting only 58% of strong passwords. While considered to be a best practice, our results suggest that the PCP alone is insufficient in preventing users from choosing weak passwords; websites need to have additional safeguards—such as blocklists—to filter out the remaining 38%. This was not the case at nine *1class8* websites, including SoundCloud, Eventbrite, and Trello; they allowed all of the *leaked* and *easiest-guessed* passwords we tried, like “1234qwer”, “1234567890”, and “babygirl123” (cross-referencing our findings from § 3).

3. **Most websites with insecure PCPs do not prevent insecure password choices.** 22 / 33 websites in the poor security cluster—including Amazon, Fox News, Etsy, and Dropbox (all with a *1class6* PCP)—do not block users from choosing passwords like “abc123” and “qwerty”—which we found with our password blocking analysis (§ 3)—and two more have insufficient blocking strategies (Slack and Yelp).

Our key finding is that PCPs are unsatisfactory in one or more ways. None of the 120 PCPs had more than 60% security and usability simultaneously. We hypothesize that there is no perfectly secure and usable PCP; all composition policies must make a tradeoff between user convenience (minimum-length PCPs) and strong passwords (character-class PCPs). Future studies should further investigate this hypothesis with different password corpora and methods. While websites with lenient PCPs can moderate the security gap with additional interventions like blocklists, we see this is not typically the case. A majority of these sites allow *leaked* and *easiest-guessed* passwords to be used.

7 Limitations

7.1 Limitations of analyzing the most popular websites

In these studies, we focused on the most popular websites. Since we did not additionally examine password policies of websites at the long tail (due to the work required to manually visit each website), we cannot be confident that our findings generalize to all websites. But note that previous research suggests that long-tail websites are likely to have even weaker security policies [20].

In Appendix B, we detail the access failures encountered at 142 websites in the ranked list we used. While future research can make an effort to study some of their password policies (like at government websites), we don’t believe their exclusion here affects our overall finding: most top websites are not following best practices in their password policies. Moreover, 83 / 142 excluded websites did not collect passwords, shared authentication with a website we already analyzed, or were unreachable (e.g., DNS, measurement links).

7.2 Limitations in the PCP security / usability analysis

In the PCP security / usability analysis, we rated the strength of all 10 million Xato passwords using PGS under no policy, which served as our ground truth. As PGS conservatively simulates an adversary cracking passwords, it also offers to configure guess number calculations under a particular PCP, since the adversary—who knows the website’s PCP—can constrain their search space to guess passwords more efficiently (the default option is no policy). Uploaded passwords that were compliant with the selected policy (17 options at the time of writing) would then be guessed with modified approaches using each of the cracking tools [27]. Since we did not select a policy to use, our results may lead to slight overreportings of both the fraction of strong passwords accepted and the fraction of weak passwords rejected for some of the PCPs. Obtaining more accurate ground truth measures would be challenging: PGS limits submissions to 30,000 passwords in order to ensure fair use of their free service, and their cracking tools can take a few weeks to complete.⁴

We did, however, further investigate the ramifications of using the no-policy guess numbers in our security / usability analysis, and found our main findings still hold true. We found that the *false positive risk*—the probability that a password rated strong was actually weak—was less than 4.56% at the 10⁶ guesses threshold we used, for all 120 PCPs.⁵ We randomly sampled 30,000 compliant passwords—weighted by their frequency—for each PCP and obtained their “PCP-aware” guess numbers from PGS in order to make the pairwise comparisons.

⁴The Passwords Research Team allowed us to submit all 10 million Xato passwords at once—for cracking under no policy—as a courtesy.

⁵Here we are concerned with the probability of a positive result being false [35]. This is different from the type 1 error rate (the false positive rate).

7.3 Limitations in scale

Our study required a significant amount of manual work to learn all of the password policies. For example, in our blocklist analysis alone, we attempted 4,800 password changes to determine whether websites allowed *leaked* and *easiest-guessed* passwords (~200 hours of work). Since we manually visited each website to reverse-engineer their password policy, we were only able to test 120 of the top English-language websites. We hence did not try to draw statistically valid conclusions about differences between industry sectors (e.g. news vs. social media websites) because of the small number of websites. We leave those topics (e.g., how the rates of compliance with best practices might vary by rank, geographic location, or sector) as future research directions.

We initially attempted two automated approaches which we ultimately abandoned due to concerns with completeness and data quality. We include our experiences in Appendix C to hopefully serve as useful notes for those who want to extend our work.

8 Other related work

Some previous empirical works have partially looked at password policies in practice. Bonneau and Preibusch (2010) extracted the PCPs of 150 websites across 3 different site categories: identity providers, content providers, and e-commerce sites [20]. They found that identity providers were significantly more likely to have minimum-length requirements (>1 character password), character-class requirements, and basic dictionary checks whenever a user changes their password. Overall, they found poor adoption of industry standards for password implementations, such as using TLS, CAPTCHA, and rate-limiting password guesses.

de Carnavalet et al. (2015) studied the password strength meters used at 11 popular websites [5]. They extracted or reverse-engineered the meter implementation at each site to local scripts and ran large-scale automated tests to get strength readings of known passwords, running a total of 53 million tests. They found most meters were only measuring password complexity, with only one implementation—`zxcvbn`—going beyond to penalize dictionary words. They also found that among the password strength meters in use on the web, most of them were inaccurate and inconsistent; passwords rated weak were often rated strong at other websites, and vice versa.

We built on the work done in both studies to deliver new additional insights. For password strength meters, we found websites with minimum-length PCPs that were using their strength meters as character-class nudges (§ 4). We also focused on investigating consistency between meter feedback at the client and password acceptance at the server by attempting to set the 20 *easiest-guessed* passwords we tested, and found more than half of websites were inconsistent. For PCPs, we resurveyed the landscape over a decade later, and found changes in the types of requirements used (§ 5). We also de-

veloped a new method to measure the security and usability of PCPs, and tentatively found none of them had decent security and usability simultaneously (§ 6).

9 Conclusion

Even with the gains in user authentication methods over the past two decades, passwords remain essential for online access, and replacing them in the near future seems improbable [36]. For these reasons, online services—especially the websites in which we found flaws—need to focus on password security and usability. Websites with insufficient blocklisting strategies, an outdated character-class PCP, or a misconfigured password strength meter should review the best practices summarized in Table 1 and make adjustments to their password policies. We further encourage them to review the research behind the guidelines in order to avoid misconfigured interventions that are inconsistent with one another (e.g., § 4).

We also suggest future research that directly engages with system administrators, in order to understand their mindset on password security. Researchers may then be able to uncover the reasons for the disconnect between industry and the academic community, and take steps towards reconciling the disparity. Some hypotheses include:

- Password policy is security theater: measures such as character-class PCPs, even if ineffective, may give users a false sense of security, and websites use them for this reason.
- Websites have shifted their attention to adopting other authentication technologies, such as multi-factor authentication (MFA), and believe that it is unnecessary to strengthen their password policies. (Note that there are severe weaknesses in SMS-based MFA, so this view might be overoptimistic [37, 38]).
- Websites need to pass security audits, and the firms who do these audits, such as Deloitte, recommend or mandate outdated practices.
- Websites face some other practical constraint that the academic community does not know about.

We have made our dataset available for other researchers at: <https://passwordpolicies.cs.princeton.edu/>.

Acknowledgements

We are grateful to Ryan Amos, Ben Kaiser, Malte Möser, and our anonymous reviewers for their helpful feedback on our writeup. We thank Arunesh Mathur and Prateek Mittal for discussions on research questions at the beginning of the study. We are also grateful to Ross Anderson, Richard Clayton, and other participants at the Cambridge security seminar for their valuable feedback.

References

- [1] Paul A. Grassi et al. *NIST Special Publication 800-63B Digital Identity Guidelines. Authentication and Lifecycle Management*. June 22, 2017. DOI: 10.6028/NIST.SP.800-63b.
- [2] Patrick Gage Kelley et al. “Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms”. In: *Proceedings of the 33rd IEEE Symposium on Security & Privacy (S&P)*. May 2012. DOI: 10.1109/SP.2012.38.
- [3] Joshua Tan et al. “Practical Recommendations for Stronger, More Usable Passwords Combining Minimum-Strength, Minimum-Length, and Blocklist Requirements”. In: *Proceedings of the 27th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Oct. 2020. DOI: 10.1145/3372297.3417882.
- [4] Dan U. *Passwords, passwords everywhere*. National Cyber Security Centre. Apr. 21, 2019. URL: <https://www.ncsc.gov.uk/blog-post/passwords-passwords-everywhere> (visited on 12/21/2021).
- [5] Xavier De Carné de Carnavalet and Mohammad Mannan. “From Very Weak to Very Strong: Analyzing Password-Strength Meters”. In: *Proceedings of the 21st Network and Distributed System Security Symposium (NDSS)*. Feb. 2014. DOI: 10.14722/ndss.2014.23268.
- [6] Blase Ur et al. “Design and Evaluation of a Data-Driven Password Meter”. In: *Proceedings of the 2017 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2017. DOI: 10.1145/3025453.3026050.
- [7] Richard Shay et al. “A Spoonful of Sugar? The Impact of Guidance and Feedback on Password-Creation Behavior”. In: *Proceedings of the 2015 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2015. DOI: 10.1145/2702123.2702586.
- [8] Joseph Bonneau. “The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords”. In: *Proceedings of the 33rd IEEE Symposium on Security & Privacy (S&P)*. May 2012. DOI: 10.1109/SP.2012.49.
- [9] Matteo Dell’Amico, Pietro Michiardi, and Yves Roudier. “Password Strength: An Empirical Analysis”. In: *2010 Proceedings IEEE INFOCOM*. Mar. 2010. DOI: 10.1109/INFOCOM.2010.5461951.
- [10] William Melicher et al. “Fast, Lean, and Accurate: Modeling Password Guessability Using Neural Networks”. In: *Proceedings of the 25th USENIX Security Symposium (USENIX Security)*. Aug. 2016. URL: <https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/melicher>.
- [11] Dinei Florêncio, Cormac Herley, and Paul C. van Oorschot. “An Administrator’s Guide to Internet Password Research”. In: *Proceedings of the 28th Large Installation System Administration Conference (LISA14)*. Nov. 2014. URL: <https://www.usenix.org/conference/lisa14/conference-program/presentation/florenccio>.
- [12] Saranga Komanduri et al. “Of Passwords and People: Measuring the Effect of Password-Composition Policies”. In: *Proceedings of the 2011 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2011. DOI: 10.1145/1978942.1979321.
- [13] Verizon DBIR Team. *2020 Data Breach Investigations Report*. May 19, 2020. URL: <https://www.verizon.com/business/resources/reports/dbir/2020/> (visited on 03/22/2022).
- [14] ESET. *ESET Threat Report T3 2021*. Feb. 9, 2022. URL: https://www.welivesecurity.com/wp-content/uploads/2022/02/ezet_threat_report_t32021.pdf (visited on 03/22/2022).
- [15] Steven Furnell. *Stop blaming people for choosing bad passwords – it’s time websites did more to help*. The Conversation. Jan. 3, 2022. URL: <https://theconversation.com/stop-blaming-people-for-choosing-bad-passwords-its-time-websites-did-more-to-help-172257> (visited on 01/26/2022).
- [16] Hana Habib et al. “Password Creation in the Presence of Blacklists”. In: *Proceedings of the 2017 Workshop on Usable Security (USEC)*. Feb. 2017. DOI: 10.14722/usec.2017.23043.
- [17] Richard Shay et al. “Can Long Passwords Be Secure and Usable?” In: *Proceedings of the 2014 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2014. DOI: 10.1145/2556288.2557377.
- [18] Richard Shay et al. “Encountering Stronger Password Requirements: User Attitudes and Behaviors”. In: *Proceedings of the 6th Symposium On Usable Privacy and Security (SOUPS)*. July 2010. DOI: 10.1145/1837110.1837113.
- [19] Blase Ur et al. “‘I Added ’!’ at the End to Make It Secure”: Observing Password Creation in the Lab”. In: *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS)*. July 2015. URL: <https://www.usenix.org/conference/soups2015/proceedings/presentation/ur>.

- [20] Joseph Bonneau and Sören Preibusch. “The password thicket: technical and market failures in human authentication on the web”. In: *The Ninth Workshop on the Economics of Information Security*. June 2010. URL: https://econinfosec.org/archive/weis2010/papers/session3/weis2010_bonneau.pdf.
- [21] Matt Weir et al. “Password Cracking Using Probabilistic Context-Free Grammars”. In: *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*. May 2009. DOI: 10.1109/SP.2009.8.
- [22] Kurt Thomas et al. “Protecting accounts from credential stuffing with password breach alerting”. In: *Proceedings of the 28th USENIX Security Symposium (USENIX Security)*. Aug. 2019. URL: <https://www.usenix.org/conference/usenixsecurity19/presentation/thomas>.
- [23] Bijeeta Pal et al. “Might I Get Pwned: A Second Generation Compromised Credential Checking Service”. In: *Proceedings of the 31st USENIX Security Symposium (USENIX Security)*. Aug. 2022. URL: <https://www.usenix.org/conference/usenixsecurity22/presentation/pal>.
- [24] Daniel Lowe Wheeler. “zxcvbn: Low-Budget Password Strength Estimation”. In: *Proceedings of the 25th USENIX Security Symposium (USENIX Security)*. Aug. 2016. URL: <https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/wheeler>.
- [25] Matt Weir et al. “Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords”. In: *Proceedings of the 17th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Oct. 2010. DOI: 10.1145/1866307.1866327.
- [26] Victor Le Pochat et al. “Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation”. In: *Proceedings of the 26th Network and Distributed System Security Symposium (NDSS)*. Feb. 2019. DOI: 10.14722/ndss.2019.23386.
- [27] Blase Ur et al. “Measuring Real-World Accuracies and Biases in Modeling Password Guessability”. In: *Proceedings of the 24th USENIX Security Symposium (USENIX Security)*. Aug. 2015. URL: <https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/ur>.
- [28] Mark Burnett. *Today I Am Releasing Ten Million Passwords*. Feb. 9, 2015. URL: <https://xato.net/today-i-am-releasing-ten-million-passwords-b6278bbe7495> (visited on 01/06/2022).
- [29] Troy Hunt. *Introducing 306 Million Freely Downloadable Pwned Passwords*. Aug. 3, 2017. URL: <https://www.troyhunt.com/introducing-306-million-freely-downloadable-pwned-passwords/> (visited on 12/22/2021).
- [30] Troy Hunt. *The 773 Million Record "Collection #1" Data Breach*. Jan. 17, 2019. URL: <https://www.troyhunt.com/the-773-million-record-collection-1-data-reach/> (visited on 12/22/2021).
- [31] William Burr et al. *NIST Special Publication 800-63-2 Electronic Authentication Guideline*. Aug. 2013. DOI: 10.6028/NIST.SP.800-63-2.
- [32] Blase Ur et al. “Do Users’ Perceptions of Password Security Match Reality?” In: *Proceedings of the 2016 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. May 2016. DOI: 10.1145/2858036.2858546.
- [33] Apple. *Password Rules Validation Tool*. URL: <https://developer.apple.com/password-rules/> (visited on 02/14/2022).
- [34] Mark Burnett. *Ten Million Passwords FAQ*. Feb. 10, 2015. URL: <https://xato.net/ten-million-passwords-faq-3b2752ed3b4c> (visited on 01/06/2022).
- [35] David Colquhoun. “The reproducibility of research and the misinterpretation of p-values”. In: *Royal Society Open Science* 4 (12 Dec. 2017). DOI: 10.1098/rsos.171085.
- [36] Joseph Bonneau et al. “The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes”. In: *Proceedings of the 33rd IEEE Symposium on Security & Privacy (S&P)*. May 2012. DOI: 10.1109/SP.2012.44.
- [37] Kevin Lee et al. “An Empirical Study of Wireless Carrier Authentication for SIM Swaps”. In: *Proceedings of the 16th Symposium On Usable Privacy and Security (SOUPS)*. Aug. 2020. URL: <https://www.usenix.org/conference/soups2020/presentation/lee>.
- [38] Kevin Lee and Arvind Narayanan. “Security and Privacy Risks of Number Recycling at Mobile Carriers in the United States”. In: *Proceedings of the 2021 APWG Symposium on Electronic Crime Research (eCrime)*. Dec. 2021. DOI: 10.1109/eCrime54498.2021.9738792.

A Visualization of best practices

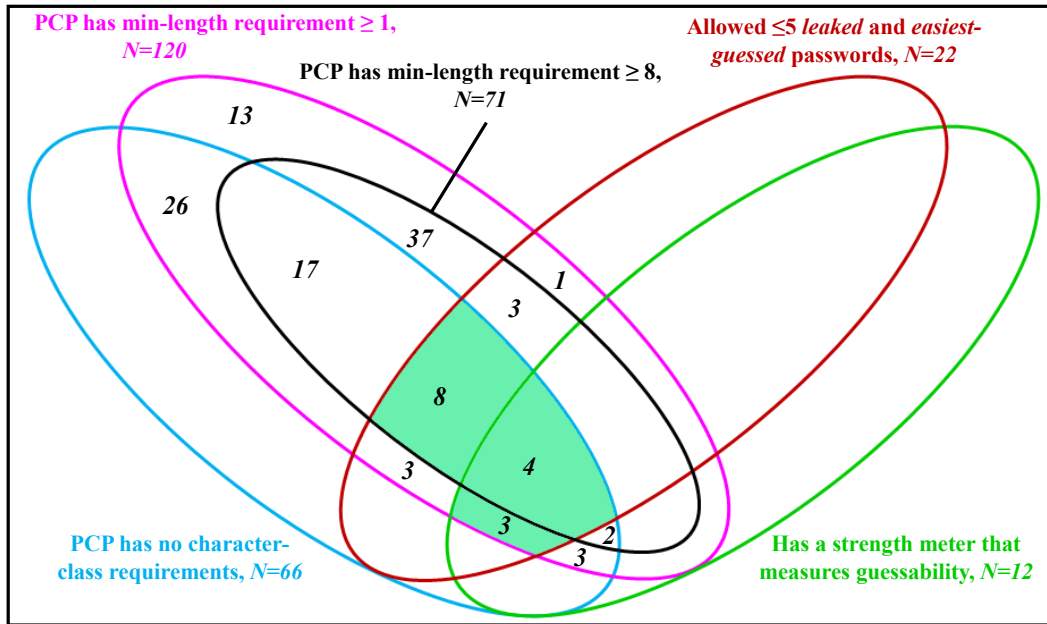


Figure 5: Websites following best practices are in the shaded green area. Unlabeled areas contain 0 websites.

Fig. 5 shows the breakdown of websites we considered to be following best practices. We considered a website to be following best practices if it allowed 5 or fewer of the 40 most common leaked passwords and easiest-to-guess passwords we tried, required passwords be no shorter than 8 characters, and did not impose any character-class requirements. We also considered websites with a shorter minimum-length requirement as following best practices if they satisfied the other two recommendations and further employed an accurate password strength meter to guide users to choosing strong passwords.

B Access failure details

Reason	Websites (N=142)
Inaccessible	69
No registration page	26
No passwords for auth	3
Government website	2
University website	4
Purchase required	7
Never received registration SMS	1
Non-U.S. phone number required	1
Site unreachable from browser	25
Explicit material	6
Non-English	38
Shared reg page w/ already-visited site	29

Table 4: Breakdown of the websites we skipped in our study.

We tried visiting the top 262 websites on the Tranco list in order to obtain the 120 websites for our study. Table 4 lists the reasons we skipped the other 142 websites.

C Lessons from our attempts at automation

Some readers may wonder why we pursued manual data sourcing methods in this study instead of an automated approach, since doing so may have enabled us to scale up the number of websites tested. As a matter of fact, we initially attempted two automated approaches which we ultimately abandoned due to concerns with completeness and data quality. We include our experiences in this writeup to hopefully serve as useful notes for those who want to extend our work.

We first tried building and using a Selenium-based web crawler to automatically extract PCPs from websites. Our crawler consisted of scripts tasked with parsing and navigating the sites of given domains to find the registration form and the PCP on the form. We leveraged search engine keyword searches to find registration pages (e.g., “join”, “create”, “signup”), as well as pattern detection of HTML tags and keywords to find and extract the PCP. However, we soon found that it was practically infeasible to develop any general solution; the unstandardized registration flows across websites required us to constantly add code to handle an extremely wide range of UI designs.

Our second approach utilizing MTurk was more successful,

	Fraction of accepted <i>leaked</i> passwords by stratum					Hypothesized minimum-strength threshold	
	1-10	11-100	101-1,000	1,001-10,000	10,001-100,000		
bit.ly	0/2	1/3	1/4	1/6	3/5	indeed.com	10 ³
chase.com	0/1	1/3	1/4	5/7	4/5	linkedin.com	10 ^{2.3}
espn.com	0/2	1/3	2/4	6/6	5/5	microsoft.com	10 ^{2.4}
facebook.com	0/2	1/3	0/4	3/6	4/5	roblox.com	10 ¹
instagram.com	0/2	1/3	0/4	3/6	4/5	reddit.com	10 ¹
slack.com	0/2	1/3	1/4	6/6	5/5	twitter.com	10 ^{3.4}
spotify.com	0/2	2/3	1/4	4/6	5/5	wetransfer.com	10 ^{1.5}
surveymonkey.com	0/2	2/3	1/4	4/6	4/5		
tripadvisor.com	0/2	0/3	0/4	6/6	5/5		
yelp.com	0/2	1/3	4/4	5/6	4/5		

(a) Looking at accepted *leaked* passwords by stratum, we hypothesized 10 websites were using shorter versions of the NCSC-HIBP-100k list.

(b) Minimum-strength thresholds we hypothesized were being used at seven websites. For reference, the threat of online guessing attacks ends at 10⁶ guesses.

Table 5: We found 10 websites that seemed to be blocking passwords based on a shorter common passwords list, and found 7 websites that seemed to be blocking passwords that did not meet a minimum-strength requirement.

but still produced data of dubious quality. We developed and published two separate MTurk Tasks to workers on the marketplace: one to identify registration pages from a given domain, and a second to extract the PCP from a given registration page. For each Task, the Worker—our hired user—was given the domain or registration page, and given a form to input information found such as the minimum-length requirement and character-class requirements. We also included quality assurance questions on the forms to confirm that the Worker had understood the given task and was paying attention. Despite the additional quality assurance measures, we found widespread inconsistency in the clarity of information collected across websites and even at the same website (we made sure to create two Tasks for each website). We concluded that our assurances were not rigorous enough, and that we had also underestimated the difficulty of educating Workers about extracting PCPs.

D Password blocking trends from § 3

Table 5 shows two trends we found in our password blocking study (§ 3). In Table 5a, we hypothesized 10 websites were using shorter versions of the most common passwords list we used. In Table 5b, we hypothesized 7 websites had a minimum-strength requirement.

E Additional findings from § 5

1. **Symbol definitions varied among the 37 websites requiring them.** 13 websites counted all 33 symbols we used towards their requirement, and half of the websites counted all but one symbol. The remaining websites below the median counted far fewer symbols, however, including 1 website that counted only #, \$, &, and @ (4 symbols), and 2 websites that counted only 9 and 10 symbols, respectively. The most commonly excluded

symbol was the space character, which counted at only 15 / 37 websites, followed by the ' , " , and ` characters, each counted as symbols at 28 / 37 websites.

13 websites placed even more restrictions on certain symbols by outright disallowing them in passwords, including the 2 websites that counted only 9 and 10 symbols; any symbol that did not count was not allowed to be in the password.

2. ***Iclass8* is the most common PCP.** 24 / 120 websites (20%) were using this PCP, followed by *Iclass6* (22 / 120). *3class8* is the most common character-class PCP (and third-most popular overall), we found it on 17 websites, followed by *4class8* and *DigSym6*, each found being used on 10 websites.
3. **Some websites were using maximum-length requirements that are too short.** 17 websites had a maximum-length requirement below 64 characters—the baseline recommended by NIST—including 1 website with a 14-character maximum length, 3 with a 15-character maximum, and 4 with a 20-character maximum [1]. Setting too short of a maximum length hurts security by preventing users from choosing long passwords that are hard-to-guess.

F Additional background

F.1 Password security is better modeled through adversarial guessability

Password strength has traditionally been measured using Shannon entropy, a function of the counts of lower- and uppercase letters, digits, and symbols (LUDS). While previously recommended by NIST, entropy—also commonly referred to as complexity—turned out to be a poor proxy for password security [31]. Researchers soon found mismatches between password entropy scores and time needed for attackers to

crack a password (or a set of passwords) [25]. The information security community has since favored using guessability as a measure of password security [8, 9].

Guessability more closely resembles the only practically important sense of password strength: the actual number of guesses an adversary would require to correctly guess the password. Unlike Shannon entropy, guess number metrics can factor in contextual information such as common passwords, human predictability and composition rules presented at password creation [25]. However, the attack method and configuration matters: many previous studies—facing time and resource constraints—have only been able to model specific attackers by using only one attack method with limited training data. A necessary drawback to the guessability approach is its inherent subjectivity. Whereas entropy is an objective measure, there is no objective guess number for any password; adversarial guessing is a strategic problem and different strategies will produce different results over the same password set input. For this reason, comparisons between studies using different guessing algorithms can be difficult at best, and moot at worst.

In an effort to harmonize future studies, in 2015, the Passwords Research Team at Carnegie Mellon University released Password Guessability Service (PGS)—a free service that rates the strength of submitted passwords [27]. PGS simulates a real attacker guessing passwords; it leverages multiple (5 at the time of our study) cracking tools to arrive at the user-provided plaintext password. Using each tool, PGS calculates the guessability (i.e., the guess number) as the password’s strength rating. PGS also offers the `min_auto` configuration, which returns the minimum guess number for each password across all 5 tools. Previous research has found that the `min_auto` approach provides a conservative estimate for the performance of an unconstrained professional attacker [27].

G Overall findings for all 120 websites

Table 6: PCP, blocklist results, strength meter, and security / usability ratings.

Website	Rank	Stores payment information	Stores PII	PCP	Allowed leaked	Allowed easiest-guessed	Strength meter	Percentage weak passwords rejected (security proxy)	Percentage strong passwords accepted (usability proxy)
google.com	1	•	•	Iclass8	0	0	Models guessability	62%	59%
netflix.com	2	•	•	Iclass6	20	20		11%	94%
facebook.com	4	•	•	Iclass6	8	2	Models complexity	11%	94%
microsoft.com	5	•	•	2class8	14	0		92%	39%
twitter.com	6	•	•	Iclass8	12	2		62%	59%
instagram.com	7	•	•	Iclass6	8	3		11%	94%
linkedin.com	9	•	•	Iclass8	14	0		62%	59%
apple.com	11	•	•	3class8	4	0	Models complexity	99%	7%
wikipedia.org	12	•	•	Iclass8	5	1		62%	59%
amazon.com	16	•	•	Iclass6	20	20		11%	94%
yahoo.com	17	•	•	Iclass7	0	0	Models guessability	47%	76%
pinterest.com	21	•	•	Iclass6	4	1		11%	94%
adobe.com	22	•	•	3class8	0	0		99%	8%
vimeo.com	24	•	•	2class8	18	15		100%	1%
wordpress.com	27	•	•	Iclass6	3	1		11%	94%
reddit.com	31	•	•	Iclass8	16	4		62%	59%
zoom.us	33	•	•	3class8	20	20		99%	7%
github.com	34	•	•	Iclass15 or 2class8	0	0		92%	36%
amazonaws.com	36	•	•	3class8	20	20		99%	8%
bit.ly	37	•	•	Iclass6	6	3		11%	94%
tumblr.com	43	•	•	Iclass8	0	0	Models guessability	62%	59%
vk.com	48	•	•	Iclass6	1	1		11%	94%
nytimes.com	49	•	•	Iclass6	20	20		11%	94%
flickr.com	51	•	•	Iclass12	20	20		100%	9%
dropbox.com	53	•	•	Iclass6	20	20	Models guessability	11%	94%
soundcloud.com	56	•	•	Iclass8	20	20		62%	59%
spotify.com	59	•	•	Iclass8	12	6		62%	59%
myshopify.com	60	•	•	Iclass5	20	20	Models guessability	4%	97%
cnn.com	65	•	•	4class8	20	20		100%	0%
forbes.com	66	•	•	4class8	20	20		100%	0%
ebay.com	68	•	•	DigSym6	11	9		85%	53%
theguardian.com	69	•	•	Iclass8	0	0		62%	59%
w3.org	70	•	•	Iclass8	0	0	Models complexity	62%	59%
paypal.com	72	•	•	DigSym8	15	5		77%	40%
twitch.tv	73	•	•	Iclass8	0	0	Models guessability	62%	59%
sourceforge.net	74	•	•	Iclass10	0	0		98%	21%
cloudflare.com	75	•	•	2class8	20	20		100%	1%
archive.org	76	•	•	Iclass3	20	20		0%	100%
imdb.com	77	•	•	Iclass8	20	20		62%	59%
bbc.co.uk	89	•	•	2class8	17	14		92%	37%
issuu.com	91	•	•	Iclass4	20	20		0%	100%
weebly.com	92	•	•	Iclass8	0	0		62%	59%
aliexpress.com	95	•	•	2class6	20	20	Models complexity	84%	57%
washingtonpost.com	96	•	•	Iclass8	20	20		62%	59%

Continued on next page

Table 6: (Continued) PCP, blocklist results, strength meter, and security / usability ratings.

Website	Rank	Stores payment information	Stores PII	PCP	Allowed leaked	Allowed easiest-guessed	Strength meter	Percentage weak passwords rejected (security proxy)	Percentage strong passwords accepted (usability proxy)
stackoverflow.com	98		•	2class8	20	20		92%	37%
etsy.com	99	•	•	Iclass6	20	20	Models complexity	11%	94%
reuters.com	103			4class8	20	20		100%	0%
tinyurl.com	106	•		Iclass6	20	20		100%	94%
tiktok.com	108		•	3class8	20	20		100%	1%
wsj.com	109		•	2class5	20	20		85%	53%
wix.com	113	•	•	Iclass6	20	20	Models complexity	11%	94%
bloomberg.com	114	•	•	Iclass8	2	1		62%	59%
sciencedirect.com	118		•	4class8	20	20	Models complexity	100%	0%
slideshare.net	120		•	Iclass5	20	20		4%	97%
imgur.com	121		•	DigSym6	20	20		85%	53%
oracle.com	122		•	4class8	20	20		100%	0%
opera.com	123		•	Iclass8	0	0		62%	59%
booking.com	125	•	•	3class10	20	20		100%	3%
indeed.com	126		•	Iclass8	9	1		62%	59%
businessinsider.com	127	•	•	3class8	20	20		99%	7%
canva.com	132	•	•	Iclass8	2	0	Models guessability	62%	59%
godaddy.com	135	•	•	Iclass9	20	20	Models guessability	94%	30%
godaddy.com	135	•	•	DigSym6	20	20		100%	1%
cnet.com	140		•	3class8	20	20		99%	7%
ibm.com	143	•	•	Iclass6	20	20	Models complexity	38%	82%
researchgate.net	144		•	3class8	20	20		100%	3%
digicert.com	145	•		Iclass5	20	20		5%	96%
dailymail.co.uk	148		•	Iclass5	20	20		11%	94%
slack.com	150	•	•	Iclass6	13	4		0%	100%
fandom.com	154			Iclass1	20	20		92%	37%
nature.com	157		•	2class8	20	20		92%	37%
force.com	159	•	•	2class8	19	18		100%	0%
cnbc.com	160		•	3class8	20	20	Models complexity	5%	97%
usatoday.com	161	•	•	Iclass5	20	20		92%	34%
chase.com	163	•	•	2class8	11	9		99%	7%
walmart.com	164	•	•	3class8	20	20		99%	8%
hp.com	166	•	•	Iclass8	20	20		62%	59%
surveymonkey.com	168	•	•	Iclass7	11	2	Models guessability	47%	76%
aol.com	170		•	Iclass7	0	0	Models complexity	11%	94%
yelp.com	171		•	Iclass6	14	6	Models guessability	62%	59%
eventbrite.com	173	•	•	Iclass8	20	20	Models guessability	62%	59%
telegraph.co.uk	174	•	•	Iclass8	20	20		100%	0%
opendns.com	176		•	4class8	20	20		0%	100%
cpanel.net	177	•	•	Iclass4	7	7	Models guessability	0%	53%
springer.com	186		•	DigSym6	20	20		85%	0%
time.com	187	•	•	3class8	20	20		100%	0%
npr.org	189		•	Iclass5	20	20		4%	97%
ted.com	190	•	•	Iclass8	20	20		62%	59%
samsung.com	191	•	•	3class8	19	14		99%	8%
myspace.com	194		•	2class8	20	20		92%	39%

Continued on next page

Table 6: (Continued) PCP, blocklist results, strength meter, and security / usability ratings.

Website	Rank	Stores payment information	Stores PII	PCP	Allowed leaked	Allowed easiest-guessed	Strength meter	Percentage weak passwords rejected (security proxy)	Percentage strong passwords accepted (usability proxy)
dailymotion.com	196		•	3class8	20	20		100%	1%
theforest.net	198	•	•	Iclass8	0	0		62%	59%
huffingtonpost.com	199		•	3class8	20	20		99%	8%
wired.com	200	•	•	Iclass6	20	20		11%	94%
mailchimp.com	201	•	•	4class8	20	20		100%	0%
espn.com	202		•	DigSym6	14	10	Models complexity	85%	53%
addthis.com	204		•	Iclass6	20	20		11%	94%
techcrunch.com	205		•	Iclass7	0	0	Models guessability	47%	76%
scribd.com	208	•	•	Iclass8	20	20		62%	59%
zillow.com	211		•	4class8	20	20		100%	0%
goodreads.com	212		•	Iclass6	20	20		11%	94%
unsplash.com	213		•	Iclass6	20	20		11%	94%
indiatimes.com	214		•	DigSym6	20	20		100%	1%
trello.com	219		•	Iclass8	20	20		62%	59%
grammarly.com	220	•	•	Iclass8	0	0		62%	59%
tripadvisor.com	221	•	•	Iclass6	11	3		11%	94%
freepik.com	222	•	•	DigSym6	0	2		100%	0%
independent.co.uk	225	•	•	DigSym6	20	20		99%	9%
roblox.com	226	•	•	Iclass8	18	6		62%	59%
squarespace.com	230	•	•	Iclass6	20	20		11%	94%
foxnews.com	232	•	•	Iclass6	20	20		11%	94%
zendesk.com	237	•	•	Iclass5	20	20		4%	97%
latimes.com	239	•	•	DigSym6	20	20		85%	53%
line.me	245	•	•	DigSym6	20	20		85%	52%
shutterstock.com	246	•	•	Iclass8	20	20		62%	59%
livejournal.com	247		•	DigSym6	19	10		99%	9%
wetransfer.com	248	•	•	3class8	19	4		99%	8%
intuit.com	250	•	•	4class8	20	20		100%	0%
intel.com	254		•	3class8	20	20		100%	1%
stackexchange.com	256		•	2class8	20	20		92%	37%
w3schools.com	262	•	•	4class8	20	20		100%	0%

Do Password Managers Nudge Secure (Random) Passwords?

Samira Zibaei
Ontario Tech University

Dinah Rinoa Malapaya
Ontario Tech University

Benjamin Mercier
Ontario Tech University

Amirali Salehi-Abari
Ontario Tech University

Julie Thorpe
Ontario Tech University

Abstract

Passwords are the most popular authentication method due to their simplicity and widespread adoption. However, the prevalence of password reuse undermines its security. A promising strategy to mitigate the risks of password reuse is to use random passwords generated and stored by password managers, yet many users do not use them. Many web browsers have built-in password managers that employ *nudges* at the time of password creation. These nudges aim to persuade the selection of more secure random passwords; however, little is known about which designs are most effective. We study ($n = 558$) the efficacy of nudges used by three popular web browsers: Chrome, Firefox, and Safari. Our results suggest Safari’s nudge implementation is significantly more effective than the others at nudging users to adopt a randomly generated password. We examine factors that may contribute to the adoption of randomly generated passwords, reasons that people adopt a randomly generated password (or not), as well as discuss elements of Safari’s nudge design that may contribute to its success. Our findings can be useful in informing both future password manager nudge designs and interventions to encourage password manager use.

1 Introduction

Authentication with passwords, despite its security [14, 52] and memorability [21, 38] shortcomings, remains widespread with applications such as online banking, e-commerce, personal devices, servers, etc. The average person is estimated to have at least 26 accounts [38] and possibly more than 100

accounts [1]. The burden of remembering many passwords often leads users to rely on insecure coping methods [33], such as using the same, simple, or similar passwords [48]. To prevent these insecure coping mechanisms, password managers have become an instrumental tool for storing and generating random, complex passwords to assist users with password security and memorability. The passwords generated by password managers are expected to be less vulnerable to credential stuffing [50]—a serious concern due to password leaks [18]—and to password guessing attacks [54]. However, password managers have not fully delivered their security promises in practice [5, 39].

Password managers, despite being recommended by security experts [20], are still not adopted by many users [39, 48]. Even when people make use of password managers, only a minority use the random password generation feature that enables its secure use [39]. One might wonder how to further encourage users to adopt password managers and also to accept randomly generated passwords as their password. One potential promising solution is *nudging* techniques [16] to influence adoption of password managers and their security features (e.g. randomly generated passwords) without limiting user choices [29]. While nudging has been explored in human-computer interaction [9] and some computer security contexts [58], research on nudging in the context of adopting password managers or their security features (e.g., randomly generated passwords) is sparse.

In this paper, we initiate studying the effect of nudging on the adoption of security and storage features of password managers. In particular, we explore how effective current browser-based password managers are at nudging users to adopt their randomly generated passwords and storage features. We also aim to gain a deeper understanding of why people choose to adopt generated passwords (or not). Our specific research questions are:

*Contact author: Samira Zibaei <samira.zibaei@ontariotechu.ca>.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.

August 7–9, 2022, Boston, MA, United States.

- Q1 How do the three most popular browser-based password managers (Chrome, Firefox, and Safari) compare to each other in nudging users to adopt randomly-generated passwords?
- Q2 Does the complexity of a website’s password policy contribute to the adoption rate for randomly generated passwords?
- Q3 What factors contribute to the adoption rate for randomly generated passwords and saving passwords in the password manager?
- Q4 What are the rationales of users to (not) adopt a randomly generated password?

To investigate these questions, we conducted a user study ($n = 558$) to evaluate the effectiveness of the generated password nudges employed by Chrome, Firefox, and Safari. Participants were asked to register for a new website, so we can observe their interaction with the password managers. Following registration, participants complete a questionnaire that asks their reasons for adopting the generated password (or not). Our website assigned participants one of two password policies (1C8 and 3C12)¹ to evaluate its impact on users’ decisions when confronted with simple or complex password requirements. We perform both quantitative and qualitative analyses on our collected statistics and participant’s free-form comments regarding their use of the randomly generated password during their account registration.

Our contributions and findings include: (i) Analysis of which browser password manager nudges are most effective. We discuss differences between the nudge designs of the password managers we study, and possible reasons for our findings, which can be useful in informing future password manager nudge designs. (ii) Identification of a number of factors that influence users’ decision to adopt a randomly generated password, such as previous use of a password manager, former familiarity/use of a generated password, and whether they noticed the nudge. (iii) Investigation of reasons why people believe they did (not) use the generated password. This information can also be useful in informing both future password manager nudge designs and interventions to encourage password manager use.

The paper is structured as follows. Section 2 discusses previous work that relates to our research. Section 3 elaborates on the purpose of our study, how we recruited our participants, our study’s structure, how we collected our data, statistical testing methods, and qualitative analysis methods employed. The results of our study are presented in Section 4, as well as participant demographic information. We discuss the results and limitations of our study in Section 5, and conclude in Section 6.

¹ 1C8 is a password policy that only requires a minimum of 8 characters. 3C12 is a password policy that requires a minimum of 12 characters and at least 3 character classes. Character classes include lowercase characters, uppercase characters, special characters, and numbers.

2 Related Work

We first briefly review password shortcomings, then discuss related work on password managers and nudging.

Many passwords to manage. Passwords remain the most popular authentication method for computer systems [8]. Unfortunately, with the proliferation of online services, the number of passwords that each user needs to remember has increased exponentially. The average person has between 70–80 passwords [56]. Creating strong, unique, and complex passwords that are easy to remember is an unavoidable challenge for users. As a result, users resort to making weak passwords that are easy to remember (sometimes, with their personal information) or reuse their passwords for multiple accounts [15, 31]. Both of these practices yield lower security. With password reuse, the leak of a password from one account renders other potentially high-risk accounts vulnerable [48]. Passwords with personal information are more vulnerable to guessing attacks [54]. Also, recently many advances have been made towards more effective guessing attacks, which leverage the reoccurring password patterns in large-scale leaked password datasets and machine learning techniques [22, 25, 32, 35, 36, 53, 54].

Password managers and usability. Password managers can generate, store, and remember random passwords for users to enhance their password security. Several usability issues have been reported in studies conducted on password managers such as poor user interface design [3] and lack of important functionalities (e.g., recovering changed or deleted credentials) [5]. It is shown that the use of technical terms when describing features (e.g., “password generator”) makes password managers seemingly complicated for users [47]. Recently, a cognitive walkthrough indicated some features of password managers (e.g., autofill, user interface design, and linking credentials to multiple sites) might help foster their adoption [46]. Some attempts have been made to improve overall usability of password managers by minimizing the user’s action and enhancing their user interfaces [7, 49].

Adoption of password managers. The adoption of password managers has faced challenges beyond their usability issues. The low adoption rates of password managers is blamed on the lack of: user’s trust [5, 45] in this technology, willingness to be dependent on technology [41], and awareness of its benefits [12, 45]. Convenience is yet another reason found for users not using password managers [24]. Other research found that a barrier to password manager adoption was not having enough accounts to protect, believing their accounts are not valuable enough to require using a password manager, lack of accessibility of passwords on multiple devices, and concern of the password manager’s single point of failure [39]. Older adults (above 60) were found to have low adoption of password managers due to concerns about where their password is stored, and whether others might have access

to their accounts [41]. Also, impediments to adoption of standalone password managers include users not having time to install the software [6], not understanding the sense of its urgency [6], or being unwilling to hand over the control of their own passwords [10]. Other research indicates that cybersecurity knowledge is an important factor in the adoption of a password manager [5].

Adoption of randomly generated passwords. The low adoption of randomly generated passwords from password managers is a concern, which has downgraded the potential security impact of password managers. The under-deployment of randomly generated passwords might be due to a lack of awareness, interest, or trust [45]. Pearman et al. [39] in an interview study ($n = 30$) found that only one out of 12 participants who used a “built-in” or browser-based password manager adopt randomly-generated passwords, whereas all 7 participants with stand-alone password managers adopt random passwords.

Nudging. Nudging, a concept in behavioral science, aims to influence decisions without limiting people’s choices [23]. Nudging has been employed in many contexts, and is of interest to a broad range of human-computer interaction (HCI) topics [9]. In cybersecurity, it has been used in many security decisions [58], including which Wi-Fi network to join [57], social network posts to make [55], and emails to trust [11]. Nudging has also been applied to tackle the problem of password creation in alphanumeric passwords (through password strength meters [43, 44]) and graphical passwords [37, 51], and password manager adoption [2, 4]. Nudging has also been studied in the context of promoting users to accept randomly-generated passwords [23]; although the studied nudges were unsuccessful, they were quite different than those employed by current password managers.

Our work. Nudging is employed by a number of popular browser’s built-in password managers: Chrome, Firefox, and Safari (see Figure 1b-d). However, the efficacy of these nudges has not yet been studied, to the best of our knowledge. In this paper, we study the efficacy of these browser nudges, factors that may influence their efficacy, and users’ reasoning for accepting (or not accepting) the nudge. Our goal is to deepen our understanding of what nudges work best in this context, and why, which can be used to help improve the state-of-the-art.

3 Methodology

Our primary goal is to evaluate the effectiveness of nudges employed by the three most popular browsers: Chrome, Firefox, and Safari, in terms of their ability to encourage the use of randomly generated passwords. We review the browser nudges studied in Section 3.1. We created a mock-up of a new e-commerce website for purchasing local produce (Fig-

ure 1a) to examine user behavior when creating an account on the website. We collected and analyzed quantitative data composed of users’ decisions while creating an account (e.g., if a randomly generated password is adopted) and both quantitative and qualitative data from users’ responses to a questionnaire. Our study was reviewed and approved by our institution’s Research Ethics Board. We explain the structure of our study further in Section 3.2. Our recruitment method is described in Section 3.5 and resulting demographics are summarized in Section 3.6. We outline our analysis approach in Section 3.7.

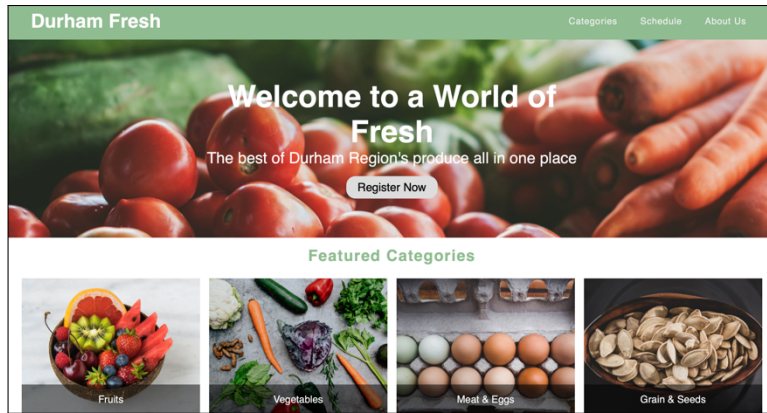
3.1 Nudges in Chrome, Firefox, and Safari

Each browser uses nudges to encourage people to use a random password generator. Chrome’s *just-in-time nudge* (see Figure 1b) is displayed when a user clicks the password field. This nudge suggests the user a 15-character randomly generated password to encourage its adoption. Chrome displays the suggested random password with the message of “Use suggested password”, and includes the following statement, “Chrome will save this password in your Google Account. You won’t have to remember it.” The focus of the nudge appears to be more on convenience than on security with an emphasis on remembering passwords for user. Chrome’s nudge is simple and does not seek to grab the user’s attention. Firefox’s nudge (see Figure 1d) is also a just-in-time nudge and very similar to that of Chrome, even in terms of the length of passwords. Firefox uses the term “Securely” in its message of “Use a Securely Generated Password”, followed by the statement of “Firefox will save this password for this website.” The main difference in word choice is that Firefox’s nudge puts emphasis on security as well as convenience.

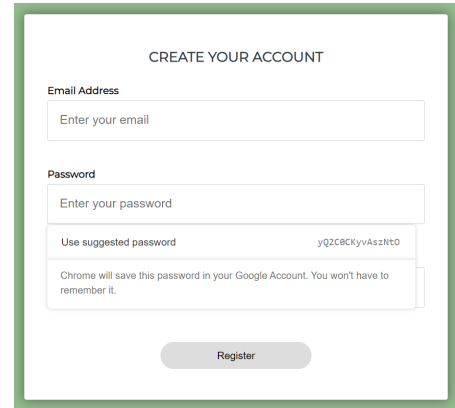
Safari (Figure 1c) uses a different method of nudging known as a *default nudge*. A default nudge works by selecting the desired option by default. To encourage the selection of a random password, Safari automatically populates the password field with an 18-character random password when the user clicks in it. Safari’s nudge is accompanied by a pop-up message of “Safari created a strong password for this website—This password will be saved to your iCloud Keychain and will AutoFill on all your devices. Look up your saved passwords in Safari Password preferences or by asking Siri”. Safari’s nudge is the most visually diverse and puts emphasis on both password strength and convenience. Safari’s use of color and a default nudge is a clear attempt to grab user’s attention. Furthermore, Safari’s description of its password manager’s functionality aims to educate users and persuade them to use it.

3.2 Study Structure

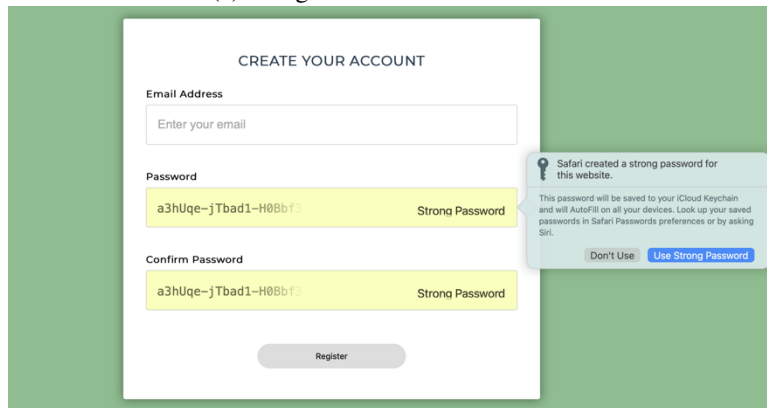
We designed our study to employ deception in order to keep our website registration as realistic as possible, without



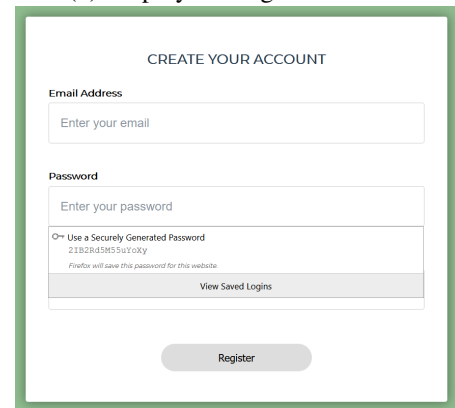
(a) Designed e-commerce website



(b) Employed nudge in Chrome



(c) Employed nudge in Safari



(d) Employed nudge in Firefox

Figure 1: Our mock-up website for which study participants were asked to register an account is shown in (a). The browser nudges studied are: (b) Chrome, (c) Safari, and (d) Firefox.

drawing additional attention to the nudge. The users are first falsely informed that the purpose of our study is evaluating the user interface and functionality of our (fake) e-commerce website. However, participants were debriefed with the actual purpose of the study in a secondary consent form and participant data is only collected if they agree to it; otherwise, they were considered to have opted out. For our study, participants were specifically required to do the following tasks:

Task 1: First consent form. Participants were provided with a deceptive consent form that explained the purpose of the study is to help evaluate the usability of our website's registration and login processes (see Appendix A). It did not reveal the study's true focus on passwords and nudging.

Task 2: Account registration. The participants were asked to test the usability of our website's registration and login process by creating an account using a valid email and a password that conforms the password policy. Users have the freedom to create their own passwords or use the browser's password manager for a randomly generated password. Regardless of how users create their passwords, users are given the option to store their passwords using their browser's password manager.

Task 3: Post-registration questionnaire. We asked participants

to answer 5 demographic questions including their age, gender, education, their primary area of study or work, and their first language. See Appendix B for full details.

Task 4: Login. Participants were asked to log in to their accounts created in Task 2 using their email address and chosen passwords. If the users have stored their passwords in the browser's password manager, the password manager would autofill their stored password.

Task 5: Post-study questionnaire. Participants were asked to answer questions focusing on users' behavior relating to the nudges, password managers, and password creation. See Appendix C for full details.

Task 6: Second consent form. Participants were provided with a second (real) consent form (see Appendix D) that explains the true purpose of the study before submission.

3.3 Ethical Considerations

If participants initially knew the purpose of our study, it would bring unrealistic focus to the randomly generated password nudge. Therefore, we used deception by telling participants that they are testing the usability of a new e-commerce

website’s registration and login process. Participants were debriefed through a secondary consent form (see Appendix D), which they were asked to read carefully before agreeing to submit their data. Participant data is only stored after they agree to this secondary consent form. To mitigate the risk of users reusing one of their passwords, we only collect/store passwords/data with anonymous identifiers, and only after obtaining secondary consent. Our study was reviewed and approved by our institution’s Research Ethics Board.

3.4 Implementation Details

For the account registration task of our study, the password policy for a user is randomly set to be either a 1C8 or 3C12 password policy. A 1C8 policy only requires a minimum of 8 characters, whereas a 3C12 policy requires a minimum of 12 characters with at least 3 character classes of lowercase characters, uppercase characters, special characters, and numbers. We used these two different password policies to analyze users’ password decisions when confronted with simple and complex password requirements. Based on the user’s browser (Chrome, Firefox, or Safari), our website shows a simulation of the browser’s password manager and records their interactions with the simulated password manager. We also ensure the actual password managers are not invoked when the simulated password managers are presented to the user. The simulated password managers are designed to appear identical to the actual password managers, but are intended to facilitate data collection of user’s interaction with the password manager. Figure 1b-d are taken from our simulated password managers, which show how carefully they were designed to be identical to the actual password managers.

Our system enforced a number of rules relating to study completion. To improve data quality, users can only have one tab/instance of our study running at a time and are only able to complete our study once. These rules are to prevent biased results from users who have already completed our study. For ethical reasons, we only collect data once users have submitted both consent forms. This means participants who leave the study before providing the final consent will have their data deleted after 10 minutes of inactivity. Once the data is removed from our servers, users will have to restart our study. However, before the 10 minutes is up, users have the option of continuing our study by restoring their closed tab. Between starting the study and seeing the second consent form, a total of 100 participants opted out. To ensure participants’ anonymity, we do not collect their emails, we only collect their passwords.

3.5 Recruitment

We tested our study through a pilot study with 5 participants and asked them to provide us with their feedback. We improved our study’s design and user interface based on their

comments. Our user study was conducted with 561 participants recruited through Amazon’s Mechanical Turk (MTurk) website. Participants were limited to those living in the United States. The estimated completion time for this study was about 5 minutes. To be consistent with minimum wage in the United States (\$7.25 USD per hour), participants were compensated \$0.60 USD for the completion of our study. Participants could choose to sign up for any one of the three MTurk groups, and we used the user-agent header to determine the correct browser is in use.

3.6 Participant Demographics

Table 1 presents an overview of the participant demographics for our study collected through the post-registration questionnaire (see Appendix B). Our participants were composed of 48.4% female, 49.6% male, and 2% who preferred not to specify their gender.

Participants’ ages range from 18 to over 50 years old. The majority of participants (39.6%) fell within the 26–35 age group, followed by the age group of 36–50 making up 27.8% of participants. Regarding participants’ education level, most participants (54.1%) had a Bachelor’s degree, followed by a high school degree (26.3%). The majority of participants in our study (30.1%) belonged to the business and IT field of education or work.

3.7 Analysis

We analyze our results to find whether there are significant differences between the adoption rate of randomly generated passwords for: (1) the three browsers studied, (2) the two implemented password policies (1C8 and 3C12), (3) participants who noticed the nudge vs. those who did not notice the nudge, (4) participants who used a password manager before vs. those who have not, (5) participants who used a random password generator before vs. those who have not, (6) participants who are using their main (daily use) browsers in our study vs. those who did not. We also analyze our results for whether there are significant differences between the rate of saving passwords in the password manager for: (7) the three browsers studied, and (8) participants who noticed the nudge vs. those who did not notice the nudge. Since all of these analyses involve comparing proportions, we use the χ^2 test to find whether there are significant differences between them. Tests were conducted using Bonferroni adjusted alpha levels of 0.006 per test (0.05/8).

We performed a qualitative analysis on the free-form data from our post-study questionnaire, to find underlying reasons participants did (or did not) use randomly generated passwords. We asked our participants, “Can you describe the reason why you used/did not use the random password generator?” Participants’ comments were analyzed using an emergent coding approach, and two researchers coded all

		Chrome	Firefox	Safari			Chrome	Firefox	Safari
Gender	Female	45.5%	41.5%	58.7%	Study/Work	Social Sci. & Humanities	5.2%	8.5%	6.7%
	Male	52.4%	56.9%	39.1%		Science	6.3%	5.9%	7.8%
	N/A	2.1%	1.6%	2.2%		Health Science	7.9%	4.3%	13.4%
Age	18-25	11.5%	12.8%	27.4%		Engineering & Applied Sci.	8.9%	9.6%	4.5%
	26-35	42.4%	38.8%	37.4%		Energy & Nuclear Sci.	0.0%	1.1%	1.1%
	36-50	25.7%	30.9%	26.8%		Education	8.4%	5.3%	14.5%
	50+	19.4%	16.5%	8.4%		Business & IT	38.2%	30.9%	20.7%
	N/A	1.0%	1.0%	0.0%		Other	16.7%	25.4%	24%
Education	High school	23.6%	30.3%	25.1%		N/A	8.4%	9%	7.3%
	Bachelor's	58.6%	53.7%	49.7%		Language	English	95.8%	96.8%
	Master's	14.1%	9.6%	18.4%	French		0.5%	0.0%	0.6%
	PhD/higher	1.6%	3.7%	3.4%	Other		2.7%	2.1%	10.6%
	N/A	2.1%	2.7%	3.4%	N/A		1.0%	1.1%	1.1%

Table 1: The user demographics across the three browsers

participants' comments independently by categorizing their statements [28]. Some participants described multiple reasons for (not) using randomly generated passwords, and we applied multiple codes to these comments. To measure the reliability of our coding process, we used Cohen's Kappa [28]. Our resulting $\kappa = 0.98$, suggesting near-perfect agreement between the two researchers.

4 Results

We recruited a total of 561 paid users on Amazon MTurk to participate in our study. We removed three responses due to inconsistent answers to an attention check question that asked users to select a specific number from the list (see Question 5 in Appendix C). We examine our research questions using the remaining 558 responses (191, 188, and 179 participants for Chrome, Firefox, and Safari respectively). The difference in group sizes is partly due to the Safari condition taking the longest to fill, whereas Chrome was the fastest.

4.1 Efficacy of Generated Password Nudge

Our results on the effectiveness of the nudges for each built-in password manager are shown in the first row of Table 2. To determine whether any one of these nudges are more effective than others while registering for our website, we test the following hypothesis:

H_0 The randomly generated password adoption rates are similar between the three browser groups.

H_a The randomly generated password adoption rates differ between the three browser groups.

To test this hypothesis, with the browser groups of Chrome, Firefox, and Safari, we used a χ^2 test ($df = 2$, $N = 558$). We reject the null hypothesis H_0 ($\chi^2 = 32.972$, $p < 0.001$) after Bonferroni multiple-test correction. The effect size is

	Chrome	Firefox	Safari
RGPs (1C8+3C12)	35.2%	41%	61.5%
RGPs (1C8)	26.5%	34.7%	61.3%
RGPs (3C12)	38.7%	47.3%	61.6%
Saved password	49.2%	55.3%	70.4%

Table 2: Percent of participants who adopted the randomly generated passwords (RGPs), were influenced by the complexity of the website's password policy (1C8 or 3C12), and saved their passwords in a password manager.

moderate (Cramer's $V = 0.24$). Therefore, we accept our alternative hypothesis H_a that the generated password adoption rates differ between the Chrome, Firefox, and Safari browser groups. As shown in Table 2, more users adopted the Safari nudge than the other two browsers. We will discuss possible reasons for Safari's nudge effectiveness in Section 5.1.

4.2 Efficacy of Nudge to Save Passwords

As shown in Table 2, 49.2% of Chrome users, 55.3% of Firefox users, and 70.4% of Safari users saved their passwords in their respective browser-based password manager. All participants who used a random password generator stored them in a password manager, as well as some additional participants who created their own passwords. We analyzed the saved passwords to determine if users were saving their own passwords or randomly generated passwords. 87.3% of the Safari users who saved their passwords saved a randomly generated password. While 74% of Firefox users and 66% of Chrome users saved randomly generated passwords. To determine if any of these nudges are more effective to encourage users to save their passwords, we test this hypothesis:

H_0 The rates of password storage are similar between browser groups.

H_a The rates of password storage differ between browser groups.

Using the χ^2 test ($df = 2, N = 558$), we reject the null hypothesis H_0 ($\chi^2 = 15.90, p < 0.001$), with weak effect size (Cramer’s $V = 0.16$). We conclude that participants did not have similar behavior regarding storing their passwords in a browser-based password manager, and Safari users were more likely to save their passwords.

4.3 Impact of Website’s Password Policy

Our results on the effectiveness of the nudges for each built-in password manager under a simple password policy (1C8) and a complex password policy (3C12) are shown in the two middle rows of Table 2. To determine whether the complexity of the website’s password policy influences user choice to adopt a generated password, we test the following hypothesis:

H_0 The randomly generated password adoption rates are similar between website password policies.

H_a The randomly generated password adoption rates differ between website password policies.

To test this hypothesis, with the website password policy groups of 1C8 and 3C12, we used a χ^2 test ($df = 1, N = 558$). We fail to reject the null hypothesis ($\chi^2 = 3.921, p = 0.047$) after Bonferroni correction ($\alpha < 0.006$), so we accept the null hypothesis and suggest that the website’s password policy likely does not create enough pressure to impact user’s adoption of a generated password.

4.4 Analysis of Possible Adoption Factors

Our goal in this analysis is to understand whether some factors may contribute to user’s adoption of generated passwords and the password manager to save passwords. Our post-study questionnaire features several questions related to participants’ familiarity with random password generators and password managers. We also ask participants if they noticed the nudges while registering and their reason for using/not using a random password generator.

More specifically, we investigate the following factors to determine their impact on adopting a randomly generated password: (i) noticing the browser’s generated password nudge, (ii) experience with using a password manager, (iii) experience with using a random password generator, and (iv) being a regular (daily) user of the browser, since repeated exposure to the nudge may make it easier to ignore. We also investigate (v) whether noticing the browser’s nudge could be a factor in users saving their password in the password manager. Table 3 shows the overall frequencies of participants’ responses to the post-study questionnaire questions on factors (i)-(iii). Data related to factor (iv) is shown in Table 4. In the following subsections, we test whether each of

	Chrome	Firefox	Safari
Used password manager before	68.6%	64.4%	65.9%
Used password generator before	48.2%	53.2%	57.5%
Noticed the nudge	70.2%	71.8%	88.8%

Table 3: Frequencies of participant characteristics based on post-questionnaire data.

	Chrome	Firefox	Safari
Daily	89%	70.2%	57%
Weekly	6.3%	9.6%	11.2%
Monthly	0.5%	4.8%	9.5%
A few times per year	0.5%	11.2%	16.2%
Never used	3.1%	3.2%	4.5%

Table 4: Frequencies of participant’s usage of the browser used in our study (from post-questionnaire data).

these factors were related to the adoption of generated passwords in our study. Our analysis suggests that noticing the nudge has an impact on both adopting the randomly generated password and on saving it. Our analysis also suggests that previous use of a password generator impacts users’ adoption of randomly generated passwords. However, previous use of a password manager does not influence users’ adoption of randomly generated passwords. We found that being a regular (daily) user does not significantly impact the rate of adopting the randomly generated password.

4.4.1 Noticing the Nudge on Generated Password

To determine whether participants noticed the nudge, we asked them “Did you notice the recommendation to use a random password while registering on our website?” in our post-study questionnaire (see Question 4 in Appendix C). We investigate their answers to find if there is a significant difference between participants who noticed the nudge and those who did not regarding using random password generators. Table 3 shows that Safari’s nudge was most successful at being noticed by participants. We also found that 43.3%, 50.4%, and 59.7% of Chrome, Firefox, and Safari participants who noticed the presence of the nudges used a random password generator in our study. Overall, almost half (48.4%) of the total number of participants who noticed the nudges in our study decided to create their own passwords, regardless of the nudges’ urge to use a randomly generated password. Although noticing the nudge increases the acceptance rate of randomly generated passwords, in Safari the acceptance rate decreases slightly (approx. 1%). However, note that only a small number of Safari users (19/179) didn’t notice the nudge.

To determine whether noticing the nudge to use a random password influences user choice to adopt a generated password, we test the following hypothesis:

H_0 The randomly generated password adoption rates are similar between participants who noticed vs. did not notice the nudge.

H_a The randomly generated password adoption rates differ between participants who noticed vs. did not notice the nudge.

To test this hypothesis, we used a χ^2 test ($df = 1, N = 558$). Our finding indicates a significant difference between above-mentioned groups of participants in terms of using a random password generator ($\chi^2 = 39.265, p < 0.001$). The effect size is moderate (Cramer's $V = 0.26$).

4.4.2 Previous Password Manager and Generator Use

Based on our findings, 53.3%, 53%, and 67% of Chrome, Firefox, and Safari participants who have experience with using password generators before used a random password generator in our study. Accordingly, in terms of having experience with using password managers, 38.2%, 43%, and 64.4% of Chrome, Firefox, and Safari users used a random password generator in our study. To determine whether using a password generator before influences user choice to adopt a generated password, we test the following hypothesis:

H_0 The randomly generated password adoption rates are similar between the participants who have used vs. have not used password generators before.

H_a The randomly generated password adoption rates differ between the participants who have used vs. have not used password generators before.

Additionally, to determine whether using password managers before influences user choice to adopt a generated password, we test the following hypothesis:

H_0 The randomly generated password adoption rates are distributed similarly between participants who have used vs. have not used password managers before.

H_a The randomly generated password adoption rates are distributed differently between participants who have used vs. have not used password managers before.

To test this hypothesis, we used a χ^2 test ($df = 1, N = 558$). Based on our results, participants who were familiar with the password generator are more likely to use it while creating an account ($\chi^2 = 43.842, p < 0.001$). The effect size is moderate (Cramer's $V = 0.28$). However, there is no significant difference between users who used a password manager before our study in terms of using a random password generator ($\chi^2 = 5.154, p = 0.023$).

4.4.3 Regular Use of Browser

We define a user's *regularly-used browser* as a browser used on a daily basis. If a participant regularly uses a browser,

it is possible that they are used to the nudge, and it may be less effective for them. To evaluate whether this might be a factor, we asked participants how often they use the browser they used for our study. Table 4 indicates the percentage of how often the browser was used in each browser group (Chrome, Firefox, and Safari). Further analysis of our data indicated that 42.4% of participants who use Firefox daily used a random password generator in our study. While 68% of participants who use Safari daily used a random password generator. The percentage of daily Chrome users who used a random password generator in our study is 29.4%. Overall, 72.6% of participants in our study indicated that the browser they used for this study is one they use daily. Only 3.6% of participants stated they had no experience using the browser they used to complete our study. Among all participants who were using a regularly used browser to complete our study, 43.5% of them generated their password using a random password generator, which means that more than half of the participants do not adopt the randomly generated password when they are using a regularly used browser. To determine whether using a regularly used browser influences user choice to adopt a generated password, we test the following hypothesis:

H_0 The generated password adoption rates are similar between the participants who used a regularly-used browser vs. the participants who used an infrequently-used browser.

H_a The generated password adoption rates differ between the participants who used a regularly-used browser vs. the participants who used an infrequently-used browser.

To test this hypothesis, we used a χ^2 test ($df = 1, N = 558$). Based on the results ($\chi^2 = 0.81, p = 0.366$) there is not a significant difference between these two groups regarding using a random password generator in our study.

4.4.4 Noticing the Nudge on Password Storage

Since storing a password in a password manager is the second primary function of the password manager, we investigate our result to find whether noticing the recommendation to use a random password affects the user's decision to store their password in a browser's password manager. To determine whether noticing the nudge influences user choice to save a password in a browser-based password manager, we test the following hypothesis

H_0 The rates of password storage are similar between the participants who noticed vs. did not notice the nudge.

H_a The rates of password storage differ between the participants who noticed vs. did not notice the nudge.

To test this hypothesis, we used a χ^2 test ($df = 1, N = 558$). Interestingly, participants who saved their passwords in a password manager mostly belong to the group of participants who

noticed the nudge, and the difference between participants who noticed the nudge and then saved their passwords and participants who did not notice the nudge and saved their password in a password manager is remarkable ($\chi^2 = 33.321$, $p < 0.001$). The effect size is moderate (Cramer's $V = 0.24$).

4.5 Why (not) Random Passwords?

The codebook with the frequencies, along with examples for each code is provided in Table 5. When analyzing participants' reasons for using a random password generator, 19.89% of participants from this group reported convenience vs. 12.19% for security. The next most common response was password storage feature (5.56%), meaning random password generators' main appeal is convenience and security. When analyzing participants' reasons for not using a random password generator, 23.66% of participants in this group reported random passwords are too hard to remember. The next most common response was participants preferred to create their own passwords (11.47%), which indicates that the endowment effect may also be a major reason for rejecting randomly generated passwords. It is possible that this reluctance to use randomly generated passwords is rooted in participants feeling unsafe when they are unable to memorize their passwords. Our study confirms others' findings that the main reasons for adopting randomly generated passwords are convenience [39, 45], and security [45], but differs regarding the save password feature's importance [31]. Moreover, Our study confirms other's findings that the main reasons for rejecting randomly generated passwords are memorability issues and user preferences [31, 45], but differs regarding the importance of a lack of awareness [39, 45], trust [45], or concern [39]. We discuss the implications of these findings in Section 5.

5 Discussion

Our study empirically tests the effects of nudges employed by the three most popular browsers: Chrome, Firefox, and Safari. We were also interested in understanding the factors that influence users' decisions while creating a password. The results from our server logs and questionnaires suggest that the majority of the participants from each browser group completed our study using a browser they use regularly, and that regular use of the browser didn't influence adoption of the randomly generated password.

5.1 Possible Reasons for Safari's Effectiveness

Safari had the most effective password manager nudge in terms of influencing participants to use a random password generator and save their passwords. Additionally, our results indicate that Safari has the most noticeable nudge when compared to Chrome and Firefox. Safari's nudge (Figure 1c)

is clearly more visually striking than Chrome's or Firefox's nudge. Safari's use of color, an additional pop-up box, and automatically populating the password field with a randomly generated password makes their nudge much more prominent. Chrome and Firefox take a subtle approach to suggest that people use randomly generated passwords. In contrast, Safari's pop-up message includes some information on the storage and autofill features. This is useful for users who are unfamiliar with password managers and may help people become more comfortable adopting this feature. Additionally, Safari uses a default nudge which takes the liberty of populating the password field with a randomly generated password and emphasizes its strength with the message, "Strong password". A quantitative review of 100 publications on nudging which aimed to determine the effectiveness of various nudging techniques states that, "default nudges are the most effective" [17]. The effectiveness of default nudges is also shown in two other studies [19, 30]. Therefore, a reason Safari is effective at convincing people to use random passwords could be attributed to the fact that Safari decides for you. Unless users take the effort to create a password themselves, simply using the password provided is more convenient. Alternatively, Safari making the choice to input a random password by default may convince users that it is the recommended action. Safari's nudge clearly expresses that the generated password is strong, implying to the user that it is the optimal password to use. Another interesting element of Safari's design is that it contains a visual effect on the last six characters of the password, giving the impression that the password contains even more characters than are seen. It is possible that this visual effect is interpreted by the user as the password offering even more security, as it appears longer and as though there are parts that couldn't be observed through shoulder-surfing.

In general, we found higher rates of randomly generated password adoption and awareness than another study [39], which found that 14% of Safari users used randomly generated passwords, while Chrome users were unaware of randomly generated passwords. Our results found higher Safari user adoption rates (61%) and also Chrome user awareness of randomly generated passwords (30% adopted randomly generated passwords); this may be due to changes in user behavior over time (2018-2022), or differences in methodology, as their study [39] was conducted through semi-structured interviews ($n = 30$).

5.2 Reasons Participants Used Password Manager Features (or Not)

Our post-study questionnaire (Appendix C) asked participants to specify their reasons for using (or not using) a random password generator. Emergent coding was then used to analyze the free-form, self-reported data from our questionnaire and categorize participants' comments. By categorizing

	Code	Frequency		Examples of participants' reasons on why they used/not used password generator
Reasons for using a random password generator	Convenience	111	19.89%	"It seemed convenient to use a securely generated password." "I used it because it seemed faster than creating a new password."
	Security	68	12.19%	"I figured the random password was strong enough so I accepted it." "Random passwords seem more secure, since they cannot be guessed by intruders."
	Remember Password Feature	31	5.56%	"I did because it was saved to Chrome and I can go back in and edit it later if I want." "I used it because it saves my password for the next time I would login to the site."
	Didn't care about the website	26	4.66%	"I selected the random password generator because it is a tempt site." "I didn't want to think of an actual password for this site."
	Avoid reusing passwords	23	4.12%	"I did not want to use one of my regular passwords." "I rather not give a random site a password I would usually use."
	Noise	9	1.61%	"NONE" "I did use it."
	Strict password policy	9	1.61%	"I used it because I couldn't really think of a 12 character password." "I did use it. I used it because it tried to require a 12 digit password, and that is too long to make up myself."
	Preferred to use a generator	7	1.25%	"I used the generator because I usually always do." "Habit. I've always generated/used single use passwords per website/service."
	Incongruous	5	0.90%	"If the password manager were to fail I would lose all my passwords." "I wanted to create my password from scratch and not use anything else."
	Unsure	5	0.90%	"I am confused and not aware of this option." "I was not looking for it!"
Reasons for self-chosen password	Memorability issue	132	23.66%	"Random is hard to remember if you need to login on another device." "I did not use the random password generator because I am afraid I will forget it!"
	Prefer to create their own passwords	64	11.47%	"I would prefer using a word i am more familiar with than any suggestions." "I'd rather use a unique password that I create."
	Didn't notice the nudge	24	4.30%	"I didn't realize I could." "Didn't notice the option."
	Trust issue	23	4.12%	"I don't believe in random password generator. May be the website hacks my details. So I'll be careful in this." "I don't trust that technology. I'd rather create my own password and then write it down."
	Noise	20	3.58%	"None" "My pet name"
	Security concerns	16	2.87%	"I don't feel safe using a generator that I have not used before." "Because it wasn't strong enough."
	Didn't care about the website	13	2.33%	"I didn't use it because I'm not going to be using this site again." "Because i do not plan to use this site so it's not relevant"
	Incongruous	13	2.33%	"I used the password manager because there are too many things sites that I use that need different passwords and I couldn't remember all of them." "I used it because its the safest way to create a password."
	The desire to reuse password	9	1.61%	"Because I usually keep one password to all..." "I just prefer to use the same password for stuff so that it is easier to remember."
	Lack of knowledge of password manager	3	0.54%	"I didn't know how to use it."

Table 5: Codebook: Reasons for adopting/not adopting a randomly generated password. As multiple codes were assigned to several comments, the summation of frequencies for each reason is more than the number of participants.

participants' comments, we could spot trends in user behavior. For instance, convenience and security were the most common reason participants adopted the randomly generated password, while memorability issues were participants' main reason for not using a random password generator. The purpose of random password generators is to provide a convenient method for creating secure passwords, which coincides with participants' reasons for using them. However, random passwords are complex and difficult to remember to prevent brute-force and guessing attacks [54]. Since random passwords are hard to remember, password generators are accompanied by password managers, which store the generated passwords. If password managers solve the issue of random password memorability, why do people reject using them? Based on participants' comments from our post-study questionnaire, people prefer to remember their passwords in order

to use them on different devices. One participant commented, "Random [password] is hard to remember if you need to login on another device." Browser-based and third-party password managers sync passwords across devices, ensuring users always have access to their passwords. Safari's nudge includes the message, "Safari created a strong password for this website—This password will be saved to your iCloud Keychain and will AutoFill on all your devices." This message informs users that they will have access to their passwords across devices that use iCloud Keychain. Chrome and Firefox, however, do not have messages explicitly stating that users will have access to their passwords across devices, which may be why people are hesitant to save their passwords. Participants' comments also expressed a distrust of password managers due to a lack of knowledge of the technology, which corroborates Fagan et al.'s study [12]. Safari was the most effective pass-

word manager likely because it explains the feature to remove doubt from users. Chrome and Firefox’s convenient, minimalist approach to nudging lacks a detailed explanation of their password manager’s features, leaving unanswered questions in people’s minds. A solution to the low adoption of password managers could be to improve their design by adding a more thorough explanation of their features. Doing so might educate users about the benefits of the technology, help build trust with users, and ultimately improve the adoption of password managers.

5.3 Limitations

Our study is categorized as a quasi-experiment because participants were not randomly assigned to each browser, but could sign up for one of the three groups. Thus, it is possible that participants’ behavior may be due to differences between Chrome, Firefox, and Safari users, rather than the differences between browser nudges. However, randomly assigning participants to each group posed its own issues: if participants were assigned to an unfamiliar browser, they may be more likely to (a) drop out since it is not installed or (b) notice the nudge more often since they aren’t familiar with the browser. These issues would affect users’ behavior and therefore the accuracy of our results. We also considered emulating each browser’s nudge on a single browser (e.g., Chrome); however, users who are familiar with the browser may notice the change in the browser’s nudge design, suspect our intentions, and alter their behavior accordingly. Therefore, we decided on a quasi-experiment design for this study.

There are some limitations from running our study on Amazon MTurk. First, our study had limited diversity because participants were all Amazon Mechanical Turk workers from the United States. MTurk workers are younger and more tech-savvy than the average population [42]. However, previous research implies that online privacy and security behavior studies can estimate the general population’s behavior despite this flaw [42]. Second, the Amazon MTurk platform’s prevalence of poor data quality has been increasing [26]. As a result, we used various countermeasures, such as validating participants’ MTurk IDs and putting a verification question to catch invalid study attempts. These countermeasures excluded invalid data from further analysis and prevented participants from taking our study more than once. However, it is possible that the nature of the study (single session/one device, no requirement to return) encouraged the use of the password manager more than longer-term scenarios. Also, it is possible that MTurk workers may encounter more account creation scenarios than most, leading to a higher adoption rate of randomly generated passwords.

Having the questionnaires and consent forms in English required participants to be fluent in English, and may have resulted in a language or cultural bias. Regarding questionnaire responses, like any self-reported data, they may be vulnerable

to a social desirability bias [13] and may differ from natural behavior due to privacy paradox [27]. To ensure participants answer honestly, the true intent of the study is not revealed until all tests and questionnaires have been completed. Initially, participants are told they are testing the registration system for a new website and are unaware of our goal to test the effectiveness of browser nudges. This allows us to test how participants would naturally create a new account for a website and helps eliminate social desirability bias.

Some users in our study may have been making use of other password managers and/or random password generators. We analyzed participants’ passwords to determine if third-party software may have been used to generate passwords as an alternative to browsers’ built-in password generators. For this purpose, we check whether users typed or pasted their password in a password field. According to our data, 7.9% of Chrome participants, 4.8% of Firefox participants, and 1.7% of Safari participants used alternative methods to generate passwords and paste them into the password field while registering.

6 Conclusion

We conducted a user study on the nudges employed by the built-in password managers in Chrome, Firefox, and Safari by using a mock e-commerce website. We investigated the effectiveness of the nudges in terms of their ability to encourage users to adopt a randomly generated password while registering. Moreover, we investigated whether a number of factors influence users’ toward adopting a randomly generated password. Our findings indicate that Safari works better in terms of its ability to encourage people to use a random password generator. Notably, participants in the Safari group believed that the nudge employed by Safari is more noticeable. Some reasons for Safari’s nudge being more noticeable are that (a) Safari is using a default nudge, which automatically populates the password field with a suggested password, (b) it uses color and a pop-up message, and (c) it implements interesting visual effects on the randomly generated password. We were surprised to find that implementing a strict password policy does not seem to influence participants to use a random password generator. Although one would assume selecting a random password is easier than creating a password that conforms to a 3C12 password policy, it would appear many people are still more comfortable creating their own passwords. Our results show that “default nudges” also work well for password managers, which is consistent with other studies suggesting that default nudges are the most effective nudge type across many fields [17, 19, 30].

Future work includes dissecting the reasons for Safari’s nudge performing better, to identify exactly which design elements are most impactful. This could be done by trying different variations of the nudge, where each implements only one of the design elements. It is possible that the default

aspect of the nudge is most important, or alternatively it could be due to the prominence of the nudge. While one of our findings was that users who noticed the nudge were more likely to accept a randomly generated password, future studies involving more prominent nudges should be aware of potential risks such as habituation. Additionally, some research suggests that personalizing nudges to match a user's decision-making behavior results in more impactful nudges [34]. However, implementing personalized nudges is a challenging endeavor that requires several phases [40]. This may be an interesting avenue for future work in password manager nudges. Since this study was conducted on Amazon Mturk, long-term studies with a non-crowdsourced population are needed. Also, the effectiveness of other forms of nudging [9, 17] for adoption of randomly generated passwords could be explored.

References

- [1] Rowe Adam. Study reveals average person has 100 passwords. <https://tech.co/password-managers/how-many-passwords-average-person>. Accessed: 2022-06-03.
- [2] Yusuf Albayram, John Liu, and Stivi Cangonj. Comparing the effectiveness of text-based and video-based delivery in motivating users to adopt a password manager. In *European Symposium on Usable Security 2021*, pages 89–104, 2021.
- [3] Nora Alkaldi and Karen Renaud. Why do people adopt, or reject, smartphone password managers? In *European Workshop on Usable Security*, 2016.
- [4] Nora Alkaldi and Karen Renaud. Encouraging password manager adoption by meeting adopter self-determination needs. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.
- [5] Fahad Alodhyani, George Theodorakopoulos, and Philipp Reinecke. Password managers—it's all about trust and transparency. *Future Internet*, 12:189, 2020.
- [6] Sal Aurigemma, Thomas Mattson, and Lori Leonard. So much promise, so little use: What is stopping home end-users from using password manager applications? In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.
- [7] Jannatul Bake Billa, Anika Nawar, Md Maruf Hasan Shakil, and Amit Kumar Das. Passman: A new approach of password generation and management without storing. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE, 2019.
- [8] Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Frank Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (S&P)*, pages 553–567, 2012.
- [9] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019.
- [10] Sonia Chiasson, Paul C van Oorschot, and Robert Biddle. A usability study and critique of two password managers. In *USENIX Security Symposium*, volume 15, pages 1–16, 2006.
- [11] Molly Cooper, Yair Levy, Ling Wang, and Laurie Dringus. Subject matter experts' feedback on a prototype development of an audio, visual, and haptic phishing email alert system. *Online Journal of Applied Knowledge Management*, 8(2):107–121, 2020.
- [12] Michael Fagan, Yusuf Albayram, Mohammad Khan, and Ross Buck. An investigation into users' considerations towards using password managers. *Human-centric Computing and Information Sciences*, 2017.
- [13] Robert J. Fisher. Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2):303–315, 1993.
- [14] Dinei Florencio and Cormac Herley. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web*, pages 657–666, 2007.
- [15] Shirley Gaw and Edward W. Felten. Password management strategies for online accounts. In *Proceedings of the Second Symposium on Usable privacy and Security*, pages 44–55, 2006.
- [16] David Halpern. *Inside the nudge unit: How small changes can make a big difference*. Random House, 2015.
- [17] Dennis Hummel and Alexander Maedche. How effective is nudging? a quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, 80, 2019.
- [18] Troy Hunt. Have i been pwned: Check if your email has been compromised in a data breach. <https://haveibeenpwned.com/>. Accessed: 2022-02-16.

- [19] Moritz Ingendahl, Dennis Hummel, Alexander Maedche, and Tobias Vogel. Who can be nudged? examining nudging effectiveness in the context of need for cognition and need for uniqueness. *Journal of Consumer Behaviour*, 20(2):324–336, 2021.
- [20] Iulia Ion, Rob Reeder, and Sunny Consolvo. No one can hack my mind: Comparing expert and non-expert security practices. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*, pages 327–346, 2015.
- [21] Blake Ives, Kenneth R. Walsh, and Helmut Schneider. The domino effect of password reuse. *Communications of the ACM*, 47(4):75–78, 2004.
- [22] Shouling Ji, Shukun Yang, Anupam Das, Xin Hu, and Raheem Beyah. Password correlation: Quantification, evaluation and application. In *Proceedings of the IEEE Conference on Computer Communications*, pages 1–9, 2017.
- [23] Shipi Kankane, Carlina DiRusso, and Christen Buckley. Can we nudge users toward better password management? an initial study. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.
- [24] Ambarish Karole, Nitesh Saxena, and Nicolas Christin. A comparative usability evaluation of traditional password managers. In *Proceedings of the 13th International Conference on Information Security and Cryptology*, 2010.
- [25] Patrick Gage Kelley, Saranga Komanduri, Michelle L. Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and L Julio. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (S&P)*, pages 523–537, 2012.
- [26] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, 8(4):614–629, 2020.
- [27] Spyros Kokolakis. Privacy attitudes and privacy behaviour. *Computer Security*, 64:122–134, 2017.
- [28] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.
- [29] Thomas C. Leonard, Richard H. Thaler, and Cass R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*, 2008.
- [30] Yiling Lin, Magda Osman, and Richard Ashcroft. Nudge: concept, effectiveness, and ethics. *Basic and Applied Social Psychology*, 39(6):293–306, 2017.
- [31] Sanam Ghorbani Lyastani, Michael Schilling, Sascha Fahl, Michael Backes, and Sven Bugiel. Better managed than memorized? studying the impact of managers on password strength and reuse. In *Proceedings of the 27th USENIX Conference on Security Symposium*, pages 203–220, 2018.
- [32] William Melicher, Blase Ur, Sean M. Segreti, Saranga Komanduri, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Fast, lean, and accurate: Modeling password guessability using neural networks. In *Proceedings of the 25th USENIX Security Symposium*, pages 175–191, 2016.
- [33] Burak Merdenyan and Helen Petrie. Perceptions of risk, benefits and likelihood of undertaking password management behaviours: Four components. In *Human-Computer Interaction – INTERACT 2019*, pages 549–563. Springer International Publishing, 2019.
- [34] Stuart Mills. Personalized nudging. *Behavioural Public Policy*, 6(1):150–159, 2022.
- [35] Bijeeta Pal, Tal Daniel, Rahul Chatterjee, and Thomas Ristenpart. Beyond credential stuffing: Password similarity models using neural networks. In *IEEE Symposium on Security and Privacy*, pages 417–434, 2019.
- [36] Zach Parish, Connor Cushing, Shourya Aggarwal, Amirali Salehi-Abari, and Julie Thorpe. Password guessers under a microscope: An in-depth analysis to inform deployments. *International Journal of Information Security*, 2021.
- [37] Zach Parish, Amirali Salehi-Abari, and Julie Thorpe. A study on priming methods for graphical passwords. *Journal of Information Security and Applications*, 62:102913, 2021.
- [38] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let’s go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 295–310, 2017.
- [39] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why people (don’t) use password managers effectively. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security*, 2019.

- [40] Eyal Peer, Serge Egelman, Marian Harbach, Nathan Malkin, Arunesh Mathur, and Alisa Frik. Nudge me right: Personalizing online security nudges to people’s decision-making styles. *Computers in Human Behavior*, 109:106347, 2020.
- [41] HIRAK RAY, FLYNN WOLF, RAVI KUBER, and ADAM J. AVIV. Why older adults (don’t) use password managers. In *Proceedings of the 30th USENIX Security Symposium*, 2021.
- [42] ELISSA M. REDMILES, SEAN KROSS, and MICHELLE L. MAZUREK. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343, 2019.
- [43] KAREN RENAUD, VERENA ZIMMERMANN, JOSEPH MAGUIRE, and STEVE DRAPER. Lessons learned from evaluating eight password nudges in the wild. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2017)*, pages 25–37, 2017.
- [44] KAREN RENAUD and VERENA ZIMMERMANN. Nudging folks towards stronger password choices: providing certainty is the key. *Behavioural Public Policy*, 3(2):228–258, 2018.
- [45] SUNYOUNG SEILER-HWANG, PATRICIA ARIAS-CABARCOS, ANDRÉS MARÍN, FLORINA ALMENARES, DANIEL DÍAZ-SÁNCHEZ, and CHRISTIAN BECKER. "i don’t see why i would ever want to use it" analyzing the usability of popular smartphone password managers. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1937–1953, 2019.
- [46] JAMES SIMMONS, OUMAR DIALLO, SEAN OESCH, and SCOTT RUOTI. Systematization of password manager use cases and design paradigms. In *Annual Computer Security Applications Conference*, pages 528–540, 2021.
- [47] ELIZABETH STOBERT and ROBERT BIDDLE. A password manager that doesn’t remember passwords. In *Proceedings of the 2014 New Security Paradigms Workshop*, pages 39–52, 2014.
- [48] ELIZABETH STOBERT and ROBERT BIDDLE. The password life cycle. *ACM Transactions on Privacy and Security*, 21(3), 2018.
- [49] ELIZABETH STOBERT, TINA SAFAIE, HEATHER MOLYNEAUX, MOHAMMAD MANNAN, and AMR YOUSSEF. Bypass: Reconsidering the usability of password managers. In *International Conference on Security and Privacy in Communication Systems*, pages 446–466. Springer, 2020.
- [50] KURT THOMAS, JENNIFER PULLMAN, KEVIN YEO, ANANTH RAGHUNATHAN, PATRICK GAGE KELLEY, LUCA INVERNIZZI, BORBALA BENKO, TADEK PIETRASZEK, SARVAR PATEL, DAN BONEH, et al. Protecting accounts from credential stuffing with password breach alerting. In *28th USENIX Security Symposium (USENIX Security ’19)*, pages 1556–1571, 2019.
- [51] JULIE THORPE, MUATH AL-BADAWI, BRENT MACRAE, and AMIRALI SALEHI-ABARI. The presentation effect on graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2947–2950, 2014.
- [52] BLASE UR, FUMIKO NOMA, JONATHAN BEES, SEAN M. SEGRETI, RICHARD SHAY, LUJO BAUER, NICOLAS CHRISTIN, and LORRIE FAITH CRANOR. I added ‘!’ at the end to make it secure: Observing password creation in the lab. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*, pages 123–140, 2015.
- [53] RAFAEL VERAS, CHRISTOPHER COLLINS, and JULIE THORPE. A large-scale analysis of the semantic password model and linguistic patterns in passwords. *ACM Transactions on Privacy and Security*, 24(3), 2021.
- [54] DING WANG, ZIJIAN ZHANG, PING WANG, JEFF YAN, and XINYI HUANG. Targeted online password guessing: An underestimated threat. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1242–1254, 2016.
- [55] YANG WANG, PEDRO LEON, KEVIN SCOTT, XIAOXUAN CHEN, ALESSANDRO ACQUISTI, and LORRIE CRANOR. Privacy nudges for social media: an exploratory facebook study. In *Proceedings of the 22nd International Conference on World Wide Web*, 2013.
- [56] SHANNON WILLIAMS. Average person has 100 passwords - study. <https://securitybrief.co.nz/story/average-person-has-100-passwords-study>. Accessed: 2022-02-17.
- [57] IRYNA YEVSEYEVA, CHARLES MORISSET, and AAD VAN MOORSEL. Modeling and analysis of influence power for information security decisions. *Performance Evaluation*, 98:36–51, 2016.
- [58] VERENA ZIMMERMANN and KAREN RENAUD. The nudge puzzle: Matching nudge interventions to cybersecurity decisions. *ACM Transactions on Computer-Human Interaction*, 28(1), 2021.

Appendix A First Consent Form

Title of Research Study: Evaluating the Usability of the Registration/Login Process of an E-Commerce Website.

Introduction: You are invited to participate in a research study entitled Evaluating the Usability of the Registration/Login Process of an E-Commerce Website. Please read the information about the study presented in this form. The form describes the study's procedures, risks and benefits that you should know before you decide if you would like to take part. You should take as much time as you need to make your decision. You should ask the Principal Investigator (PI) or study team to explain anything that you do not understand and make sure that all of your questions have been answered before signing this consent form. Before you make your decision, feel free to talk about this study with anyone you wish including your friends and family. Participation in this study is voluntary. This study has been reviewed by the University of Ontario Institute of Technology (Ontario Tech University) Research Ethics Board 16544 on October 17, 2021.

Purpose: You have been invited to participate in this study because your participation can contribute to our evaluation of the usability of the registration/login process for our website.

Procedure: This study will take about 5 minutes, and you will be provided with \$0.60 USD upon completion of our study and survey. The study tasks include:

- You will be taken to the Registration page of our website and asked to register.
- We will ask you some demographic questions.
- You will be taken to the Login page of our website and asked to login.
- We will ask you some additional questions and to provide feedback.

Potential Benefits: You will be compensated with \$0.60 USD for participation and completion of your task and survey.

Potential Risk or Discomforts: There are no known or anticipated risks to you from participating in this study.

Use and Storage of Data: The data includes demographic information and feedback (i.e., gender, age, and education level). All the data is anonymous and the data doesn't include any personal, confidential, or valuable information.

Confidentiality: Your MTurk ID will be kept confidential. Collected data will be anonymous and it will not include any information that reveals your identity. Please note that to maintain your registration experience on our website, you will be asked to enter an email address, but this information will not be stored. Your privacy shall be respected. No information about your identity will be shared or published without your permission, unless required by law. Confidentiality will be provided to the fullest extent possible by law, professional practice, and ethical codes of conduct. Please note that confidentiality cannot be guaranteed while data is

in transit over the Internet. This research study includes the collection of demographic data which will be aggregated in an effort to protect your anonymity. Despite best efforts it is possible that your identity can be determined even when data is aggregated.

Voluntary Participation: Your participation in this study is voluntary. You may also decide not to be in this study, or to leave the study at any time. You will be given information that is relevant to your decision to continue or withdraw from participation. You may refuse to answer any question you do not want to answer.

Right to Withdraw: If you withdraw from the research project prior to your final submission and the end of the study tasks, any data will be removed from the study and you do not need to offer any reason for making this request. You can withdraw within one week of submitting your data by contacting the researchers directly by email.

Compensation, Reimbursement, Incentives: You will be compensated with \$0.60 USD for participation and completion of your task and survey. You won't be compensated if you do not submit your data at the end of the study.

Debriefing and Dissemination of Results: If you are interested in learning of the results, please contact Samira Zibaei at Samira.Zibaei@ontariotechu.net.

Participant Rights and Concerns: Please read this consent form carefully and feel free to ask the researcher any questions that you might have about the study. If you have any questions about your rights as a participant in this study, complaints, or adverse events, please contact the Research Ethics Office at (905) 721-8668 ext. 3693 or at researchethics@ontariotechu.ca. If you have any questions concerning the research study or experience any discomfort related to the study, please contact the researcher Samira Zibaei at Samira.Zibaei@ontariotechu.net.

Secondary Use of Research for Future Research Purposes: Please note, if you agree to participate (and do not withdraw from the study), your anonymous data may also be used for future studies relating to our research.

Consent to Participate:

1. I have read the consent form and understand the study being described.
2. I have had an opportunity to ask questions and my questions have been answered. I am free to ask questions about the study in the future.
3. I freely consent to participate in the research study, understanding that I may discontinue participation at any time without penalty.
4. I understand the possible need for secondary research uses of my research data for future research use and provide consent for the use of my data to be used in future studies.

I agree

Appendix B Post-Registration Questionnaire

1. What gender do you identify as?
 - Female
 - Male
 - Prefer not to answer
2. What is your age?
 - 18 – 25 years old
 - 26 – 35 years old
 - 36 – 50 years old
 - 50 +
 - Prefer not to answer
3. What is the highest degree or level of education you have completed?
 - High school
 - Bachelor's degree
 - Master's degree
 - PhD or higher
 - Prefer not to answer
4. What is your first language (i.e., mother tongue)?
 - English
 - French
 - Other: _____
 - Prefer not to answer
5. What is your primary area of study or work?
 - Social Sciences and Humanities
 - Science
 - Health Science
 - Engineering and Applied Science
 - Energy and Nuclear Science
 - Education
 - Business and IT
 - Prefer not to answer

Appendix C Post-Study Questionnaire

1. How often do you use the browser you used in this study?
 - I use this browser daily
 - I use this browser weekly
 - I use this browser monthly
 - I use this browser a few times per year
 - I have never used this browser before today
 - Prefer not to answer
2. Have you ever used a password manager before registering on our website today?
 - Yes
 - No
 - Prefer not to answer
3. Have you ever used a random password generator before registering on our website today?
 - Yes
 - No
 - Prefer not to answer
4. Did you notice the recommendation to use a random password while registering on our website?
 - Yes
 - No
 - Prefer not to answer
5. Please select "Seven" from the following list.
 - 1
 - 5
 - 7
 - 3
6. Can you describe the reason why you used/did not use the random password generator?
Answer: _____
7. We are interested in any other comments you might have concerning your experience during registration. Please write any thoughts you'd like to share with us.
Answer: _____

Appendix D Second Consent Form

Title of Research Study: A Study of Nudging to Encourage Random Password Generation

Introduction: You are participating in this research study, and you were asked to evaluate the registration and login process of our proposed E-commerce website. However, this research is studying whether your web browser encourages use of generated passwords and storing them in your browser's password manager. Participation in this study is voluntary, and if you prefer not to submit at this step, your data is automatically withdrawn.

This study has been reviewed by the University of Ontario Institute of Technology (Ontario Tech University) Research Ethics Board 16544 on October 17, 2021.

Purpose: The actual purpose of this study is to test the efficacy of web browser nudges, which try to encourage you as a user to use a random password generator while you register on a new website. Using a randomly generated password and storing it in a password manager is considered a more secure strategy than reusing passwords (even partially) across accounts. Be aware that this strategy is recommended for many web accounts (e.g., e-commerce sites), but not for sensitive accounts (e.g., banking and email). For more information about password managers, please see: <https://cyber.gc.ca/en/guidance/password-managers-security-itsap30025>.

Potential Benefits: You will be compensated with \$0.60 USD for participation and completion of your task and survey. By reading the above information, you may have learned about how to improve your password security by using password generators and password managers.

Potential Risk or Discomforts: There are no known or anticipated risks to you from participating in this study.

Use and Storage of Data: The data includes whether you used the random password generator or not, the password you entered, demographic information, and feedback (i.e., gender, age, and education level). All the data is anonymous and the data doesn't include any personal, confidential, or valuable information. Data will be anonymous and your e-mail address will not be saved in our database.

Confidentiality: Your MTurk ID will be kept confidential and deleted upon completion of the study. Collected data will be anonymous and it will not include any information that reveals your identity. Your privacy shall be respected. No information about your identity will be shared or published without your permission, unless required by law. Confidentiality will be provided to the fullest extent possible by law, professional practice, and ethical codes of conduct. Please note that confidentiality cannot be guaranteed while data is in transit over the Internet. This research study includes the

collection of demographic data which will be aggregated in an effort to protect your anonymity. Despite best efforts it is possible that your identity can be determined even when data is aggregated.

Voluntary Participation: Your participation in this study is voluntary. You may choose to submit your information next in order to complete the study, or withdraw by simply exiting the session.

Right to Withdraw: You may withdraw from the research project by not submitting your data next. Also for the next week, you can still withdraw by contacting the researchers by email. Any data will be removed from the study and you do not need to offer any reason for making this request.

Compensation, Reimbursement, Incentives: You will be compensated with \$0.60 USD for participation and completion of your task and survey. You won't be compensated if you do not submit next and your collected data will be deleted permanently from our database.

Debriefing and Dissemination of Results: If you are interested in learning of the results, please contact Samira Zibaei at Samira.Zibaei@ontariotechu.net.

Participant Rights and Concerns: Please read this consent form carefully and feel free to ask the researcher any questions that you might have about the study. If you have any questions about your rights as a participant in this study, complaints, or adverse events, please contact the Research Ethics Office at (905) 721-8668 ext. 3693 or at researchethics@ontariotechu.ca. If you have any questions concerning the research study or experience any discomfort related to the study, please contact the researcher Samira Zibaei at Samira.Zibaei@ontariotechu.net.

Secondary Use of Research for Future Research Purposes: Please note, if you agree to participate (and do not withdraw from the study), your anonymous data may also be used for future studies relating to our research.

Consent to Participate:

1. I have read the consent form and understand the study being described.
2. I have had an opportunity to ask questions and my questions have been answered. I am free to ask questions about the study in the future.
3. I freely consent to participate in the research study, understanding that I may discontinue participation at any time without penalty.
4. I understand the possible need for secondary research uses of my research data for future research use and provide consent for the use of my data to be used in future studies.

I agree

Let The Right One In: Attestation as a Usable CAPTCHA Alternative

Tara Whalen
Cloudflare Inc.

Thibault Meunier
Cloudflare Inc.

Mrudula Kodali
Cloudflare Inc.

Alex Davidson
Brave

Marwan Fayed
Cloudflare, Inc.

Armando Faz-Hernández
Cloudflare Inc.

Watson Ladd
Sealance Corp.

Deepak Maram
Cornell Tech

Nick Sullivan
Cloudflare Inc.

Benedikt Christoph Wolters
Cloudflare Inc.

Maxime Guerreiro
Cloudflare Inc.

Andrew Galloni
Cloudflare Inc.

Abstract

CAPTCHAs are necessary to protect websites from bots and malicious crawlers, yet are increasingly solvable by automated systems. This has led to more challenging tests that require greater human effort and cultural knowledge; they may prevent bots effectively but sacrifice usability and discourage the human users they are meant to admit. We propose a new class of challenge: a Cryptographic Attestation of Personhood (CAP) as the foundation of a usable, pro-privacy alternative. Our challenge is constructed using the open Web Authentication API (WebAuthn) that is supported in most browsers. We evaluated the CAP challenge through a public demo, with an accompanying user survey. Our evaluation indicates that CAP has a strong likelihood of adoption by users who possess the necessary hardware, showing good results for effectiveness and efficiency as well as a strong expressed preference for using CAP over traditional CAPTCHA solutions. In addition to demonstrating a mechanism for more usable challenge tests, we identify some areas for improvement for the WebAuthn user experience, and reflect on the difficult usable privacy problems in this domain and how they might be mitigated.

1 Introduction

In a CAPTCHA challenge, a client is presented with a human-targeted puzzle requiring an interaction that no algorithm should be able to provide. A puzzle solved correctly is understood to be a puzzle solved by a human.

In practice, the association between puzzle and person has

been broken by advancements in machine learning and artificial intelligence techniques that solve CAPTCHAs with high degrees of accuracy [39]. In response, new CAPTCHAs emerge with increasingly specific (or challenging) signals and characteristics to distinguish human users from bots. The natural consequence of puzzles that focus on very specific traits of “humanness” is a set of laborious tests that can be solved by a decreasing number of humans [22]. This creates a cycle of increasing user frustration.

How, then, can the burden of proof that a client is not a bot be reduced for the human user? One approach reduces the number of challenge-response tests by extensive server-side user behaviour modeling and analysis [26]. This is accomplished with the use of cookies to track and profile users, alongside automated tests such as canvas rendering [35]. These tools are used to fingerprint client behaviour at the cost of privacy.

Alternatively, we can revisit the question: How can a human prove that they are not a program? The motivation to do so stems from two observations. First, CAPTCHA challenges are fundamentally connected to a design [38] born in a decades-old Internet ecosystem, more culturally homogeneous and with less capable hardware and software. Second, today’s Internet infrastructure consists of, indeed relies upon, cryptographic constructs and systems. Remote attestation is one such bedrock of increasing importance to Internet systems and protocols [31]. In cryptography, remote attestation involves supplying evidence to an appraiser over a network, in support of a claim about the properties of a target [10].

In this paper we explore remote attestation as the foundation for a new class of challenge-response that can attest to the presence of a person. Rather than identify tasks that bots are incapable of completing, our focus shifts to tasks that a human can complete. Thus we ask the following question: *what is the smallest task that separates a human from a bot?* The answer, we claim, is a *physical* interaction such as a touch or a look. We note that support for such interactions

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

is increasingly ubiquitous on even lower-end mobile phones and computers via *privacy-preserving* biometric sensors, and is additionally supported by USB and NFC hardware keys. These are authenticator devices that “attest” to the interaction. Their functionality is also widely accessible via the World Wide Web Consortium’s (W3C) Web Authentication API (WebAuthn) [23].

Motivated by these observations, we architected and implemented a challenge in which the response is a cryptographic and WebAuthn-compliant attestation. We note that WebAuthn functionality is increasingly available on the lower-end devices that are the primary means for connecting to the Internet for most of the world’s population [9, 32, 20]. Our design is guided by the W3C guidelines and requirements for replacing CAPTCHAs, with privacy-preservation made an explicit priority [22].

We evaluated the feasibility of our WebAuthn-based challenge, called the “Cryptographic Attestation of Personhood” (CAP), through a set of user studies. After a pilot study using USB security keys, we created a demo compatible with a wider range of hardware and released it for public testing and feedback. We found that, given the required hardware and browser environment, users were able to quickly and easily pass a challenge, and most said they were likely to use CAP if it were available as an option.

Our results were drawn from an analysis of 1896 sessions in which users tried our CAP challenge, testing it with their own hardware; a subset of these users (n=93) provided additional details via a survey. In our demo evaluation, a large proportion of users were able to complete the CAP challenge, with approximately half of the attempts being successfully passed. Task completion was quick, at 10.6 seconds—approximately half the time needed to solve a picture-selection CAPTCHA. Our survey results indicated that the majority of respondents (75%) were likely to use CAP when possible.

Overall, CAP shows great promise as a usable CAPTCHA alternative, although there are some barriers to adoption. These include privacy concerns (which are a challenge for WebAuthn in general); the difficulty of clear communication; and inconsistencies across different browsing environments.

2 Background: Users vs. CAPTCHAs

CAPTCHAs have been routinely identified as problematic by both researchers and others in the wider technical community [22]. The first CAPTCHA defined the puzzle as a challenge-response mechanism that involves a user and a challenge provider [38] (most often a content server or service). The puzzle has one requirement: a correct solution should assure the provider of an interaction that only a human could have performed. Interactions that could be completed by bots and algorithms are excluded by definition. The

definition and intention notwithstanding, automated solvers have since emerged [34], prompting increasingly complex CAPTCHAs that place ever-higher demands on people to solve them.

One major problem is accessibility. CAPTCHA tasks frequently involve visual identification, which makes them unusable by users with visual impairments [22]. Audio recognition tasks [15] may be an improvement for some user needs but still demand a heavy task burden. In addition, many task types rely on language or cultural knowledge that is far from universal. This can create barriers—for example, if taxicabs in the images look nothing like those in the user’s country [13] or for users who have never seen a fire hydrant. Mathematics, seemingly universal, is a far from trivial type of challenge for many users [18].

Privacy is another area of concern. For example, reCAPTCHA v3 calculates an “adaptive risk analysis” to assess the likelihood that a site visitor is a human, and may refrain from presenting a task if there is high enough score [26]. These approaches rely on background data collection—the specific details of which are rarely made public [33, 19]. In this context, some loss of privacy may be unavoidable, despite being undesirable.

The reliability of CAPTCHAs increasingly suffers in response to AI algorithms that continue to improve. Levels of complexity have been added to tests in response, as well as server-side tracking and profiling mechanisms to reduce their appearance to users. Reliability is an important requirement as any test that insufficiently prevents a bot from solving it has little utility as a security mechanism in the Internet setting. This worsens, in turn, accessibility. Among audio CAPTCHAs, for example, the gap between human and robot performance has shrunk dramatically, with bots reporting higher scores than humans [37, 2, 36].

In contrast to the available set of CAPTCHAs, our hardware challenge establishes proof of personhood with no cognitive burden and relies instead on a minimal set of possible physical interactions. The criteria, the components, and overall architecture are presented in the next section.

3 A WebAuthn challenge architecture

In this section, we describe a challenge platform with the Web Authentication standard’s API for attestation [23]. The platform is intended to be easily deployable so that smaller service providers can benefit.

3.1 Design requirements

Our design is guided by work at the W3C [22] and the experience with CAPTCHAs at Cloudflare, a service that provides

security features, including bot management, for a large proportion of the Internet [1]. Based on these, we believe any proof of a person attached to a device must meet the following goals:

1. **Ephemerality:** Solutions cannot be precomputed.
2. **Browser-based:** The challenge task must work in the browser without client modifications.
3. **Usability:** Internet-using humans should be able to prove their proximity to the device with minimal burden.
4. **Integrity:** The task has no solution without a human, otherwise the task fails to ensure security.

Standard CAPTCHAs clearly adhere to the two criteria of being *ephemeral* and *browser-based*. Each puzzle is randomly generated, and usually consists of a visual or audio challenge that can be displayed in an Internet browser. However, CAPTCHAs often fail to adhere to *usability* and *integrity*, as previously discussed.

In response to the diminishing *integrity* of CAPTCHAs, tools such as reCAPTCHA v3 [26] use sophisticated server-side modelling of client behaviour and anomaly detection. In some cases, this may preserve a degree of usability, but transforms the independent presentations of a challenge across websites into a connected web of user tracking, and motivates an additional requirement:

5. **Privacy:** Tests and challenges should reveal no information about users, nor be substitute identifiers.

We note that *privacy* is one attribute in which CAPTCHAs excel if executed in isolation. Absent the extra analytics pipelines that are, or can be, built on top of them, there is no information to tie a puzzle solution to an identity. Given the Internet context that we are operating in, the main privacy considerations that we examine in this work relate to ensuring that user identities are never revealed. In addition, we regard as unacceptable any challenge framework that can track users across visits. Even in situations where a user's identity is never directly revealed, the presence of such tracking potential may be used to identify the user via other means.

3.2 A challenge that trusts cryptographic attestation of human signals

We propose that one simple task that can differentiate a human from a bot is a *physical* interaction. Interactions may include biometric verification of a fingerprint or face, or a registering a touch on a secure hardware key. In this context an interaction challenge is deferred to a trusted platform to correctly and cryptographically *attest* to some attribute or action. This idea is the bedrock of trusted computing platforms.

Internet browsers have recently acquired the interfaces needed

to support cryptographic attestation, which are exposed via the W3C's Web Authentication standard (WebAuthn) [23]. The WebAuthn protocol is supported in all major Internet browsers [12]. It is also supported by many authenticator devices, including FIDO-supporting touch hardware keys, as well as biometric sensors increasingly available in Android, iOS, macOS, and Windows devices.

The WebAuthn protocol consists of two information flows, one for registration and another for authentication. The authentication flow is used to log in to an account without a password after an account has been created or registered. For our purposes, the authentication flow is ignored, thus there is no account against which to authenticate. Our design *relies solely on the attestation flow*. The attestation flow is similar to the registration flow (see Figure 1), but omits information that would bind an account to a user, such as an email address or name. Since there is no account-related information, there is no relationship between an account and a user for the attestation to expose.

We instead isolate the cryptographic attestations from within the WebAuthn framework's standard registration flow, as depicted in Figure 1. The standard flow has three high-level stages: (i) A server first requests an attestation challenge from a client in response to a username; (ii) the client then requests a credential from the WebAuthn-supported device, for which a person must take an action; (iii) the client receives a credential containing a proof of the action (usually, an attestation in the form of a digital signature), and sends it to the server, where the attestation is verified and stored.

Our changes omit the first and last stages of the WebAuthn standard registration flow. Figure 1 shows the standard flow, with greyed boxes depicting our omissions. The standard registration process is initiated by a username. Instead, our challenge is initiated by the server, which requests an attestation from the client without being prompted by a username. Note that this invocation is otherwise a standard WebAuthn registration interaction. During the last stage, the public keys are discarded once the attestation is verified, in contrast to their being stored after registration. These omissions preserve the integrity of the attestation itself.

In our challenge platform, the contents of the challenge string include a timestamp to limit the validity period of the response, together with information about the browser and user such as IP address, enabled Javascript APIs, etc. This prevents use of the response from any user agent in subsequent scripted interactions. Furthermore, successful completion of this or any other challenge to prove humanity only grants access if other aspects of the request are consistent with human interaction.

We emphasize that the cryptographic elements of the flow are untouched, so security aspects are preserved. Conversely, the

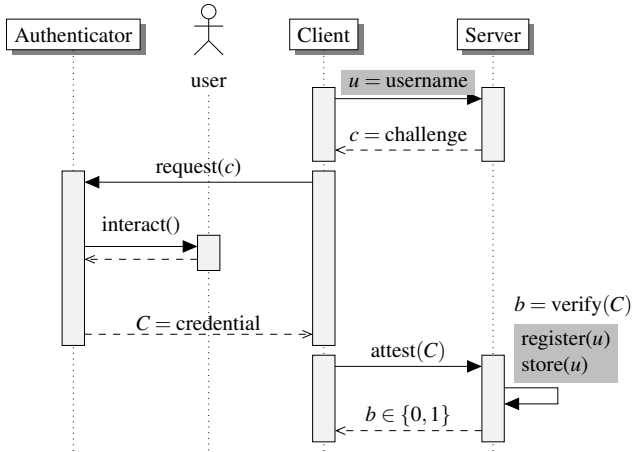


Figure 1: A high-level overview of the WebAuthn registration flow, with the minor omissions that enable our challenge: Portions encapsulated in grayed boxes are required for registration, and unnecessary for attestation verification. Ignoring the registration components preserves the privacy of users. Our challenge flow is otherwise identical to the standard.

omissions from the typical WebAuthn flow pertain *only* to user data. The deviations from the exact specification of the protocol leave the attestation and its verification untouched. Our hardware challenge is then characterized by the following properties:

- No user data is stored at the server.
- There are no user identifiers: users never specify a username, display names are replaced with generic text and unique IDs with random values that go unused.
- Attestations are provided directly by authenticators to ensure that they can be validated.

The availability of WebAuthn as a web API among Internet browsers enables us to build a human attestation system with the same *ephemerality* provided by a CAPTCHA. It is instructive to revisit the ability of our challenge to fulfill the remaining design goals, below and summarized in Table 1.

Usability Our WebAuthn challenge supports the same set of devices as does the W3C standard API, including Apple and Android biometric sensors and hardware security keys. The user gesture, such as presenting one’s face or touching a USB key, was expected to be easier to perform than a CAPTCHA interaction, and fits the profile of CAPTCHA alternatives envisioned in the recent W3C technical report highlighting CAPTCHA inaccessibility [22]. The usability assessment forms the bulk of this paper, in which we confirmed that this interaction was quick and easy in the majority of cases for

Table 1: Design requirements comparison between our approach and CAPTCHAs.

Challenge	Usability	Security	Privacy
CAPTCHA	✗	✓ ¹	✓ ²
Hardware attestation	✓ ³	✓	✓

¹ Reliant on continual upgrade of CAPTCHA challenges to prevent attacks from bots of ever-increasing capability.
² Only for those CAPTCHAs that do not use wider user browsing analytics to make inferences on the user’s humanness.
³ Usability is ensured for those that own applicable hardware.

which users had the necessary hardware and web-browsing environment.

The drawback of using this approach as a challenge is somewhat obvious: it is only available to those individuals with applicable hardware that implements the WebAuthn standard. As mentioned previously, WebAuthn is currently supported by a variety of devices including security keys, smartphones, and personal computing devices. With this in mind, we believe that it is reasonable to expect that WebAuthn, and Internet-based hardware attestation, will become more prevalent across the globe in the near future.

Integrity Our challenge should be difficult for bots to bypass. The integrity of the interaction is tied to the integrity of the WebAuthn standard, and devices’ ability to maintain keys securely. The attestation is generated by the device using a secret key that is embedded in hardware, tamper-resistant, and can only be extracted manually. Such a task is engineered to be difficult by design. Notably, the secret key is embedded in a batch during manufacture across a cohort of devices. In this manner, a batch key is shared, for example, among the same device model or devices manufactured in the same year. The key batching makes it possible for the challenge provider to only accept attestations from selected classes of devices. Similarly, attestations for device classes can be revoked if they fail some set of criteria, or if keys are known to have leaked. One weakness in touch devices *may* be that they can be circumvented by an automated physical device¹, against which biometric sensors are resilient.

Privacy Any viable challenge solution must reveal no details about the user identity, nor provide avenues for tracking the user across multiple websites or challenges. Our challenge reduces the registration to an attestation that is non-specific to the user. As a result, the attestation reveals no personally identifiable information. However, each attestation does reveal a hard-coded certificate associated with the device class. If the certificate were unique, it could be used to track a user’s

¹See, for example, <https://bert.org/2020/10/01/pressing-yubikeys/>.

attestation across multiple challenges and make inferences about that user’s browsing patterns.

Fortunately, the expected privacy impact incurred by revealing this certificate is very small, as described in the standardisation document [23, Section 14.4.1]. The standard recommends that these certificates (and their associated cryptographic keys) are batched and shared across multiple manufactured devices. The result is that each user belongs to a large *anonymity set*, as no given hardware device can be identified by the revealing of this certificate alone. For example, the FIDO UAF standard [4] requires that at least 100000 authenticator devices share the same attestation certificate in order to produce sufficiently large groups. (When considering mobile device classes we expect the anonymity set to be orders of magnitude larger.) The knowledge revealed to a provider is limited to the type of device and its batch or model.

Note that the WebAuthn challenge proposed here is built on an open standard. This is not a proprietary solution, but can be deployed by anyone needing to roll out a human challenge in their systems. They can learn from our evaluation (detailed below), and adapt and extend this solution in the ways that are most suitable for their own requirements.

4 Pilot study

We explored the possibilities of our hardware attestation mechanism in the context of Cloudflare, which provided opportunities for real-world evaluation as well as the potential of large-scale deployment as a challenge solution for a substantial number of websites. As a starting point, we carried out a pilot study to assess whether our idea had merit, particularly in terms of its usability.

We evaluated our hardware attestation mechanism with a usability experiment, assessing effectiveness, efficiency, user satisfaction and gathering feedback about the overall user experience. We compared this hardware key method, using Yubico YubiKey 5 Series security keys, against a standard CAPTCHA method currently protecting millions of sites: hCaptcha [21]. hCaptcha presents a 3x3 grid of pictures and prompts the user to select a subset matching specific criteria (e.g., “Please click each image containing a boat”).

In the experiment, 17 participants (Cloudflare employees) performed a simple webpage access task, where they visited two public webpages protected by hCaptcha or hardware attestation. For hCaptcha, participants identified objects from a set; for hardware attestation, they launched the proof-of-concept challenge and touched their YubiKey when prompted by their browser.

Each task was followed by a System Usability Scale (SUS) [7] satisfaction questionnaire. Participants were also provided

with a post-session questionnaire to measure preference between the two methods during a short, closing debrief.

Results of the usability experiment We instrumented the testing environment used by our participants to record *errors*, measure *success rate* (task completion), and *time-on-task*. Effectiveness was high for both conditions: all participants successfully completed both the YubiKey and hCaptcha conditions with no errors. Our participants rated the hardware challenge as easier to use with an SUS score of 77.1, and hCaptcha with an SUS score of 65.3. For SUS scores, “better products scor[e] in the high 70s to upper 80s”, and “[p]roducts with scores of less than 70. . . should be judged to be marginal at best” [5].

Measurements and analysis indicated significantly shorter completion times for the hardware challenge. A Wilcoxon signed-ranks test indicated a mean task time of 13.5 s for the hardware challenge, and 25.0 s for hCaptcha: $V=115$, $p < 0.001$. Note that the hardware challenge completion time is not just the time taken for the physical interaction with the key, but also includes the time taken to read and respond to informational pop-up messages spawned in-browser by the WebAuthn flow.

15/17 participants (88%) preferred the hardware challenge, with only one participant preferring hCaptcha and another having no preference. Participants who preferred the YubiKey expressed frustrations with CAPTCHAs and commented on the ease and speed of the YubiKeys. The two participants who did not prefer the YubiKey voiced concern and fear about security and privacy. Similar concerns were shared by some participants who favoured the hardware key. Participant feedback also identified wider user-communication challenges with browser prompts and messaging, where the information presented was viewed as uninformative or confusing.

5 Evaluation: Public demo study

The results of the preliminary user study indicated that our proposed solution was promising enough to develop further. We therefore developed a public demo of our “Cryptographic Attestation of Personhood” (CAP), which we deployed at Cloudflare for wider evaluation. Unlike our pilot study, which was limited to YubiKeys, the challenge on the demo site could be passed using a wide variety of hardware, such as biometric readers (e.g., Face ID and Touch ID) and multiple models of secure hardware keys. The site accepted any USB or NFC key certified by the FIDO Alliance, as long as it had no known security issues according to the FIDO Alliance Metadata service (MDS 3.0) [3]. A summary of supported hardware can be found in Table 2.

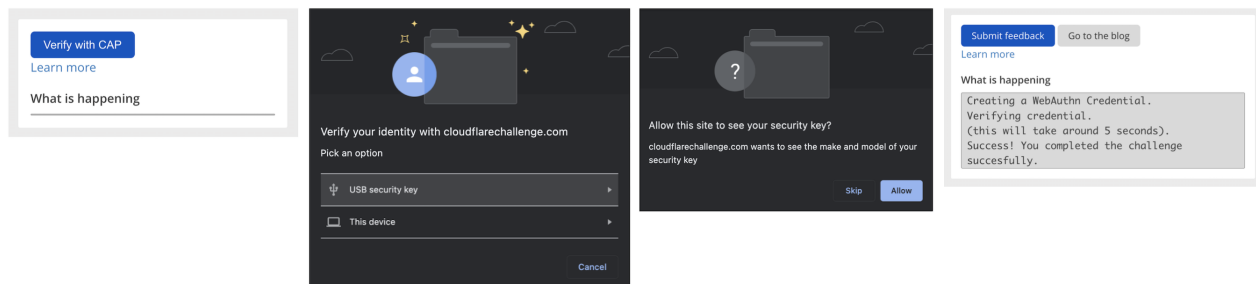


Figure 2: CAP demo: stages of CAP interaction, including browser WebAuthn prompts

5.1 Experiment details

We created a demo website where users could click on a button to “Verify with CAP”, which would prompt them to complete the WebAuthn challenge, whose main stages are illustrated in Figure 2. The start panel (with the “Verify” button) was displayed on a web page; this panel included a “Learn More” link, which brought the user to a separate “Frequently Asked Questions” page for assistance. The space below the button, labelled “What is happening” displayed progress through the verification process, until its conclusion: success or failure.

Once the user clicked on the button, additional pop-up windows were spawned by their browser as part of the standard WebAuthn process. Note that the specific text and design of these pop-ups is determined by the browser, and not by CAP. The examples in Figure 2 are from Chrome v98 on MacOS 12. The first browser pop-up prompts the user to “verify your identity” on the Cloudflare demo website, and gives them a list of WebAuthn authenticator options. In this example, the user can pick from the built-in Touch ID sensor on their Macbook, or they could use a portable USB key. The user selects their preferred option, and performs the user gesture (e.g., touches the fingerprint sensor). Because an attestation has been requested for this WebAuthn interaction—as this is an integral part of CAP—a prompt is displayed that asks whether the user wishes to disclose the make and model of their security key to the site. If the user selects “allow”, then the attestation is sent for verification; if it passes, then the user successfully passes the challenge (as shown in the final image in Figure 2). The user might also fail to pass, in which case an error message is shown in the panel. Technical details of the error are shown, and the user is informed “It seems there was an error completing the challenge! You can retry or share your feedback with us.” After each challenge attempt, users have the option of clicking “Retry” or “Submit Feedback”. The latter takes them to a user survey (described in 5.3 below).

This demo site was launched in conjunction with a blog post about CAP published on Cloudflare’s blog, which often dis-

cusses new features and experiments being run [28]. It was expected that this post would spark readers’ interest in trying out CAP, which would provide us with useful information. The blog post explained how the underlying WebAuthn technology worked, at both a non-expert and a more technical level for those who might prefer such details. The post included information about privacy considerations, which we anticipated would be of concern to users (and had been demonstrated in our preliminary study). The privacy explanation highlighted the size of the anonymity set (e.g., your key is indistinguishable from a large batch of others) and stressed how WebAuthn strictly limits what is sent to the server—for example, biometric data never leaves the device. The blog post concluded with a link to the demo site, and invited people to try out this experimental feature. In a later expanded version of the demo, with an associated blog post [14], when a user completed an attempt at a CAP verification, they had the option of giving feedback through a survey. This provided richer information on what they liked and disliked, general concerns, and suggestions for improvement. We evaluated the demo version of CAP through a combination of data logged from online users and feedback gathered from the online survey. For each interaction with the CAP demo, timing and error data was logged.

Note that we adopted a minimal data collection approach for the data logging of these interactions. Because we were concerned that users might be uneasy about disclosing information when testing this new feature (despite the protections being provided), particularly given privacy sensitivities (e.g., biometrics), we strongly limited what we stored. This meant that we did not collect details such as browser user agent strings and did not store any information about the authenticator (such as make or model); this attestation information was not logged.

Ethics Institutional review boards (IRBs) are uncommon in most workplaces, including ours. Nonetheless, care was taken to follow appropriate experimental procedures throughout (e.g., obtaining user consent for participation and data collection). No identifying information was logged in the interaction

phase, only timing and error data related to each stage of the CAP process. (This also means that duplicates may occur in our dataset, since repeat visits could not be identified.) For the survey, respondents provided explicit consent and were not required to provide any identifying information. They were permitted to provide an email address if they wanted to be contacted for further studies; they also were given the separate option of providing details on their environment (such as their browser’s user agent). They were also asked whether they consented to having their responses quoted, without attribution, in research publications. No participant compensation was offered.

5.2 Logged Interaction Data

We analyzed 1896 user sessions, collected over eight days. A single session was defined as any instance in which a user clicked on the “Verify with CAP” button at least once, which ended in a failed or successful verification, and include multiple attempts at verification (if any) within the same session.

5.2.1 Results

Completion Time For cases in which a person successfully validated with CAP at any point in a session, the mean completion time—from button click to completed validation—was 10.6 s. In the case of a failed validation, the mean time was 2.8 s. Failure is faster than success because the process terminated earlier without completing further steps; note that this also includes cases in which there is no further user response after the button click, which leads to a failure upon timeout (whose duration is environment-dependent).

For comparison purposes, we analyzed the time taken to complete real-world hCaptcha interactions (which could be from bots or humans), based on a sample of 8262 interactions recorded in Cloudflare’s logs. (hCaptcha uses an object identification challenge involving a 3x3 picture grid [21].) The mean hCaptcha solving time was 24.5 s, over twice the time of a successful CAP challenge; this timing difference is statistically significant (Wilcoxon rank-sum test: $W = 1476154$, $p < 0.001$).

Success Rate Out of 1896 sessions, 919 (48.5%) included a successful validation, with the majority of these (818, 43%) having no errors. (Recall that a person could retry multiple times per session.) In most cases, people tried only once: in the 1078 sessions with at least one failed attempt, only 24% had more than one failed attempt. (Note that we do not have any details about users’ environments in this dataset, owing to our minimal data collection in this part of the experiment.)

5.3 User Feedback: Survey

When a person completed a CAP validation attempt, they were given the option of completing a survey to provide additional feedback. This survey was deliberately kept brief, to encourage people to complete it, and focused on the key elements we wished to measure. We collected 93 survey responses during our evaluation period.

Likert scales The first set of questions asked for responses to 5-point Likert scales (“strongly agree” to “strongly disagree”):

- I am likely to use this when possible (I have a security key/biometric sensor)
- Assuming I have what I need, I prefer using this instead of a CAPTCHA
- It’s frustrating how often I have to prove I’m a human
- I feel confident that this preserves my privacy

The Likert scale responses are shown in Figure 3.

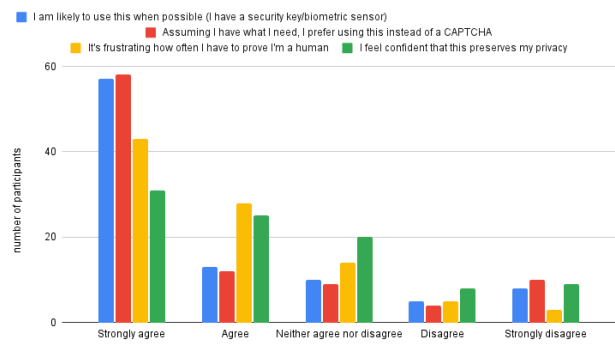


Figure 3: Results from 5-point Likert questions in CAP survey

The majority of respondents indicated they were likely to use CAP when their hardware allowed this option: 70 (75%) agreed or strongly agreed. Similarly, 70 respondents (75%) said they preferred CAP to a CAPTCHA (agreeing or strongly agreeing). Respondents indicated a high level of annoyance with having to complete human challenges, with 71 respondents (76%) agreeing or strongly agreeing that it was frustrating to do this task often. On the question of privacy, responses were more mixed. Respondents had some confidence in CAP’s privacy protections, although this was not as high as for the other items: 56 respondents (60%) agreed or strongly agreed with the statement “I feel confident that this preserves my privacy”; a further 20 (22%) neither agreed nor disagreed.

Free-form responses Respondents could provide free-form responses to four further questions: (i) What do you like

the most about this? (ii) What one thing would you change about how this works? (iii) If you have any accessibility needs, please let us know how well or poorly this caters to those needs, and (iv) If there's anything else you'd like us to know, please tell us here.

If desired, the respondent could send information about their environment: browser user agent; hardware device issuer; attestation format and type. Additionally, we collected the time taken for their most recent verification attempt and the number of errors encountered during their session. The free-form responses provided us with greater detail on what users liked and disliked about CAP. These were manually coded, which involved three researchers collectively identifying a set of initial themes, then coding independently and finally comparing results to achieve consensus.

The most commonly cited strengths ("What do you like the most about this?") were ease of use; speed; and improvement over other types of challenges (e.g., traditional picture-selection CAPTCHAs):

- "Honestly, it's quite fast. Works great, while proving the same thing that regular captchas do" [P6]
- "this is much much quicker than selecting all the buses....and trucks..." [P15]
- "Passed the challenge with just my fingerprint. Very convenient." [P43]
- "Easy as ABC. Love it!" [P54]

There were a number of suggestions that people had about how to improve CAP, primarily around clarifying the communication; preventing errors and failures; and reducing UI pop-ups. On the theme of communication, respondents recommended improvements in explaining some aspects of CAP, primarily the privacy protections:

- "Maybe making it clearer that the model of your key doesn't go out to the internet?" [P6]
- "Maybe add some explanation of how this works, what information do you guys collect during this process"[P21]
- "will Cloudflare store my 2FA key?" [P39]
- "how is this not a unique identifier? and how are you gonna explain that this is not surveillance to 'the normie folks'?" [P91]

Others suggested a need for better explaining some of the WebAuthn process and components, which may be hard to understand:

- "The options that are available on Android can be overwhelming for a non-technical audience. Most people

won't know what a Yubikey is or understand that 'unlock with screen lock' means finger print sensor." [P37]

Some users had problems completing the CAP challenge because they did not have a compatible setup, so they wanted better support for their devices (e.g., "Make it work with Windows Hello PIN" [P3]).

Although we did not specifically evaluate the accessibility aspects of CAP in this phase of study, we did wish to solicit feedback from anyone with these user requirements. Three suggestions were provided: two for larger UI elements and one for improved contrast.

Finally, respondents could provide us with any additional comments. Again, there was a call for extended support (on more devices and browsers), particularly to avoid failed attempts; recommendations for clearer communication to users; and requests for removing inefficiencies (such as pop-ups) where possible.

- "Chrome Android requires few more steps to actually choose which authentication system to use (NFC, security key or fingerprint). It doesn't automatically save my preferences so that I don't have to choose again" [P43]

Some other people wanted to simply express satisfaction with our approach:

- "I hope every website on the internet adopts this method" [P22]
- "I do wonder how well this will work to prevent farms of Captcha solving bots [...] if you can truly prevent that or stop it, this will be an amazing alternative." [P6]

Environment and Completion Time Of the 93 survey respondents, 82 provided details about their environment (browser, security hardware) along with the number of errors encountered during their entire CAP session and the time taken for their most recent verification attempt. Based on User Agent String, 39 were on mobile devices and 43 on desktop; the most commonly-reported browser was Chrome (41), followed by Safari (17), Firefox (10), and Edge (9). In terms of errors and timing, 50% of these respondents (41) had no errors at all in their session; 27 (33%) had one error in their session, and the remaining 14 (17%) had two or three errors. This is similar to the distribution found in the larger set of logs described previously, although there is a slightly higher success rate in the survey respondents.

Task completion timing was recorded, but note this measured duration from the initial JavaScript load event until the verification attempt ended, while the task time in the log dataset previously discussed was measured only from when the user clicked the button, which is a much shorter set of events. We analyzed the log data to give the same baseline for comparison: for a successful attempt, the survey participants took

15.1 s (vs 15.7 s) and 8.9 s for a failed attempt (vs. 7.0 s); again, this is similar to the larger dataset.

6 Discussion

6.1 Availability and Ease of Use

For successful validation cases, the completion time is quick (half that of hCaptcha), with few errors, and has high perceived efficiency. However, this is not the situation for all users: as noted, about half of them were unable to validate. The main difference of note comes down to environment: the biggest obstacle was having (or using) the correct combination of security hardware, OS, and browser. A summary of supported hardware and browser combinations is shown in Table 2. Survey respondents reported problems with validation when using MacOS with non-Safari browsers, and on Android mobile outside of Chrome, along with a few users having Windows compatibility issues.

Table 2: Overview of hardware support (based on testing in this study)

Hardware	Browser support	WebAuthn support	Secure attestation ¹
macOS (11 onwards)	Safari	✓	✓
	Major browsers	✓	✗
iOS 15 devices	Major browsers	✓	✓
Windows Hello	Microsoft Edge	✓	✓
	Other browsers	✓	✗
Android mobile	Chrome	✓	✓
	Other browsers	✗	✗
Hardware keys in FIDO MDS (e.g., YubiKeys)	Major browsers	✓	✓

¹ Secure attestation refers to attestation formats [23] that allow validation with a global issuing certificate.

For example, a person with a MacBook equipped with Touch ID would need to use Safari with CAP in order for the attestation to work properly; if they tried with Chrome, it would fail, as the Apple attestation sent with the Touch ID platform authenticator for WebAuthn is only compatible with Apple’s browser (Safari). In some cases, the user might lack the necessary hardware, although this is becoming less of an issue given the deployment of built-in WebAuthn-compatible devices in mobile devices (e.g., Face ID), and the growing adoption of hardware security keys for multi-factor authentication [6].

As noted in the survey responses, the majority of respondents (75%) were likely to use CAP if they had the necessary hardware. These results suggest that CAP is a good solution in the

right circumstances: given the appropriate environment, users prefer it to traditional CAPTCHAs. However, CAP is best positioned as an alternative challenge method for those equipped to take advantage of it, rather than it being presented as the sole option, given the number of users for whom it would not be possible or practical to use for a human challenge.

6.2 Communication Challenges

Explaining functionality Although the majority of our survey participants stated that they were likely to use CAP when possible, and many commented on how easy it was to use, it is important to consider that CAP involves a number of elements that are likely to be unfamiliar to many users. This is an entirely new human challenge method, which does not resemble the more familiar puzzle-based tasks. WebAuthn is itself a fairly new technology as well, and even those who may be comfortable with WebAuthn may be confused by its application in this unusual way. Those trying the CAP demo had the opportunity to review a substantial blog post with explanations of the technology before they tried it out; this would not be the most common scenario in a real-world deployment. Users need to know what this new feature does, and whether or not they are equipped to use it, as well as any additional considerations (such as privacy, discussed below). This is a lot to convey in a limited user interface. We have used the results of the study to refine our design and to augment customer support materials to assist users; these additions will be evaluated and refined iteratively as we continue to test CAP in deployment, as discussed in Section 7.

CAP as novel WebAuthn application CAP leverages the capabilities of WebAuthn and extends its functionality into the human challenge space; this is a benefit, and could provide additional incentives for people to obtain and use hardware security keys in order to mitigate their frustrations with CAPTCHAs. However, there are always challenges with novel technology, and in the CAP scenario, WebAuthn is being used for quite a distinct purpose from its usual application. Most people using WebAuthn are doing so for *authentication* purposes, and elements such as browser messages are designed with that in mind. As one example, consider the pop-up example shown when describing the public demo, in Figure 2. Note the text used when prompting the user: “Verify your identity with [example.com]”. Often, this is what users are doing: verifying their identity as part of an authentication process, such as logging into their account. Because CAP does not include this component (as it never registers a user and does not handle credentials), this message does not properly describe what is about to happen. It is understandable that the WebAuthn browser designs prioritize the majority use case, but it would be helpful to accommodate other applications.

WebAuthn: inconsistent experiences Additionally, the design choices of CAP are only one part of the entire WebAuthn user experience; many of the messages displayed during user interaction are under the control of the browser, not CAP. If there is a confusing message displayed, or excessive popups, this also has an effect on the overall user experience. At best, one can anticipate and explain some of the confusing elements of the WebAuthn ecosystem. This is compounded by the number of different configurations that a person may be using: WebAuthn via Face ID on an iPhone using Safari is not identical to WebAuthn via Yubikey on a Windows laptop using Chrome. These are similar, but not identical, and the inconsistencies in these experiences can lead to a sub-optimal user experience: some may lead to a failed verification, while others might simply provide unclear information.

6.3 Privacy Considerations

The survey results indicated that not all users were confident in the privacy protections provided by CAP. While the overall sentiment was positive (with 60% of respondents expressing confidence), this shows an area where improvement is needed. Very few respondents (only four) who had low confidence in privacy provided any comments about this topic at all, so the source of their concerns is not clear. Two of these discussed privacy in the content of how communication might be improved, whereas the other two were more concerned about the actual data collection risks (i.e., what is the website collecting?). In one case, the participant was confused by the specific Firefox messaging that appears when attestation is requested: “Firefox displays a warning that the site ‘is requesting extended information about your security key, which may affect your privacy’. I wouldn’t necessarily trust this if I didn’t know for sure that the request was coming from Cloudflare (which, in general, as a user, one doesn’t).” [P72]. In other browsers, the message is different, despite it being for the same type of request: for example, as shown in Figure 2, Chrome v98 says that the site “wants to see the make and model of your security key”. This example shows the importance of communication, and also the stark differences that users can experience between different environments.

7 Enhancements and Future Work

The findings from our user studies have highlighted areas where CAP could be improved, along with some promising new directions. We have also identified some research questions that we will continue to explore.

7.1 Improving Communication

When we conducted our usability evaluations, we provided explanatory material (such as blog posts) that assisted users in learning about this new human challenge approach that

is enabled by secure hardware; this also explained the underlying technology and its privacy and security capabilities. This was workable for experiments, but is not realistic outside of this situation. In the more usual scenario, a user would be browsing the web and then encounter a human challenge, such as an interstitial page containing a CAP prompt. A first-time user would have no previous experience with this type of attestation challenge, and perhaps would have no previous experience with WebAuthn at all. They would need to know how they might pass this challenge, including whether or not they had the right hardware and environment to do so successfully. They might also wonder about the security and privacy risks associated with using secure hardware to pass this challenge. Note also that those with WebAuthn experience in its more common *authentication* situation might have specific expectations about CAP that are not true: for example, they may expect they need the same hardware device to pass a challenge on repeat visits to a particular website.

This situation presents many significant challenges for user communication, and we are continuing to work on solutions. We began by revising the visual elements of the CAP prompt panel, so that it gives a suggestion that this is a task you perform with secure hardware; the first version of our new design displays a graphic with a fingerprint (to suggest a biometric reader) plus a USB key. We are also developing new customer support materials, which might involve videos to demonstrate the technology and how to use it; this would be readily accessible from the challenge page, in context with the CAP prompt. Providing explanations for WebAuthn through richer interactions, such as video, is consistent with recommendations provided in recent research on WebAuthn adoption; this was shown to be beneficial for mitigating misconceptions (such as where biometrics are stored) [25]. We expect to iterate on our designs as we have begun to run small-scale tests in a production environment and can evaluate the results.

7.2 Privacy: Zero-Knowledge Proofs

We have continued to explore how we might improve the privacy story for CAP and WebAuthn. In an extension of this work, we investigated how one might disclose the minimum possible amount of information: not the make and model of the security key, but simply the proof that the key being used is trustworthy. We developed an in-browser zero-knowledge proof to provide this functionality [17]. In brief: instead of sending the signature, the client sends a proof that the signature was generated by a key on a server-provided list. Because only the proof is sent, the server learns only that the attestation exists, and not which hardware security key generated it. An efficient proving and verification system was developed for this scenario, which is currently being evaluated. Results to date demonstrate that a zero-knowledge proof can be generated in approximately 10 seconds, which is extremely

promising as an efficient, privacy-preserving solution.

Ideally, this solution could be integrated into the WebAuthn standard, as an attestation type, so that it could signal to browsers that sending this particular attestation type would not disclose the make and model of hardware key (given the underlying zero-knowledge proof). In that case, there would be no need for the consent pop-up that users must click through, as there is no disclosure in this case. Not only would this be a more robust privacy solution, but it would also make WebAuthn interactions more efficient and less confusing for users, for any instance in which attestations were used (which is not restricted to CAP).

7.3 Exploring Privacy Concerns

While one path of our ongoing privacy improvements involves zero-knowledge proofs, we would also like to explore what some of the underlying privacy concerns are that could impede the adoption of CAP. Our survey touched on this question, as we anticipated its importance, but as noted above, this was designed to be a short questionnaire that did not delve deeply into any one specific area—including privacy. However, given the complexity and persistence of privacy considerations of WebAuthn in general, we feel it would be valuable to deepen our understanding of this problem. There are many potential sources of unease, some of which may be unrelated to the human challenge itself. For example, a person might choose not to use CAP because they do not want to use a biometric reader, and their underlying discomfort may be due to the biometric component in itself, which would be the case in *any* online context (not just for CAP). A better understanding of these factors would help us determine how to improve designs for this specific application, as well as how to contribute to WebAuthn adoption more broadly. This would be informed by, and build on, ongoing user research in this domain (e.g., [25, 29, 27]).

7.4 Security Considerations

In designing new methods for attesting to personhood, we must be mindful of security issues that arise when *malicious* clients attempt to provide false proofs of humanity. In the following, we attempt to build an overview of the threat model and potential methods for calculating adversarial costs of providing false proofs. Valuable future work would establish a thorough security analysis of using such attestations widely before establishing a large-scale deployment of these technologies.

Threat model We can split the attack surface into the following two types of attacks:

- *Human-assisted*: These type of attacks involve an adversary proxying attestation requests to a real human being,

who provides the proof based on their own inherent characteristics and returns the proof to the adversary to be returned to the requester.

- *Automated*: These attacks involve constructing mechanisms (either physically or in software) that allow generating valid attestation proofs from hardware authenticators, without a real person interacting with them.

All challenges that attempt to provide attestation of humanity—including all CAPTCHAs and related technologies—are vulnerable to human-assisted attacks. This assumes, however, that a challenge that an adversary receives can always be forwarded to a different real person that can solve the challenge instead. Currently, it is an open problem whether forwarding of hardware attestations is possible, and to what extent that compares to existing challenge systems.

In addition, software-based challenges are vulnerable to automated attacks that involve no human participation. As mentioned previously, CAP authenticators that rely solely on touch (rather than biometric identification, such as Yubico Yubikeys) may be vulnerable to automated attacks that involve constructing physical devices that generate valid interactions with the device. It is more difficult to circumvent biometric authenticators; such biometrics have not yet been mimicked in a similar manner (again, assuming it is possible to forward hardware attestations).

Adversarial costs A common way of establishing the security of a human-based challenge system is identifying the cost of buying a single valid attestation. These attestations can be provided either by real humans (who are paid for solving each challenge), or by an adversary that controls a resource that is able to provide automated proofs. Generally speaking, vulnerability of a challenge mechanism to automated attacks is quite damaging, since it is likely that such proofs can be provided much more cheaply than those that require human assistance.

In the case of CAPTCHAs, various services are known to price a single solution of a standard Google reCAPTCHA at \$0.003². Therefore, even human-assisted challenges are very cheap to acquire solutions for. Hardware-based authenticators such as Yubico Yubikeys require an initial up-front cost of between \$45 and \$85.³ Assessing the cost of launching an automated attack on top of these authenticators would be a valuable task for future work, but is likely to involve another one-time cost of setting up the tools that are required for automation, plus the much lower cost of continued usage. CAP authenticators that rely on biometrics are likely to involve much higher costs. Firstly, devices such as smartphones and

²According to <https://www.f5.com/labs/articles/cisotociso/i-was-a-human-captcha-solver>.

³See <https://www.yubico.com/us/store/> (accessed 23 May 2022).

laptops that provide valid signals incur very significant one-time costs. Moreover, they will also require paying for human subjects to provide valid proofs of personhood, which will further incur per-usage costs.

A rough analysis using the above figures could suggest that it might be economically advantageous to launch automated physical attacks on touch-based authenticators via commodity hardware. However, servers can tip the economic balance against attackers, by leveraging the asymmetry of information about the types of authenticators being used. While leveraging this asymmetry remains an open research topic, our system provides visibility into the global breakdown of device types, which attackers do not have. As mentioned previously, authenticators are typically associated with coarse-level batches of a specific model by their attestation certificates (Section 3.2). Thus, a server has the ability to collect and maintain a view of different device types. An attacker may invest in a particular model of security key that could be removed from the list of allowed devices (e.g., if it was uncommon and mainly used for attacks). This adds an additional risk for the attacker, who may find their investment wiped out with one configuration change. The attack cost is higher to maintain for a diverse profile of security keys that matches up with the global distribution. The server could remove support for specific device keys if a farm of them was discovered; it is worth noting that this would affect legitimate users that share devices within the same batch as the attacker, but the diversity of keys used in practice means CAP would still be effective for most of the other users. Note also that unlike human-assisted farms, where the cost is per-CAPTCHA, security key farms have an upfront cost that is amortized over time. The ability for the server to selectively support the feature for specific devices or regions introduces a significant downside risk for any capital investment by attackers. In summary, valuable future work would establish whether using such mechanisms as a viable mitigation is possible, without introducing significant overheads to legitimate users.

8 Related Work

CAPTCHA-related research that has motivated and guided our explorations is described throughout this work and includes studies of usability [15, 18, 24] and security [34, 35]. Recent and related streams of study on security key usability identify many strengths along with some weaknesses [16, 8]. Their results are highly encouraging and report that users are readily able to physically interact with YubiKeys. Minor and occasional problems included key touches that fail to be recognized [16], or the key being inserted incorrectly [8]. Users are also concerned about being able to locate or losing such small devices [30]. These occurrences will be familiar to any user of touch and biometric devices (e.g., mobile device fingerprint sensors). We anticipate their reductions with

practiced use, improvements in hardware sensors, and further hardware integration.

Many security key usability challenges emerge as part of a two-factor authentication (2FA) [11, 16]. Our hardware challenge task has lower barriers to entry since (i) there are no passwords or user accounts involved, and (ii) a failed challenge can fall back to a CAPTCHA. However, the same works identified inconsistencies and inadequacies in messaging and best practice for WebAuthn among Internet browsers [16]. This observation is in keeping with our own and deserves further attention.

There have also been some recent studies about FIDO and WebAuthn usability more broadly, which are helping highlight specific challenges and potential solutions. A study of mobile phones as roaming authenticators [29] suggested that users wanted to take advantage of the user presence features (such as facial recognition) available on their smartphones for authentication; the convenience of these features could also be leveraged for the attestation-only variant of WebAuthn (as in CAP).

An exploration of user misconceptions about WebAuthn biometrics [25] provides useful insights into some of the persistent points of confusion and gives recommendations for mitigating these (e.g., by providing more explicit guidance about where biometric data is stored, and providing users with more than just simple notification messages when explaining the technology). We have identified similar issues and are experimenting with ways of improving the user experience, particularly in terms of communication.

9 Concluding Remarks

The balancing act between security and usability places undue hardship on users to complete frustrating, impenetrable CAPTCHAs that have a number of serious shortcomings. Based on our user study we believe that a cryptographic attestation to a physical interaction provides a better solution for users without degrading bot detection.

We hope that others will be able to apply this solution in their own environments, leveraging the open WebAuthn standard to benefit from cryptographic attestations for human challenges. Our evaluation provides us with confidence that this is a fruitful approach for those users poised to take advantage of it; the necessary hardware is already widely available and is being rolled out even further. We have identified a number of barriers to adoption, however, primarily in the areas of privacy, clarity of communication, and consistency of user experience with WebAuthn. We will continue to pursue research into these areas, in hopes that cryptographic attestations will be more widely adopted and provide users with better ways of completing human challenges.

Acknowledgments

We gratefully acknowledge the assistance of our study participants and our anonymous shepherd and reviewers. We would also like to thank our Cloudflare colleagues for their valuable support.

References

- [1] Usage statistics and market share of Cloudflare. <https://w3techs.com/technologies/details/cn-cloudflare>.
- [2] William Aiken and Hyounghshick Kim. Poster: Deepcrack: Using deep learning to automatically crack audio CAPTCHAs. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 797–799, 2018.
- [3] FIDO Alliance. FIDO Alliance Metadata Service v3.0. <https://fidoalliance.org/metadata/>. Accessed Feb 2022.
- [4] Dirk Balfanz, Alexei Czeskis, Emil Lundberg, J.C. Jones, Jeff Hodges, Michael Jones, Rolf Lindemann, Akshay Kumar, and Huakai Liao. FIDO UAF Protocol Specification v1.0. FIDO Alliance Standard, FIDO, December 2014. <https://fidoalliance.org/specs/fido-uaf-v1.0-ps-20141208/fido-uaf-protocol-v1.0-ps-20141208.html>.
- [5] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.
- [6] Garrett Bekker and Matthew Utter. Work-from-home policies driving MFA adoption, but still work to be done, Apr 2021. "<https://pages.yubico.com/work-from-home-policies-driving-mfa-adoption>". Accessed Feb 2022.
- [7] John Brooke. SUS - A quick and dirty usability scale. *Usability evaluation in industry*, page 189, 1996.
- [8] Stéphane Ciolino, Simon Parkin, and Paul Dunphy. Of two minds about two-factor: Understanding everyday FIDO U2F usability through device comparison and experience sampling. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, 2019.
- [9] J Clement. Mobile internet usage worldwide - statistics & facts. Statista, Jul 12 2021. <https://www.statista.com/topics/779/mobile-internet/>. Accessed Feb 2022.
- [10] George S. Coker, Joshua D. Guttman, Peter A. Loscocco, Amy Herzog, Jonathan Millen, Brian O’Hanlon, John Ramsdell, Ariel Segall, Justin Sheehy, and Brian Sniffen. Principles of remote attestation. *International Journal for Information Security*, 10(2):63–81, 2011.
- [11] Sanchari Das, Andrew Dingman, and L Jean Camp. Why Johnny doesn’t use two factor: a two-phase usability study of the FIDO U2F security key. In *International Conference on Financial Cryptography and Data Security*, pages 160–179. Springer, 2018.
- [12] MDN Web docs. Web authentication API. https://developer.mozilla.org/en-US/docs/Web/API/Web_Authentication_API#Browser_compatibility. Accessed Feb 2022.
- [13] Josh Dzeiza. Why CAPTCHAs Have Gotten So Difficult. *The Verge*, Feb 2019. <https://www.theverge.com/2019/2/1/18205610/google-captcha-ai-robot-human-difficult-artificial-intelligence>.
- [14] Wesley Evans and Tara Whalen. More devices, fewer CAPTCHAs, happier users, August 2022. <https://blog.cloudflare.com/cap-expands-support>.
- [15] Valerie Fanelle, Sepideh Karimi, Aditi Shah, Bharath Subramanian, and Sauvik Das. Blind and human: Exploring more usable audio CAPTCHA designs. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 111–125, 2020.
- [16] Florian M Farke, Lennart Lorenz, Theodor Schnitzler, Philipp Markert, and Markus Dürmuth. “You still use the password after all” – Exploring FIDO2 Security Keys in a Small Company. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 19–35, 2020.
- [17] Armando Faz-Hernández, Watson Ladd, and Deepak Maram. ZKAttest: Ring and group signatures for existing ECDSA keys. In *International Conference on Selected Areas in Cryptography*, pages 68–83. Springer, 2022.
- [18] Ruti Gafni and Idan Nagar. CAPTCHA – Security affecting user experience. *Issues in Informing Science and Information Technology*, 13:063–077, 2016.
- [19] Google. Google reCAPTCHA: Register a site. <https://www.google.com/recaptcha/admin/create>. Accessed Feb 2022.
- [20] Lucy Handley. Nearly three quarters of the world will use just their smartphones to access the internet by 2025, Jan 2019. "<https://www.cnn.com/2019/01/24/smartphones-72percent-of-people-will-use-only-mobile-for-internet-by-2025.html>". Accessed Feb 2022.

- [21] hCaptcha. hCaptcha Developer Guide. Available at <https://docs.hcaptcha.com/>.
- [22] Scott Hollier, Janina Sajka, Matthew May, Michael Cooper, and Jason White. Inaccessibility of CAPTCHA. W3C note, W3C, December 2019. <https://www.w3.org/TR/2019/NOTE-turingtest-20191209/>.
- [23] J.C. Jones, Akshay Kumar, Alexei Czeskis, Vijay Bharadwaj, Dirk Balfanz, Hubert Le Van Gong, Huakai Liao, Michael Jones, Jeff Hodges, Rolf Lindemann, and Arnar Birgisson. Web Authentication: An API for accessing Public Key Credentials Level 2. W3C working draft, W3C, July 2020. <https://www.w3.org/TR/webauthn-2/>.
- [24] Kat Krol, Simon Parkin, and M Angela Sasse. Better the devil you know: A user study of two CAPTCHAs and a possible replacement technology. In *NDSS Workshop on Usable Security (USEC)*, volume 10, 2016.
- [25] Leona Lassak, Annika Hildebrandt, Maximilian Golla, and Blase Ur. “It’s stored, hopefully, on an encrypted server”: Mitigating users’ misconceptions about FIDO2 biometric WebAuthn. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 91–108, 2021.
- [26] Wei Liu. Introducing reCAPTCHA v3: the new way to stop bots. Google Webmaster Central Blog, October 2018. <https://webmasters.googleblog.com/2018/10/introducing-recaptcha-v3-new-way-to.html>. Accessed Feb 2022.
- [27] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the kingslayer of user authentication? A comparative usability study of FIDO2 passwordless authentication. In *IEEE Symposium on Security and Privacy*, pages 268–285, 2020.
- [28] Thibault Meunier. Humanity wastes about 500 years per day on CAPTCHAs. It’s time to end this madness, May 2022. <https://blog.cloudflare.com/introducing-cryptographic-attestation-of-personhood>.
- [29] Kentrell Owens, Olabode Anise, Amanda Krauss, and Blase Ur. User perceptions of the usability and security of smartphones as FIDO2 roaming authenticators. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 57–76, 2021.
- [30] Joshua Reynolds, Nikita Samarin, Joseph Barnes, Taylor Judd, Joshua Mason, Michael Bailey, and Serge Egelman. Empirical measurement of systemic 2FA usability. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 127–143, 2020.
- [31] Michael Richardson, Carl Wallace, and Wei Pan. Use cases for Remote Attestation common encodings. <https://datatracker.ietf.org/doc/html/draft-richardson-rats-usecases-08>, November 2020. Work in Progress.
- [32] Max Roser, Hannah Ritchie, and Esteban Ortiz-Ospina. Internet. *Our World in Data*, 2015. <https://ourworldindata.org/internet>.
- [33] Catherine Schwab. Google’s new reCAPTCHA has a dark side. Fast Company, June 2019. <https://www.fastcompany.com/90369697/googles-new-recaptcha-has-a-dark-side>.
- [34] Chenghui Shi, Shouling Ji, Qianjun Liu, Changchang Liu, Yuefeng Chen, Yuan He, Z Liu, R Beyah, and T Wang. Text Captcha is dead? A large scale deployment and empirical study. In *The 27th ACM Conference on Computer and Communications Security*, 2020.
- [35] Suphannee Sivakorn, Jason Polakis, and Angelos D. Keromytis. I’m not a human: Breaking the Google reCAPTCHA. In *Black Hat ASIA*, 2016.
- [36] Saumya Solanki, Gautam Krishnan, Varshini Sampath, and Jason Polakis. In (cyber)space bots can hear you speak: Breaking audio CAPTCHAs using OTS speech recognition. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec ’17*, page 69–80, New York, NY, USA, 2017. Association for Computing Machinery.
- [37] Jennifer Tam, Sean Hyde, Jiri Simsa, and Luis Von Ahn. Breaking audio CAPTCHAs. In *Proceedings of the 21st International Conference on Neural Information Processing Systems, NIPS’08*, page 1625–1632, Red Hook, NY, USA, 2008. Curran Associates Inc.
- [38] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: Using Hard AI Problems for Security. In Eli Biham, editor, *Advances in Cryptology — EUROCRYPT 2003*, pages 294–311, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- [39] Guixin Ye, Zhanyong Tang, Dingyi Fang, Zhanxing Zhu, Yansong Feng, Pengfei Xu, Xiaojiang Chen, and Zheng Wang. Yet Another Text Captcha Solver: A Generative Adversarial Network Based Approach. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS ’18*, page 332–348, New York, NY, USA, 2018. Association for Computing Machinery.

Being Hacked: Understanding Victims' Experiences of IoT Hacking

Asreen Rostami^{1,2}, Minna Vigren², Shahid Raza¹, Barry Brown^{2,3}

¹*RISE Research Institutes of Sweden*, ²*Stockholm University*

³*Department of Computer Science, University of Copenhagen*

asreen.rostami@ri.se, minna.vigren@helsinki.fi, shahid.raza@ri.se, barry@di.ku.dk

Abstract

From light bulbs to smart locks, IoT is increasingly embedded into our homes and lives. This opens up new vulnerabilities as IoT devices can be hacked and manipulated to cause harm or discomfort. In this paper we document users' experiences of having their IoT systems hacked through 210 self-reports from Reddit, device support forums, and Amazon review pages. These reports and the discussion around them show how *uncertainty* is at the heart of 'being hacked'. Hacks are sometimes difficult to detect, and users can mistake unusual IoT behaviour as evidence of a hack, yet this can still cause considerable emotional hurt and harm. In discussion, we shift from seeing hacks as technical system failings to be repaired, to seeing them as sites for care and user support. Such a shift in perspective opens a new front in designing for hacking - not just prevention but alleviating harm.

1 Introduction

The threat of being hacked is sadly a common part of technology use. This is particularly challenging for the Internet of Things (IoT), since hacked IoT can be used to cause serious physical harm or discomfort, such as locking a user out of their home, or video recording their children. While much research has focused on the technical aspects of IoT security (e.g. [1, 42, 45, 91]), there has been recent interest in how users manage their IoT security (e.g. [5, 8, 62, 87]). Building on these works, in this paper we focus on the 'user experience' of being hacked: how users discover they are hacked, cope with the hack, and manage the damage done.

Our data comes from the discussions and reports from users who believe they have been hacked, posted on online discussion forums, product support forums, and product review pages. We focus on users' experiences of hacked IoT systems, since these systems both present particular issues in terms of user interface, but in the damage that a hacked device can inflict. From these online sources we collected 210 cases of users reporting having an IoT device hacked. These first-hand experiences and stories, and the online discussion around them, gives us new insights into these hard to reach experiences. Prevalent throughout this data was users' uncertainty and doubt around 'being hacked'. This is captured well by one poster who asked if his experiences were [a] "Ghost, cat or hack?" about their experiences. Users firstly asked *if* they have been hacked, then *who* has hacked them, and lastly *why* they have been hacked. There are also situations where users suspect they have been hacked, but are unsure if it is actually an intrusion, a technical problem or unusual system behaviour. For example, one common brand of smart home light bulb will flash on and off to indicate an error condition – and when this sort of behaviour happens across multiple devices, users reported confusion, concern and worry. These '*non-hack hacks*' are a major problem since users need to deal with them as if they are hacks, and so they can cause similar amounts of disruption, worry and personal pain.

In discussion we explore how this data lets us look at the social and psychological aspects of hacks. With a focus not on the hacks, but the people whom the hacks impact, how can we design for managing the impact that hacks have on users. We discuss '*cybernoia*' - where users become unsettled through a hack or suspected hack, but also how hacks can disrupt relationships and expectations around technology. This refocuses design attention from just preventing hacks to supporting victims, and designing systems that could take a support role similar to the roles the forums provide for users. In conclusion, we discuss how hacks interfere with relationships, in particular when the person who is behind the the hack is known by the victim.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022,
August 7–9, 2022, Boston, MA, United States.

2 Related works

With our focus on hacking, users and IoT, we have drawn on two main research areas: first, research on user experiences of being hacked and second, security and privacy issues around domestic IoT. In addition, with our use of discussion forums as a source of data we provide a brief review of the use of online discussion data in HCI.

2.1 User experiences of being hacked

For as long as there have been computational systems, hackers have attempted to penetrate and hack these systems for nefarious reasons, and users of these systems have suffered from these attempts. Hackers' intentions and their experience of hacking have been a topic of interest in a number of different fields [19, 38, 44, 80, 90]. Research has also studied end-users' practice of hacking their own devices to make the device serve different purpose and interests, beyond the intention of the device manufacturer [11].

However, it has only been recently that the user's experience of being hacked - the victims' perspective - has come into focus. Tian et al. [79] document different types of cybersecurity incidents reported by users and describe victims coping mechanisms in dealing with these hacks. They in particular discuss how users felt ambiguous about some of these incidents, resulting in users denying the existence of the hack as a coping strategy. More broadly, this work has explored how victims of cybercrime [15, 21, 37] can be caught up in large-scale attacks on organisations [22]. This research also covers how a 'victim-blaming' discourse around online fraud as well as online humour and shame are used to push victims to stay silent. With the rise of online harassment [51, 82], phishing [20, 79, 85, 86] and ransomware attacks [14, 71, 73, 89, 92], many users and organisations have been targeted in hacks where they are faced to pay ransom money or experience devastating damage to their digital ownership of their properties and data. In one example, Zhao et al. [93] studied a group of attending surgeons' experience of a ransomware attack that caused the shutdown of their hospital. Their study demonstrates how such incidents caused not only disruptions in their online communication but also affected their process of carrying out of surgery and medical procedures.

One area of growing concern is technology-mediated abuse, where hacking or technology is misused as part of violence between current (or former) partners. Parkin et al. [62] document a number of different types of threats where, for example, a security camera can be used (hacked or otherwise) to facilitate intimate partner abuse. Bellini et al. [7] take a different research angle, and present how potential perpetrators use online resources to plan, discuss and legitimise their use of technology in intimate partner abuse. In a related study, Freed et al. [30] discuss how in intimate partner abuse, the attacks performed by the abuser are not necessarily technologically

sophisticated, rather the abuser uses their knowledge about the victim to crack passwords and hack into their accounts and devices. One concerning issue brought up in this study is how victims cannot always remove a compromised device from their network, since that device is used by the victim to gain access to professional and social support to deal with the hack. Removing the technology might solve the problem, but could put the victim in danger of isolation [31].

For this study, we were specifically interested in IoT hacks, in part because these systems present more constrained security user interfaces to users, but also for the ways in which IoT hacks risk a potentially higher level of material harm for users. As Slupska points out, consumer IoT products while marketed to support protection and care actually create new vulnerabilities - particularly with respect to domestic violence, something almost entirely ignored in the smart home security analysis literature [74]. McKay and Miller [58] discuss how home IoT devices can be used to perpetuate new forms of harassment and coercive control in the home environment, moving beyond hacking into 'traditional' technologies such as mobile phones or personal computers. Levy and Schneier [52] broaden this discussion by highlighting how within intimate relationships privacy and security threats can arise from the lack of appropriate design, and how cybersecurity broadly ignores the different data sensitivities that can occur within relationships. This has resulted in systems that do not protect against intimate threats [53]. The relevant and timely concerns expressed in these studies open up the study of user experiences of IoT device hacking.

2.2 Security and IoT

HCI (and related work in UbiComp) has developed an extensive body of work around the user experience of IoT and home IoT in particular [8, 16, 26, 32, 39]. HCI research have increasingly brought more privacy relevant research to the fore [31, 48, 64, 65], and called for revisiting users' privacy design with respect to vulnerable populations (e.g. [57]). In one study, Choe et al. [18] studied privacy in relation to parents' use of smart security cameras at home as part of a 'responsible parenting' strategy. In a related study, Worthy et al. [88] found that users of IoT devices are often concerned about the trustworthiness of the device, particularly in terms of the data it collects as well as the person or organisation who controls the functionality of the device. In another study, Bouwmeester et al. [8] present different steps that a home IoT user may take to identify a device infected by malware. Their study highlights how participants felt uncertain about whether they have correctly identified the infected device, and how some participants failed to fully execute the recommended actions to remediate the infected device due to design complications or simply because of lack of security knowledge and experience.

In a series of original papers, Pierce [64] uses the 'creepiness' of internet-connected security cameras in domestic con-

texts to discuss hole-and-corner applications that make use of user data out of the context of the service they should provide, in a hidden way that could harm the user. For instance, neighbours could misuse security cameras to monitor others' religious commitment to digitally gaslight and blackmail them. Other scholars have also studied how users of IoT devices perceive the security and privacy of their data being used by these devices [43, 78, 94]. For instance, Jacobsson et al. [43] highlight the importance of understanding user interaction with IoT devices "in order to create usable privacy mechanisms". In a similar vein, Zheng et al. [94] studied how users perceive privacy when using IoT and what are the different approaches they take to protect the privacy of themselves and their homes. Their study highlights that the majority of users put their trust in the hand of the well-established brands and manufacturers to protect their data and privacy, selecting the device based on the vendor's reputation as well as online reviews about the device. One interesting aspect of this study is that users of these devices have reported government and Internet Service Providers (ISP) as the most worrisome outsider actors from which they wish to protect their data, rather than hackers. At times, this perspective served, perhaps, in light of the recent critical debate on state surveillance [72] and the commodification of personal data. However, during recent years and with the increased popularity of home IoT devices in the consumer market, home IoT users have witnessed and dealt with threats and security intrusions committed by different groups of individuals. Reports on various network-connected devices being exploited by criminals and hackers, such as baby monitors [3, 83] and smart doorbells [23, 63] are few examples of these security threats posed by different groups of bad actors with different goals and agenda in mind.

2.3 Forum Studies in HCI

The main data source we use in this paper comes from forum data posted on the internet. HCI researchers have broadly used online communities and forums as part of understanding both internet behaviour and community structure [55, 70] but also more generally to investigate users' motivations for their participation in online forums and communities [47]. Collecting data from users participating in different online communities and forums have been a rich source of data and inspiration for both HCI and CSCW communities [17, 27, 33, 46, 61, 67]. Particularly in the domain of health, research has looked at how individuals adopt or disengage with online communities [56], how these forums are used by patients for recovery [54], and how patients share their knowledge and experience of the medical condition to support each other [41, 67] as well as discussing the potential challenges for clinicians to participate in these forums [40]. Previous research [28] has also discussed the ethics of using online data for research, by presenting how users of online communities, such as Twitter, can have different attitudes toward privacy, and their expecta-

tions can be highly context-dependent. In their studies, Fiesler et al. [28] & Bruckman et al. [13] suggest that researchers should take extra measures to minimise potential harms that could arise from neglecting the privacy of online users by, for example, careful anonymization of names while making sure that the published data is not linked back to the online original account.

Forum studies are effective at overcoming methodological challenges such as collecting data from communities who are isolated due to social and geographical constraints. Moreover, collecting interview data when participants need to disclose sensitive information about themselves or their experience (such as understanding survivors of sexual abuse [4]) has proven to be a challenging task, as many survivors wish to remain anonymous or find it hard to talk about their experience outside the social media context due to their traumatic experience. Previous research has also discussed how the rarity of cases as well as the stigma around speaking about the topic have affected the research and data collection strategy. Such a case can be seen in identifying early adopters and early victims of a particular technology, for example in studying victims' experience of technology-mediated abuse [58]. Research have also reflected on the difficulties of collecting interview data from participants around a phenomena due to an ongoing world crisis such as outbreaks or pandemic. For instance, Gui et al. [34] use online communities to collect data from geographically distributed travellers affected by the Zika outbreak, something that could not be collected directly due to resulting travel restrictions.

3 Methods

For this study, we were interested in understanding better the experience of hacking victims, how they make sense of the hack, and in particular the different practical and emotional resources they deploy to deal with being hacked. In planning our data collection, we were concerned that the social stigma attached to being a victim [4, 22] could affect the quality of collected data, and that interview data might be unreliable as in cases where there is a strong social desirability bias [10]. Moreover, during our initial attempts to recruit interview participants we found considerable hesitancy to talk to researchers about these sensitive issues. This led us to explore other means to collect data. Looking online for users' reports of being hacked, we were surprised to find many reports in general online discussion forums (Reddit), product support forums, and user-generated product reviews (Amazon). There were also extensive responses and discussion of the initial posts – spanning from related stories, advice around dealing with and repelling attacks, as well as broader discussion around the hack, the hacked devices, and IoT security. This led us to explore how online posts could be used as a data source.

As we discussed above, forum data has been used in user

research, often as part of investigating sensitive topics. These explorations usually take a dual path – first, the forums themselves are a topic of study, the types of contributions and how the forums benefit those who contribute or read posts. Second, the actual content on the forums provides an insight into users’ experiences outwith the online forums. Internet posts have a number of advantages over reports collected through interviews. Interviews as a method shape and prompt data in strong ways by the questions being asked, and the responses of the researchers [6]. Forum posts offer the advantage of being naturally produced data, as well as containing responses of other posters. This said, the lack of control over the data does mean that it is harder to check and validate data. Participating in an online community is essential so as to learn possible in jokes, ‘house style’ and the like that might mislead a researcher or ‘outside’ readers of the forum content. It is also important to maintain some scepticism around posts and responses and recognise that trolling is a common aspect of all online interaction.

We collected our data during January - March 2021. Before conducting our study we informally read and participated in Reddit forums around IoT device support, smart homes, cybersecurity, and hacking. Building on this we searched media coverage and security reports on vulnerabilities on IoT devices. Through this pre-screening of vulnerable devices, we were able to identify a list of 13 hacked devices from different brands that served as our seed list of devices to look for stories from users from online forums. These devices included security cameras, smart locks, doorbells, lights, TVs, speakers, and voice assistants. We used the forums’ internal search function and applied the brand names, general devices names, as well as keywords such as *smart home*, *home IoT* in combination with *hack* and *hacked* (Appendix A: Table 1). This actual data collection expanded the list of devices as the searches from these forums brought up new devices to be included in our list of hacked devices (Appendix B: Table 2). From this, we collected together a corpus of 210 hack reports and posts along with the related discussion. From Reddit we took data from subreddits with 114 posts (54% of all the data), product support forums with 84 posts (40%), and 12 Amazon reviews and product support forums discussing being hacked (6%). Overall this was more than 1000 pages of data.

It is worth mentioning that while our focus was on collecting reports of hacked IoT devices, we found that often reports of security vulnerabilities for these devices would come not from the devices themselves but from the WiFi network they were on, or the management account with the company that made the device. This meant that our data includes reports that were somewhat associated with an IoT device (as reported by the original poster), such as routers and the network where the device was installed. One motivation behind this data collection decision was – as we shall see in the analysis section – while the devices themselves could be hacked in various ways, a home IoT device exists in a complex network

of other vulnerable access points. Indeed, access to the home WiFi seemed a common vulnerability here, since home WiFi passwords are often rather weak and shared relatively broadly within the household members and even with guests. Being on the same network, therefore, could be used as a way of accessing IoT devices, sometimes without any further security needed. WiFi speakers, and TV video streamers, are a few examples that could then be vulnerable to unwanted access.

We analysed the data set using an open coding method. Our coding process was done using NVivo. Independently, two of the authors coded 50 threads each and developed their own codebook [49]. The research team then met, discussed this data and codebook, and in discussion formed a joint codebook. A small number of threads (5) were coded together in this joint session to test the joint codebook. Each author then coded the rest of their own data sets independently. During this part of coding, new codes did arise along with refinements on the old codes, which were addressed in two further joint sessions during the coding process. The final codebook (Appendix B: Table 3) included 32 main codes that we used to organise our data. These codes were aimed at categorising our data based on, for example, the type of the device, the description of the incident or hack, the (presumed) motivation behind the incident, the identified actor(s) behind the incident, the evidence available or shared about the incident, discussions on the analysis of the (presumed) hack, harms and damages caused by the incident, actions taken by the OP, and their existing security practices.

This categorisation helped us to analyse the whole corpus in a systematic manner, further identify patterns and develop themes through an in-depth iterative and collaborative analysis process. The categorised and coded data were then iteratively analysed using thematic analysis [9]. In doing so, we mapped our data to the themes that we had identified contributing to developing an understanding of how and why users share their stories of being hacked in online forums, what the different types of online interaction they engaged in during this process are, what type of hacks they have encountered and dealt with, what kind of collective sense-making is entailed around different shared cases, and what (if any) the consequences of the incident were.

3.1 Ethics

We have restricted our data collection to those forums that are available to the public without the need for registration and considered these forums as public material [77]. However, we have not collected any personal data about the users of those communities that we studied beyond their reports and posts and content provided. In presenting quotes in the results sections we have provided most quotes verbatim, resonating with Brown et al. [12] argument that anonymity should be provided when “participants *want* to be anonymous.” However, to ensure the safety and anonymity of those who have experienced

serious threats, report serious mental health issues, or accuse an individual of wrongdoing, we have paraphrased our quotes in a way that it is not easily traceable to the Original Poster (OP) and the community they were engaged with [25, 66]. In doing so, we have kept the core concerns that are mentioned in posts and the type of device they have used but removed the brand and name of the specific forum they have originally posted.

4 Results

Our analysis is broken into three sections. We start by discussing how users reported being hacked on the different type of online forums we studied, and in particular how users report their hack experiences. Second, we discuss how users dealt with uncertainties around being hacked, and the doubts they have about if, how, and who had hacked them. Lastly, we discuss how users deal with hacks, and the lengths to which they go to understand the hack and to deal with the hurt and harm it causes.

4.1 How users report being hacked

The posts we analysed spanned from users asking the simple question “Was I hacked?” to much more detailed descriptions of events, technology in use, actions taken to solve the situation, evidence to back up the incident, and social dynamics that have caused the situation (as in the case of (ex-)partner abuse or neighbour harassment). These hack descriptions give us access to users’ accounts of what it is like to be hacked, but also how users detect hacks in the first place, how they identify the nature of the hack and its potential source, as well as how they attempt to ‘fix’ the hack, and deal with the problems that the hack causes.

A prevalent feature across all our data is the way in which users tell their ‘hack story’. End-users recount ‘what happened’ by telling the general story of how they discovered the hack, the evidence of what makes them think it is a hack, and how they proceeded to deal with the impact of the hack. As an ‘exceptional event’, users often go to great lengths to explain why they suspect a hack, rather than more benign explanations. As is typical when individuals describe unusual events, the justification of an unusual event is preceded by evidence that shows how more mundane explanations could not apply [68]. As we quoted in the introduction, a user asked the question ‘cat, ghost or hack?’ in the title of their post – if it is not an accident (the cat), or the paranormal (the ghost), this it must be - as a last resort – a hack.

Post were sometimes met with scepticism as to the reliability of the story, or the user’s motives. However, more commonly those who responded took the stories seriously, offering a range of helpful responses, spanning across technical, social and practical support of different sorts. As with storytelling more broadly [69], these ‘hacked stories’ often prompt the

telling of follow up ‘hack stories’ by other posters. These first stories thus prompt ‘second stories’ of similar hacks that had happened to other users, either with the similar technology or similar in form. In the Reddit and product support forums, these second stories would take the form of follow-up posts sharing a poster’s own story. Elements of the hack then would be taken apart - discussed for their relevance in different ways. Interestingly, even with the online reviews - although the format does not lend itself to responses - reviewers would start their own reviews by referencing earlier reviewers, such as “*I also had my security camera hacked*”. Reviewers then would connect together reviews to provide cross-validation of experiences and give added credence to the warnings of vulnerable technology. Besides asking for help, these second stories are often rich in content including also warnings about the product, discussion on cybersecurity in general, complaining about bad product support and design, or admonishing the manufacturer for a faulty product.

Looking over the corpus, we can characterise three elements that reoccur in the hack posts: *warnings*: descriptions of problems with, and a warning against buying a particular device, *calls for support and help*: where a post asks directly for support with dealing with the hack, and *initiating discussion*: more conversational points where the aim seems to be more to start discussion around the hacking incident.

4.1.1 Warnings

Some parts of posts are written essentially as warnings, with a user reporting their hack as evidence of the unsuitability of an IoT device. For example:

“It was the best camera till it was compromised. I got it so I could make sure my place was secure. BUT sure enough last night [...] at about 9pm [...] I WATCHED IT MOVE 5 DEGREES TO THE LEFT [A]ND THEN BACK TO THE RIGHT! [...] I strongly urge you not to buy this camera because if one is compromised then most likely their system is too”.

Warnings were most common in the product reviews, although they were not unique to the reviews. In some posts, the reported incident was only briefly described, with the description rather focused on pointing towards describing vulnerabilities in the device, and the user offering a brief recount of the experience of the hack or problem. With warning posts, the original poster is ‘making a case’ about the device, justifying their warning by referencing their own experiences, others’ experiences, and even in some cases media reports or other online discussions.

4.1.2 Calls for support and help

A second element was the more direct asking for support, either from other users or directly from the manufacturer. These requests spanned questions on how to protect oneself against the (presumed) hack, but in more desperate cases, they

were pleas for help to know what to do next from exasperated users who had no idea what to do and what was going on. For instance, in the following example on the Chromecast support forum, a user reported about their hacked device, asking for help not only from the community members but the Google product team, while also reporting on their immediate action to call the law enforcement:

“My Chromecast was hacked and turns my tv on with creepy short videoclips.. what is going on? [...]. I called the police and the Az Attorney Generals office”.

While manufacturers do respond, they usually answer through a template of asking for particular information. So, for example, with remote cameras there is usually an attempt to obtain a copy of access logs of some sort through which the manufacturer can detect who had accessed the data and whether the incident is in fact due to unauthorised access or something else. Follow-up posts by others often offer stories of earlier reports, and earlier solutions provided by the support team, demonstrating similarities in hacks across the life-cycle of the product. Interestingly, these requests for help are not always technical in nature. Especially in the case of ‘known hacker hacks’, the requests for support and help can be more emotional and practical in nature, such as who to report a hack to, or how to deal with the relationship with the hacker.

4.1.3 Initiating discussion

A third element of the post was when posters gave more detailed descriptions of the hack, where the intention of the OP seems to be more to initiate discussion and hear if others have experienced the same. In these posts, the shared story can be seen as reporting the hack as part of asking for advice and opinions from others but more importantly as an opening for discussion. As in the following quote, the OP describes what has happened and the actions they are planning to take like contacting the manufacturer, with the ending question on wanting to know if anything similar has happened to others.

“10 minutes ago my Arlo camera in the kitchen started making pornography sounds through it speaker, it lasted around 30 seconds. Just before the event my google home speaker made the blimp noise as if someone had said the trigger words. My Arlo is linked to my Smarter things hub, google home and I think the ifttt [if this then that] service. I guessing something has been hacked! This is not a joke and is freaking me out. During the sounds the lights was on as if someone was watching. As you can imagine I have turn everything off and will be contacting Arlo support to shed light on the event. Did anyone else experience thus, surly I wasn't the only one”.

Posts like these would usually be responded to by follow up posts from other users which replied to the concern over IoT security and discussed different aspects of it but also provided practical security information – such as protecting and changing passwords, or setting up two-factor authentication (2FA). The example above received 20 replies from 15 Red-

dit users. These responses include broad security advice that describes how to ‘delete and wipe’ a network that has been compromised and ways of re-securing home technology in such a way that it would not be vulnerable to being hacked all over again. This discussion takes a number of different directions but responds to the different issues raised by the original post - the different services outlined, the hack and whether it did take place and so on.

4.2 Making sense of the hack

We move on to consider perhaps the core problem that our posts reported in dealing with hacks - the problem of knowing what is going on and how to deal with uncertainties around hacks.

In the posts, we analysed it is the users themselves who define their own experiences as ‘being hacked’ and choose to post or respond to others’ posts. This means that we are including many cases that would perhaps not fit with security professionals’ definition of being hacked, or even a common sense understanding of a hack. Uncertainty about the status of a hack is ever-present in our data, and this leads to the issue of how ‘uncertain’ hacks should be approached since they often still can cause real hurt and damage. In some cases, it becomes clear through the online discussion that the OP themselves are actually mistaken about the hack, and that it is some sort of other technical problem in the functioning of the IoT device. We characterised these cases as ‘non-hack hacks’, in that even though the OP is mistaken, and the hack did not actually occur, the ‘hack’ has a real harmful impact, described by the OP in terms of the time they waste to diagnose what has happened, the worry they have around the hack and the emotion cost of originally believing they were hacked.

In some threads, the OP later reveals that they were joking, a trick was played on them or that the post is not to be taken at face value. Clearly, paying attention to the discussion alongside the original post is important here. By analysing forum data, we have the advantage of having responses from others and follow up posts from the OP which can shed light on the original story, or at least put it into some context.

Indeed, uncertainty was a prevalent feature of users’ hack experiences - uncertainty about whether they had been hacked, uncertainty about who might have hacked them, and uncertainty about what to do about the hack. Users start by being suspicious about something that has happened with their IoT system - this might take the form of unexpected system behaviour – such as an IoT light flashing on and off without user intervention or a security camera moving or making clicking sounds. Depending on their technical skills or experience, the user usually collects ‘evidence’ about the hack, trying to find out if they have been hacked by looking at the system behaviour, checking security settings and logs in order to make sense of the incident or to investigate what is going on. In the cases we analysed, this often leads to some attempt to fix the

problem – to repel the hacker or to find some way of barring entry by the hacker (such as changing a password or turning a device off).

4.2.1 Have I been hacked?

As we described above, the hacked stories often contain a ‘cry for help’ of sorts, with OPs using the platforms with the hope of receiving an outsider’s perspective to help them make sense of the incident and better understand it. Indeed, to a non-expert user, nearly any unfamiliar or unexpected behaviour can be seen as a suspicious action or a potential hacking alert that they need to deal with. For instance, a post to Reddit describes an unexpected action occurring without a chance to verify if this is a hacking attempt:

“Is my Alexa Hacked? I have recently put an IP camera to monitor my cats activity while I’m sleeping and away [...]. A couple days ago I was reviewing the videos [...]. At 3:20am (halfway through the video) you can see the light on the Echo Dot go on. It stays on for about a minute. Nothing is said or heard except for silence. When I checked the history in the Alexa app there is absolutely nothing that triggered it to go on. There have been a few times when Alexa has started recording without being prompted but I can usually listen to what it was...this was just completely undocumented. I cant find any logs on how to check what prompted it to turn on. It’s been bothering me ever since”.

In response to this post, posters discussed the different scenarios that could be the reason behind the system’s behaviour. For example, one poster highlighted how this could be a less-known feature of the device, designed to communicate some system changes, particularly *“when software updates are applied”*. So while in this case, it is perhaps unlikely that the Alexa was hacked, the user has been “bothered”, to the extent of posting on the forum for help. This uncertainty does not always stem from a lack of technical knowledge - it can also come from the complex and ambiguous design of particular IoT systems. Without any screen to explain its behaviour, for unobtrusive alerts the Alexa can only turn its light on. The device does not give any additional explanatory feedback about ongoing background activities (such as system updates), or how the device’s visible features (such as lights) react or change in response to these activities.

Indeed, the most common question posted to the forums was in varying ways asking “have I been hacked?”. The forums have many cases where users encountered an incident, which they think might be a potential intrusion, but were unsure how this could have happened, and wanting to know if the device is ‘hackable’. In one example, a user recounts their experience of hearing a whisper coming from their security camera but being unsure if the device could be exploited by a hacker:

“I’m no tech-savvy over here, but I need to understand logically if there is a possibility that the Yi Home Camera

we are using is being hacked. Last night we decided to move the camera from the living room where my son plays to the bedroom and a couple of minutes later while putting him in bed I heard a very clear whisper of someone saying something. My husband thinks I’m hallucinating (I’m sick with a bad cold and been taking Advil from the fever) but I swear I heard a voice from the camera. Can someone tell me what to do and if we should worry, or if this is just simply nothing? I’ve disconnected and deleted the camera”.

This process of doubt and uncertainty can be quite harmful – with some users go as far as to questioning their mental health, or having their senses questioned as in the quote above. In one post to Reddit about a similar incident related to a different device (a Ring doorbell), the OP questions their mental health and suggests that another explanation to the scary event could be paranormal activity:

“So last night I had the most paranormal experience of my life. My wife had just gone to bed, I was the only person awake in my living room. I was playing a game on my iPad and had just clicked on a Netflix show. It got stuck loading at 25%. I wasn’t paying much attention to it, it was completely silent and I hear someone whisper “hey guuyysss... heeeeyyyyyy”. I immediately grab my gun and clear my house. All my windows are shut, no one is around. The voice came from inside my living room. I then remembered my ring was charging, but I have 2 factor authentication, so I would know if anyone tried to log in or attempt to change password. My wife was passed out, and it was 100% a male voice. The only solution I have to this is someone has access to my ring and was playing around. Is this a common thing? Because it scared the living shit out of me and I still have no idea where it came from. Either I’m going crazy, there’s a ghost in my house or someone got into my ring account”.

Mental health state, or relating the incident to something paranormal, is in fact a common thread in our data, and mentioned by OPs or their family members as well as suggested as an explanation by other posters as well. In further discussing the previously quoted incident, one user contributes to the discussion of possible scenarios of occurrence by suggesting the OP may suffer from a momentary ‘auditory hallucination’ implying that the hack could not have happened perhaps because of the lack of evidence of intrusion or technical possibilities of performing such a hack with 2FA being in use. This shows that the likelihood of hacks is still considered rather exceptional, even though in this case there are reports and verified cases of similar incidents in relation to this specific device: *“Are you on any medications? Auditory hallucinations are a thing that can randomly happen with or without being on medications. It doesn’t necessarily indicate a bigger issue. If similar things recur I’d look into getting medical advice though”.*

4.2.2 Who hacked me?

Once the user suspects a hack has taken place, attention then moves to other aspects of the potential hack. One key question is *who* might have hacked the user, and in particular if the hack is from someone known or unknown. Most frequently the hacker was identified as an unknown person – as ‘someone’:

“I think someone has hacked my smart TV. So I’m sitting down playing Minecraft and all [t]he sudden my screen goes black for a few seconds my TV then automatically switches into screen cast mode and porn begins to play I don’t think anyone’s in cast range”.

Although in many of the cases, the possible hacker remains unknown, it is interesting how common place it is for the OP to suspect a close person being behind the hack is. In these cases, the posters can go through in detail candidates for the hacker, for example, (ex)partner and neighbour, discussing if they have a vindictive goal and agenda. In the following post, the OP describes how they think that their security camera’s unexpected behaviour is the result of their ex-partner having possibly hacked into their network and the camera, trying to abuse and stalk them in their home. They go on to describe how their camera behaves out of control, for example, by keep changing between the day and night mode even though their room with the camera is “bright enough, and not dark all.”:

“Because I have a stalker, I purchased and installed a security camera. I am terrified [...], and needed some sort of security. It could be other individuals too, but I think it is my ex-partner and it is him who is messing with my life. [...] I think he is monitoring me on a camera and he has hacked our network... he tells our kids about what we have done at home. He moves thing around so I can realise he can come and do these things. I have installed the camera so it faces the door and records those who enter. It send me a notification if any movement happens. Then the other night I got a notification on my phone because camera shifted between the night and day mode over and over, and the blue frame in the app showed it is trying to detect something, but it couldn’t. Our room was bright enough and not dark at all. I want to check with you folks here, is it possible to hack this security camera?”.

While the OP here is trying to understand whether such a hack is possible from technical perspective, they also highlight a cause and effect relationship between their abusive ex-partner and the ongoing “odd happenings”. From their point of view the ongoing abuse and previously documented and reported harassment of their stalker ex-partner are good enough reasons, and a catalyst, to believe it is them behind the incident while also being open to the idea of the incident could be done by “other individuals” as well.

In a related way, neighbours or friends were also often suspected as being behind hacks, building on existing bad relationships. The following example is taken from a post in a community forum of a brand of security camera, where a user reports on how they think one of their neighbour is hacking

into their smart camera, perhaps because of their previous ill-mannered actions towards the OP or their suspicious actions around their property:

“I have more than two of this type of security camera and I have used them for some time now [...] Recently there were some problems and I assumed that the issues I was experiencing with my cameras were related to the updates. I now have a strong suspicion that my neighbour has hacked my accounts. I have already set 2FA and changed passwords couple of times [...] Some illegal activities have occurred that resulted in property damages and somehow that period of time is not recorded.”.

4.2.3 Why was I hacked?

Another central aspect to these reported incidents is the collective sense-making around the motivation behind the hack. As described earlier, for those cases where OP themselves could identify the hacker, or be suspicious of someone, the motivation seemed to be more personal and easier to speculate. For instance, one person reported they think their device or network is being hacked because their neighbour “dislikes” them or the hack is a form of revenge and the result of dispute over statutory nuisance:

“I am in a noise battle, playing loud music and all of that, with my neighbour downstairs. They started all of this and now they have hacked my computer or wifi. I know I am hacked because they have sent me an email to me from my my account”.

Previous HCI research [30] have already reported on several cases from which smart technologies at home being exploited by intimate partners to not only invade the privacy of the victim but to abuse them physically and psychologically. As in the example from the ex-partner’s assumed hacking described above, the intention of the hacker was identified as to scare them, to monitor and control their life or as they put “to mess with their life”.

Another set of cases in which the motivation of the hack was discussed among community members are those that OP could not identify any personal relationship with the hacker. What seems to be a common acceptance among this group of community members, including victims of these hacks, is the fact that many IoT users may suffer from data breach and hacks, simply because users tend to re-use old passwords for their accounts which can result in their credentials being exposed and exploited by ‘bored teenagers’ as one user put it: *“It sounds like bored teenagers who found some credential dumps, and started trying them against Ring until they found a victim”.* For this group of users, the incident is not seen as a threat targeted towards them, rather they are most likely victims of opportunistic exploitation of technology vulnerabilities. As one Reddit poster describes “almost certainly [they] were not specifically targeted. The slim minority of people who get targeted are either political targets, or known financial

targets, or being stalked by people they already know.” Such a perspective, can be seen in relation to several reported hacks and incidents associated with famous mass incidents such as hacks of Google devices (e.g. smart speaker, Chromecast) or Ring security camera and door bell:

“Every 20 minutes or so my TV switches to some crappy YouTube video about PewDiePie with shitty rap music and a #ChromecastHack” hashtag. Anyone know how to stop this, it’s driving me bonkers”.

The “PewDiePie hack” originated from a YouTube subscriber battle between different internet channels, with hacks of printers and video players (such as the Chromecast). In this example, reported in Reddit, the OP and other community members affected by the similar attack discuss different motivations for the hack. What is central to their discussion is that hackers, whether they are PewDiePie fans or not, have not targeted a particular individual and do not seem to have vendetta towards a specific group of users. Rather it is a mass exploitation of existing ‘feature’ in the router, namely Universal Plug and Play (UPnP) to raise attention and awareness about an identified vulnerability in UPnP. This has also been seen as a way for hackers to simply show off and ‘boast’ about their discovery and hacking skills, or bring attention to a specific YouTube channel. In our data, we have also similar discussions in relation to a series of a controversial printer hacks by TheHackerGiraffe. In this hack the motivation was described by the hacker¹ and part of the community as a way to get the public’s attention to an existing vulnerability in network printer that allowed anyone outside the network access a users printer. In this specific case, the hacker was seen both as a ‘bad actor’ and a ‘concerned citizen’ by “drawing attention to a real issue in a fairly harmless way. There is a security issue here and it should be fixed.”, as one Reddit member put it.

In contrast to the examples documented above, we learned about cases in which the device owner misread the situation by assuming the hacking incident is a prank played on them by those whom they have shared the device with– or pranks that they took as evidence as they are being hacked. One common functionality that was being used and manipulated in these incidents is the two-way talking feature of the device (e.g. security camera or door bell) that allows the hacker-prankster speak with the people in the vicinity through the device’s microphone. While this feature is mainly available to the trusted individuals who have access to the device or the related app, there are ways to gain such access through exploiting vulnerabilities available in the network or the accounts connected to the app and device:

“My wyze cam pan was sitting next to me. Motion detection and the pan setting off. It was facing 45 deg from me. Suddenly I heard the speaker come on and the camera begin to rotate around. It faced me and looked back and forth between me and my dog. I would say it was just resetting or panning, but

the speaker came on like someone was talking through it”.

In this case, while community members are sharing similar incidents and suggesting solution, OP further provides an update saying that the hack was in fact an innocent prank: *“UPDATE: LOL MY GIRLFRIEND WAS FUCKING WITH ME. MY BAD FAM”.*

Although this case turns out without any reported harm, we learned of cases where the situation was initially perceived as a joke or a prank and then it was realised as a hack. In one example of this type, the OP has initially assumed the security camera incident is a prank played by their partner, the only person who has access to the device:

“Someone hacked my ring indoor camera by screaming to try to scare me and I thought it was my boyfriend who is the only one who has access to my camera. I immediately called my bf to ask if it was some kind of joke and while I was on the phone with him they were taunting me and my bf could hear them [...] They wanted to negotiate something with me and tried telling me to hang up the phone and that it wasn’t my boyfriend. I’m shaken and called 911 and the city police to file a report. I’m actually on the phone with Ring to see what happened”.

This OP later returned with an update about the incident after discussing it with the device manufacturer’s support team. While we do not get the full details of the event, we learn that the technology has been (mis)used by the hackers to gather information or compromising material on the OP to blackmail them:

“Turns out someone from the dark web stole my info. they tried getting money out of me by “negotiating” and then threatening me.”.

4.3 Dealing with the hack

Discussions of how to deal with the hack and finding a temporary or permanent solution for the problem is another characteristics of users’ posts. Similar to the collective efforts in making sense of the hack – who, how and whys – community members shared their own practices as well as their successful or failed stories and solutions. This sharing often vary from technical advice on how to ‘patch’ the problem by resetting the password associated with the device, to a more practical conversation on how to report and deal with the situation, or how to emotionally deal with being a hack victim.

4.3.1 Getting technical support

Apart from technical advice, such as password reset, many community members provided information and advice on how to increase security measures by setting up a two-factor authentication, as well as a more educational content on how to identify similar security vulnerabilities in other devices or in their network in order to prevent future similar incidents. For instance, in one case the OP discussed how occasionally

¹<https://darknetdiaries.com/transcript/31/>

their TV would turn on in the middle of the night without their permission. In response, one user suggested to start with changing their WiFi password in a more detailed manner, helping the OP to find their way in dealing technical difficulties of such a task, going through how to change the WiFi password, and how to check on the type of wireless encryption that was enabled.

One point to make here is related to the level of technical knowledge one needs in order to deal with ‘basic’ security functions in different devices. While users may be become accustomed to the basic requirements of keeping the device working in their domestic environment, for casual and non-technical users the topic of security can be overwhelmingly technical. The complexity of these connected devices has created complex requirements user, making them security-dependent on others [24, 36] or in this case online strangers who can help them understand ‘what is going on’ and what they should do.

4.3.2 Getting social and legal support

Alongside the technical support given in response, there was also often a practical discussion of who the OP could go to get help from others. This could span across law enforcement and the government (such as security agencies), and more prosaically help from the manufacturer. Indeed, in a few cases OPs themselves have mentioned contacting the police department as a practical legal and security practice in order to investigate the case.

The majority of advice for contacting the law enforcement came from community members, particularly in relation to those cases in which the attack required professional and technical attention related to an ongoing harassment. This also included those cases where children were affected or OP’s life and safety could be in danger. While for many users, reaching out to law enforcement agencies was seen as a legal action towards solving the problem or preventing the victim from further harm, there was also considerable suspicion about whether ordinary law enforcement would have any understanding of technical issues. Rather, the expectation was that law enforcement’s response could be used to give a warning to those, or a ‘fright’ to the likely perpetrators of a hack:

“You can try and call the police and show them evidence of your WiFi being duplicated and showing them the MAC addresses of the devices connecting to your wifi access point. There’s a good chance they’ll just have a talk with your neighbours but that might make them shit their pants enough that they stop”.

4.3.3 Dealing with harm and hurt

Several users reported the financial loss associated with purchasing a device that was now useless due to a decision to uninstall the device and replacing it with a trustable device

after the incident. But perhaps the biggest harm came from *emotional* burden of being hacked. Many users who experienced their IoT device being hacked, reported on different range of feelings, from being uncomfortable in having the technology at their home after the hack, to being scared of the ‘spookiness’ of the technology failures, to having a mixed feelings of confusion, anger, and worry that comes from not knowing for certain whether they have been hacked or not.

Perhaps the most devastating feeling reported comes from being unsure if the device is hacked or is being used by someone whom the users trust to share the device with, in a way that we do not understand. Many of the home IoT devices the users discussed were acquired in the first place for reasons of security and safety - to ensure the safety of themselves and their family members or to keep their home and property secure. In some cases though the vulnerabilities and problems reported by the users of these devices become, ironically, a new source of insecurity, anxiety and stress – stalkers digitally stalking the victim even at their intimate moments in their homes, and outsiders given unauthorised access to victim’s property remotely.

Such hacking incidents becomes particularly problematic, dangerous and harmful when children are involved or affected by the incident. In one instance, a parent reported of a traumatic experience when they realised their child was potentially subjected to security camera hack. In this case, their child could hear voices from the camera installed in their bedroom assuming it was the parent asking them to act upon a presumably ‘innocent’ request:

“I just unplugged the camera in my child’s room. This morning she came back in to wake me up and said the following: “Mom why did you talk on my speaker?” What? “You talked on my speaker.” When? “Right now. You said hey go to sleep.” Right now? “Yeah and I didn’t like it. You said stop playing and go to sleep.” I asked her if it was a mommy voice or a daddy voice, and she said mommy voice and then imitated it, whispering. And she said, “and I didn’t like it so I covered my ears and came in here.[”] I am FREAKED out and promptly went in her room and unplugged it”.

While parents who reported this specific case, fortunately, did not report any other incident after they disconnected the device, several parents reported the terror of hearing a stranger’s voice in their kid’s room via a hacked baby monitor, threatening to kidnap and harm their child [83]. The use of such technology to hijack the authority of the system owner in the eyes of someone being cared for – be that a child or other dependants – can not only cause emotional and (potential) physical harm but echos many confidence scams and man in the middle hacks [29, 60] and opens the path for the same categories of maleficence.

5 Discussion: Rethinking Hacks

As work in the SOUPS community has explored [52, 75, 84], cybersecurity has political, social, psychological and economic aspects. We find this becomes more important if, as we do here, we attempt to focus not on the hacks themselves but on the people who the hacks have impact upon. By focusing on these hacked ‘users’, we have attempted here to open up a new front in understanding both how hacks operate and the ongoing impacts they have.

5.1 Designing for being hacked

The most common question that is brought to the online forums we studied was “have I been hacked?”. At the heart of the user experience of hacking is users’ own uncertainty in their need for help. This suggests that the needs of users for support go much beyond technically detecting and blocking a hack (useful though that would of course be), but of helping the user in this situation more broadly.

In terms of design, this points to a number of directions. Beyond basic security help and information [8], there is often a need for diagnosing particular issues with particular devices and listing unusual behaviours that might be mistaken for being signs of a hack. So for example, for each device or service tools could help by summarising others’ experiences around suspecting or even being hacked. This could take the form of a knowledge base, or a tool that summarises forum interactions in some way. Such a knowledge base can offer a set of actions and tasks from which users could benefit from collecting evidence around the incident, their setup and any other data that often struggle to collect by themselves. Such data could help an outsider assist, be that law enforcement or security forensics, in diagnosing and assisting a hacking victim. This can be used as a design direction for supporting the manufacturer (or third party) providing support.

Indeed, it may be at times that what is needed is something that goes beyond direct support, yet also deals with their emotional needs. The forums themselves in different ways play a role here in that they provide a venue for support from others with dealing with the hack. The role they play is a sort of ‘technical counselling’ - with support spanning from help with the technology, of course, but also how the hack interferes with social relationships, assistance from the law, emotional support and even financial assistance. One interesting, and challenging, area of design would be to focus our attention on cases where users think they have been hacked but probably have not been hacked – what we called ‘non-hacked hacks’. As we described above, it is not users’ technical incompetence here that is to blame but often poor design decisions, as well as the inscrutability of IoT systems (that can only communicate with users through a flashing light without indicating a clear direction or purpose) can fail the user in detecting the problem or result in hypervigilant reactions towards unex-

pected actions. As technology is becoming more and more embedded in people’s everyday lives, as our data suggested, there is a need for additional technical solutions that helps users with the fluidity and integration of maintenance of IoT devices in their homes. Receiving a push notification on users’ mobile device can be one solution to help them understand if the flashing light is in fact related to an ongoing update or an indication of a hack.

Another suggestion to facilitate this approach is designing security tools that are tailored toward the needs of casual IoT users rather than network and security experts (e.g. [35]). A ‘white hat’ tool could communicate with other IoT devices located on the same network, scan logs and configurations to work out if there has likely been a hack, but also to broadly assist and reassure users who might be reasonably concerned by the unusual behaviour of their systems. While such a design can technically be complex (as it requires access to a set of diverse protocols and standards such as in [2]), designing for ‘non-hacked hacks’ might focus as much on reassuring users as detecting a hack. This could be as simple as documenting the different devices on a networks, and describing their usual failures and other unusual behaviours that other users have detected. While ‘secure by design’ has been a powerful guide in the cybersecurity world, it unfortunately removes users as active agents in the security process. In designing IoT security systems there may be opportunities for supporting users to go beyond what can be ‘designed in’ to a system as part of the development process. If we contrast technology with the case of automobiles, we can see how safety is not something that can be ‘designed’ during manufacture, it is an ongoing commitment supported by product recalls, testing institutes, safety certifications and so on. In this way we would argue for users’ involvement in ‘lifecycle security’, where security comes from supporting users in detecting, repelling and dealing with being hacked throughout the life of a product.

5.2 Cybernoia

Our data lets us move beyond thinking about hacks as mainly technical objects – as something that can be prevented through better security – to thinking about them as users’ experiences through how they discover and manage them, and in their relationships with others. Hacks exist not only as breaches in the security, but also breaches in the practices and understanding of end-users. Hacks by their very nature will always exist outside the knowledge and understanding of those who they impact, beyond the understanding of victims, at least initially. Preventing and supporting users then in dealing with hacks is not only a question of design or technical specification, but also one of supporting users’ understanding and engagement with their systems when things go wrong.

Hacks can have a considerable psychological impact on users. There are unfortunate ways in which hacks can also contribute to, or be part of, ongoing mental health conditions

suffered by the user. Paranoia - a feature of different mental health conditions - can lead to imagined hacks, but also the expansion of small mishaps or mistakes in a system to major incidents of victimisation. *Cybernoia* is a feature that may be ever more pressing as paranoia and technology use go together. This cybernoia is less frequently identified by the OPs, but usually it comes from posters who accuse the OP of inventing unlikely scenarios and being part of ridiculous ‘tin foil hat’ conspiracy theories.

Yet as we have outlined here, there is the need to realise that hacks exist when they are perceived by users – even if they are actually not hacks as technically defined. In many situations users cannot determine themselves if they are actually hacked or not, with the sometimes bizarre behaviour of systems giving users a reasonable (if sometimes unreasonable) belief that they have been hacked. For these situations the impact is as if the user actually had been hacked - as the famous phrase puts it “things imagined are real in their consequences” [59].

5.3 Security and relationships

There are a number of recent papers that argue that cybersecurity needs to take an explicitly feminist direction in understanding how technology can become part of abuse and even enabling violence and discrimination [50, 58, 74]. Building on this, our data contributes to an understanding of how security is a practice embedded in users’ relationships with others – the question of ‘who’ has hacked is as important as ‘how’ users were hacked. Indeed, the ways in which security is embedded in different social relationships that take place around IoT can create new forms of harm, insecurities and dependency.

Dealing with a hack necessarily involves going beyond expectations and current knowledge, requiring somewhat a level of trust. Thinking about the social aspects of hacks thus focuses attention on the relationships between the hacked and the hacker, between the organisations that make technical systems, and users who resort to different support resources to manage them. As our data shows, the forums we studied (and the users who contribute to them) play an important role in supporting users who find that the manufacturers have let them down in whatever way. Indeed, the level of support offered for much of the IoT that we focused on here can be rather poor and users found little support from either the organisations involved when their issues became serious. This led them to resort to Internet forums which can be seen as important sites to whoever tries to understand problems and to get support.

A different relationship which cybersecurity can become part of is that between family members. As a technology that is often used at home, IoT devices are frequently shared amongst family members. An Alexa smart speaker, for example, is available to everyone in the household and will be activated when the wake word is used regardless of users’ age.

Being ‘shared by default’ – sometimes just because they are physically in a shared family space – makes IoT potentially more useful, as something that goes beyond individual usage. But this also presents new challenges for IoT security since this becomes another aspect of devices that needs to be managed and shared across a family, with likely different users having a diverse set of security skills and knowledge of understanding how IoT ecosystems work and how the security is achieved. Maintaining ‘home security’ – in terms of IoT can then become a new point of dependency between household members, and at times then a new vulnerability for those who are newly dependent. Even if the technology itself tries to be diverse and accessible – affecting users without technical background [81] – it actually results in new unwanted dependencies and inequalities. While asynchronous knowledge and control over an IoT device can create whimsical and fun moments of playing pranks on other household residents, it also can result in exposing these residents “particularly women, to unique privacy and security risks.” [76].

6 Conclusion

In this paper we have sought to return hack victims themselves to understand what it is to be hacked. Using online reports of hacks, we reviewed 210 self reports of hacks to identify the role that uncertainty plays, but also more broadly how users understand and deal with the experience of having their home IoT systems hacked in some way. Our focus on IoT lets us explore technologies which while still in flux, are increasingly embedded into our world and homes. Vulnerabilities in IoT are then especially worrying.

Indeed, the growth in their acceptance and use of IoT suggests that their use may not only because their use may become increasingly involuntary, but IoT may become as commonplace as ordinary ‘non-smart’ devices are today. This then means that the victims’ stories that we identified here may move from unusual examples, to be a much more widespread phenomena. As we talk of ‘early adopters’ of technology, our users may actually be ‘early victims’, with their stories and experiences offering a broader warning about IoT and cybersecurity more generally.

In doing so we follow the long tradition in SOUPS of putting the social back into the technical - the hack as both as social and a technical object.

Acknowledgements

This research is partially funded by a Digital Futures post-doctoral research grant, and the Swedish Research Council (Vetenskapsrådet) under the project Securing Things (2017-04804: Säkra saker: Säkra sakernas Internet).

References

- [1] Omnia Abu Waraga, Meriem Bettayeb, Qassim Nasir, and Manar Abu Talib. Design and implementation of automated IoT security testbed. *Computers & Security*, 88:101648, January 2020.
- [2] Abbas Acar, Hossein Fereidooni, Tigist Abera, Amit Kumar Sikder, Markus Miettinen, Hidayet Aksu, Mauro Conti, Ahmad-Reza Sadeghi, and Selcuk Uluagac. Peek-a-boo: i see your smart home activities, even encrypted! In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks, WiSec '20*, pages 207–218, New York, NY, USA, July 2020. Association for Computing Machinery.
- [3] Katherine Albrecht and Liz McIntyre. Privacy Nightmare: When Baby Monitors Go Bad [Opinion]. *IEEE Technology and Society Magazine*, 34(3):14–19, September 2015.
- [4] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16*, pages 3906–3918, New York, NY, USA, May 2016. Association for Computing Machinery.
- [5] Carmelo Ardito, Regina Bernhaupt, Philippe Palanque, and Stefan Sauer. Handling Security, Usability, User Experience and Reliability in User-Centered Development Processes. In David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris, editors, *Human-Computer Interaction – INTERACT 2019*, Lecture Notes in Computer Science, pages 759–762, Cham, 2019. Springer International Publishing.
- [6] Paul Atkinson and David Silverman. Kundera's Immortality: The Interview Society and the Invention of the Self. *Qualitative Inquiry*, 3(3):304–325, September 1997.
- [7] Rosanna Bellini, Emily Tseng, Nora McDonald, Rachel Greenstadt, Damon McCoy, Thomas Ristenpart, and Nicola Dell. "So-called privacy breeds evil": Narrative Justifications for Intimate Partner Surveillance in Online Forums. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):210:1–210:27, January 2021.
- [8] Brennen Bouwmeester, Elsa Rodríguez, Carlos Gañán, Michel van Eeten, and Simon Parkin. "The thing Doesn't Have a Name": Learning from emergent real-world interventions in smart home security. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 493–512. USENIX Association, August 2021.
- [9] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, January 2006.
- [10] Barry Brown, Stuart Reeves, and Scott Sherwood. Into the wild: Challenges and opportunities for field trial methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1657–1666. Association for Computing Machinery, New York, NY, USA, May 2011.
- [11] Barry Brown, Minna Vigren, Asreen Rostami, and Mareike Glöss. Why users hack: Conflicting interests and the political economy of software. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW), 2022.
- [12] Barry Brown, Alexandra Weilenmann, Donald McMillan, and Airi Lampinen. Five Provocations for Ethical HCI Research. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 852–863. Association for Computing Machinery, New York, NY, USA, May 2016.
- [13] Amy Bruckman, Kurt Luther, and Casey Fiesler. When Should We Use Real Names in Published Accounts of Internet Research? In *Digital Research Confidential: The Secrets of Studying Behavior Online*, pages 243–258. MIT Press, 2016.
- [14] Mark Button, Dean Blackburn, Lisa Sugiura, David Shepherd, Richard Kapend, and Victoria Wang. From feeling like rape to a minor inconvenience: Victims' accounts of the impact of computer misuse crime in the United Kingdom. *Telematics and Informatics*, 64:101675, November 2021.
- [15] Mark Button and Cassandra Cross. *Cyber Frauds, Scams and Their Victims*. Routledge, London, May 2017.
- [16] Rainara M. Carvalho, Rossana M.C. Andrade, Káthia M. Oliveira, and Christophe Kolski. Catalog of Invisibility Requirements for UbiComp and IoT Applications. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 88–99. IEEE, August 2018.
- [17] Anna C. Cavender, Daniel S. Otero, Jeffrey P. Bigham, and Richard E. Ladner. Asl-stem forum: Enabling sign language to grow through online collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2075–2078. Association for Computing Machinery, New York, NY, USA, April 2010.

- [18] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, Shwetak N. Patel, and Julie A. Kientz. Investigating receptiveness to sensing and inference in the home using sensor proxies. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12*, pages 61–70, New York, NY, USA, September 2012. Association for Computing Machinery.
- [19] E. Gabriella Coleman. *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton University Press, Princeton, December 2012.
- [20] Dan Conway, Ronnie Taib, Mitch Harris, Kun Yu, Shlomo Berkovsky, and Fang Chen. A Qualitative Investigation of Bank Employee Experiences of Information Security and Phishing. In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*, pages 115–129, 2017.
- [21] Cassandra Cross. No laughing matter: Blaming the victim of online fraud. *International Review of Victimology*, 21(2):187–204, May 2015.
- [22] Cassandra Cross, Megan Parker, and Daniel Sansom. Media discourses surrounding ‘non-ideal’ victims: The case of the Ashley Madison data breach. *International Review of Victimology*, 25(1):53–69, January 2019.
- [23] Brittany D. Davis, Janelle C. Mason, and Mohd Anwar. Vulnerability Studies and Security Postures of IoT Devices: A Smart Home Case Study. *IEEE Internet of Things Journal*, 7(10):10102–10110, October 2020.
- [24] Paul Dourish, Rebecca E. Grinter, Jessica Delgado De La Flor, and Melissa Joseph. Security in the wild: User strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 2004.
- [25] Brianna Dym and Casey Fiesler. Ethical and privacy considerations for research using online fandom data. *Transformative Works and Cultures*, 33, June 2020.
- [26] Milène Fauquex, Sidhant Goyal, Florian Evequoz, and Yann Bocchi. Creating people-aware IoT applications by combining design thinking and user-centered design methods. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, pages 57–62, December 2015.
- [27] Jessica L. Feuston and Anne Marie Piper. Everyday Experiences: Small Stories and Mental Illness on Instagram. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pages 1–14, New York, NY, USA, May 2019. Association for Computing Machinery.
- [28] Casey Fiesler and Nicholas Proferes. “participant” perceptions of twitter research ethics. *Social Media + Society*, 4(1):2056305118763366, 2018.
- [29] Peter Fischer, Stephen E. G. Lea, and Kath M. Evans. Why do individuals respond to fraudulent scam communications and lose money? the psychological determinants of scam compliance. *Journal of Applied Social Psychology*, 43(10):2060–2072, 2013.
- [30] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. A Stalker’s Paradis: How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, pages 1–13, New York, NY, USA, April 2018. Association for Computing Machinery.
- [31] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):46:1–46:22, December 2017.
- [32] Márcio Miguel Gomes, Rodrigo da Rosa Righi, and Cristiano André da Costa. Future directions for providing better IoT infrastructure. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct*, pages 51–54, New York, NY, USA, September 2014. Association for Computing Machinery.
- [33] Colin M. Gray, Shruthi Sai Chivukula, and Ahreum Lee. What Kind of Work Do "Asshole Designers" Create? describing Properties of Ethical Concern on Reddit. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 61–73, Eindhoven Netherlands, July 2020. ACM.
- [34] Xinning Gui, Yubo Kou, Kathleen H. Pine, and Yunan Chen. Managing Uncertainty: Using Social Media for Risk Assessment during a Public Health Crisis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17*, pages 4520–4533, New York, NY, USA, May 2017. Association for Computing Machinery.
- [35] Hassan Habibi Gharakheili, Arunan Sivanathan, Ayyoob Hamza, and Vijay Sivaraman. Network-Level Security for the Internet of Things: Opportunities and Challenges. *Computer*, 52(8):58–62, August 2019.
- [36] Richard Harper, editor. *Inside the Smart Home*. Springer-Verlag, London, 2003.
- [37] Billy Henson, Bradford W. Reynolds, and Bonnie S. Fisher. Cybercrime victimization. In *The Wiley Handbook on the Psychology of Violence*, pages 555–570. Wiley Blackwell, Hoboken, NJ, US, 2016.

- [38] Thomas Holt, Deborah Strumsky, Olga Smirnova, and Max Kilger. Examining the Social Networks of Malware Writers and Hackers. *International Journal of Cyber Criminology*, 6, January 2012.
- [39] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, New York, NY, USA, April 2020.
- [40] Jina Huh. Clinical Questions in Online Health Communities: The Case of "See your doctor" Threads. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15*, pages 1488–1499, New York, NY, USA, February 2015. Association for Computing Machinery.
- [41] Jina Huh and Wanda Pratt. Weaving clinical expertise in online health communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14*, pages 1355–1364, New York, NY, USA, April 2014. Association for Computing Machinery.
- [42] Max Ingham, Jims Marchang, and Deepayan Bhowmik. IoT Security Vulnerabilities and Predictive Signal Jamming Attack Analysis in LoRaWAN. *IET Information Security*, January 2020.
- [43] Andreas Jacobsson and Paul Davidsson. Towards a model of privacy and security for smart homes. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, pages 727–732, December 2015.
- [44] Tim Jordan and Paul Taylor. A Sociology of Hackers. *The Sociological Review*, 46(4):757–780, November 1998.
- [45] P. Karthika, R. Ganesh Babu, and P. A. Karthik. Fog Computing using Interoperability and IoT Security Issues in Health Care. In Devendra Kumar Sharma, Valentina Emilia Balas, Le Hoang Son, Rohit Sharma, and Korhan Cengiz, editors, *Micro-Electronics and Telecommunication Engineering, Lecture Notes in Networks and Systems*, pages 97–105, Singapore, 2020. Springer.
- [46] Megan Knittel, Faye Kollig, Abrielle Mason, and Rick Wash. Anyone else have this experience: Sharing the Emotional Labor of Tracking Data About Me. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):79:1–79:30, April 2021.
- [47] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. Motivations to participate in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1927–1936. Association for Computing Machinery, New York, NY, USA, April 2010.
- [48] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, Are You Listening? privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):102:1–102:31, November 2018.
- [49] Derek Layder. *Sociological Practice: Linking Theory and Social Research*. SAGE, September 1998.
- [50] Roxanne Leitão. Anticipating Smart Home Security and Privacy Threats with Survivors of Intimate Partner Abuse. In *Proceedings of the 2019 on Designing Interactive Systems Conference, DIS '19*, pages 527–539, New York, NY, USA, June 2019. Association for Computing Machinery.
- [51] Eric Rutger Leukfeldt, R. J. (Raoul) Notté, and M. (Marijke) Malsch. Exploring the Needs of Victims of Cyber-dependent and Cyber-enabled Crimes. *Victims & Offenders*, 15(1):60–77, January 2020.
- [52] Karen Levy and Bruce Schneier. Privacy Threats in Intimate Relationships. SSRN Scholarly Paper ID 3620883, Social Science Research Network, Rochester, NY, June 2020.
- [53] Isabel Lopez-Neira, Trupti Patel, Simon Parkin, George Danezis, and Leonie Tanczer. 'Internet of Things': How Abuse is Getting Smarter. SSRN Scholarly Paper ID 3350615, Social Science Research Network, Rochester, NY, March 2019.
- [54] Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15*, pages 1511–1526, New York, NY, USA, February 2015. Association for Computing Machinery.
- [55] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. Collective Sensemaking in Online Health Forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3217–3226. Association for Computing Machinery, New York, NY, USA, April 2015.
- [56] Michael Massimi, Jackie L. Bender, Holly O. Witteman, and Osman H. Ahmed. Life transitions and online

- health communities: Reflecting on adoption, use, and disengagement. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1491–1501, New York, NY, USA, February 2014. Association for Computing Machinery.
- [57] Nora McDonald, Karla Badillo-Urquiola, Morgan G. Ames, Nicola Dell, Elizabeth Keneski, Manya Sleeper, and Pamela J. Wisniewski. Privacy and Power: Acknowledging the Importance of Privacy Research and Design for Vulnerable Populations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA, April 2020. Association for Computing Machinery.
- [58] Dana McKay and Charlynn Miller. Standing in the Way of Control: A Call to Action to Prevent Abuse through Better Design of Smart Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–14, New York, NY, USA, May 2021. Association for Computing Machinery.
- [59] Robert K. Merton. The Thomas Theorem and the Matthew Effect. *Social Forces*, 74(2):379–422, 1995.
- [60] Gopi Nath Nayak and Shefalika Ghosh Samaddar. Different flavours of Man-In-The-Middle attack, consequences and feasible solutions. In *2010 3rd International Conference on Computer Science and Information Technology*, volume 5, pages 491–495, July 2010.
- [61] Andrea Grimes Parker, Ian McClendon, Catherine Grevet, Victoria Ayo, WonTaek Chung, Veda Johnson, and Elizabeth D. Mynatt. I am what i eat: Identity & critical thinking in an online health forum for kid. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2437–2446. Association for Computing Machinery, New York, NY, USA, April 2013.
- [62] Simon Parkin, Trupti Patel, Isabel Lopez-Neira, and Leonie Tanczer. Usability analysis of shared device ecosystem security: Informing support for survivors of IoT-facilitated tech-abuse. In *Proceedings of the New Security Paradigms Workshop*, NSPW '19, pages 1–15, New York, NY, USA, September 2019. Association for Computing Machinery.
- [63] Kari Paul. Dozens sue Amazon's Ring after camera hack leads to threats and racial slurs. *The Guardian*, December 2020.
- [64] James Pierce. Smart Home Security Cameras and Shifting Lines of Creepiness: A Design-Led Inquiry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Machinery, New York, NY, USA, May 2019.
- [65] Rebecca S. Portnoff, Linda N. Lee, Serge Egelman, Pratyush Mishra, Derek Leung, and David Wagner. Somebody's Watching Me? assessing the Effectiveness of Webcam Indicator Lights. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1649–1658. Association for Computing Machinery, New York, NY, USA, April 2015.
- [66] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2):20563051211019004, April 2021.
- [67] Sabirat Rubya and Svetlana Yarosh. Video-Mediated Peer Support in an Online Community for Recovery from Substance Use Disorders. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1454–1469, New York, NY, USA, February 2017. Association for Computing Machinery.
- [68] Harvey Sacks. On doing “being ordinary”. In J. Maxwell Atkinson, editor, *Structures of Social Action*, Studies in Emotion and Social Interaction, pages 413–429. Cambridge University Press, Cambridge, 1985.
- [69] Harvey Sacks. Spring 1968: April 24 Second Stories. In *Lectures on Conversation Volume I (Edited by Gail Jefferson)*, chapter 7, pages 749–805. John Wiley & Sons, Ltd, 1995.
- [70] Mattia Samory, Vincenzo-Maria Cappelleri, and Enoch Peserico. Quotes Reveal Community Structure and Interaction Dynamics. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 322–335, New York, NY, USA, February 2017. Association for Computing Machinery.
- [71] Amirali Sanatinia and Guevara Noubir. OnionBots: Subverting Privacy Infrastructure for Cyber Attacks. In *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 69–80, June 2015.
- [72] Pedro Sanches, Vasiliki Tsaknaki, Asreen Rostami, and Barry Brown. Under Surveillance: Technology Practices of those Monitored by the State. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, New York, NY, USA, April 2020.

- [73] Mike Simmonds. How businesses can navigate the growing tide of ransomware attacks. *Computer Fraud & Security*, 2017(3):9–12, March 2017.
- [74] Julia Slupska. Safe at Home: Towards a Feminist Critique of Cybersecurity. SSRN Scholarly Paper ID 3429851, Social Science Research Network, Rochester, NY, May 2019.
- [75] Julia Slupska, Scarlet Dawson Dawson Duckworth, Linda Ma, and Gina Neff. Participatory Threat Modelling: Exploring Paths to Reconfigure Cybersecurity. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–6, New York, NY, USA, May 2021. Association for Computing Machinery.
- [76] Yolande Strengers and Jenny Kennedy. *The Smart Wife: Why Siri, Alexa, and Other Smart Home Devices Need a Feminist Reboot*. MIT Press, Cambridge, MA, USA, September 2020.
- [77] Lisa Sugiura, Rosemary Wiles, and Catherine Pope. Ethical challenges in online research: Public/private perceptions. *Research Ethics*, 13(3-4):184–199, July 2017.
- [78] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. "I don't own the data": End User Perceptions of Smart Home Device Data Practices and Risks. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, 2019.
- [79] Huixin Tian, Chris Kanich, Jason Polakis, and Sameer Patil. Tech Pains: Characterizations of Lived Cybersecurity Experiences. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, pages 250–259, September 2020.
- [80] Orly Turgeman-Goldschmidt. Hackers' Accounts: Hacking as a Social Entertainment. *Social Science Computer Review*, 23(1):8–23, February 2005.
- [81] EQUALS Skills Coalition UNESCO. I'd blush if I could: Closing gender divides in digital skills through education, 2019.
- [82] Johan van Wilsem. Hacking and Harassment—Do They Have Something in Common? comparing Risk Factors for Online Victimization. *Journal of Contemporary Criminal Justice*, 29(4):437–453, November 2013.
- [83] Amy B Wang. Nest cam security breach: A hacker took over a baby monitor and broadcast threats, Houston parents say - The Washington Post. 2018.
- [84] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 1–16, New York, NY, USA, July 2010. Association for Computing Machinery.
- [85] Rick Wash. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):160:1–160:28, October 2020.
- [86] Rick Wash and Molly M. Cooper. Who Provides Phishing Training? facts, Stories, and People Like Me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery, New York, NY, USA, April 2018.
- [87] Meredydd Williams, Jason R. C. Nurse, and Sadie Creese. Privacy is the Boring Bit: User Perceptions and Behaviour in the Internet-of-Things. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pages 181–18109, August 2017.
- [88] Peter Worthy, Ben Matthews, and Stephen Viller. Trust Me: Doubts and Concerns Living with the Internet of Things. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pages 427–434, New York, NY, USA, June 2016. Association for Computing Machinery.
- [89] Ibrar Yaqoob, Ejaz Ahmed, Muhammad Habibur Rehman, Abdelmuttlib Ibrahim Abdalla Ahmed, Mohammed Ali Al-garadi, Muhammad Imran, and Mohsen Guizani. The rise of ransomware and emerging security challenges in the Internet of Things. *Computer Networks*, 129:444–458, December 2017.
- [90] Randall Young, Lixuan Zhang, and Victor R. Prybutok. Hacking into the Minds of Hackers. *Information Systems Management*, 24(4):281–287, October 2007.
- [91] Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security and Privacy Concerns with Smart Homes. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security*, SOUPS '17, pages 65–80, USA, July 2017. USENIX Association.
- [92] Leah Zhang-Kennedy, Hala Assal, Jessica Rocheleau, Reham Mohamed, Khadija Baig, and Sonia Chiasson. The aftermath of a crypto-ransomware attack at a large academic institution. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 1061–1078, 2018.
- [93] Jane Y. Zhao, Evan G. Kessler, Jihneeh Yu, Kabir Jalal, Clairice A. Cooper, Jeffrey J. Brewer, Steven D. Schwaitzberg, and Weidun Alan Guo. Impact of Trauma Hospital Ransomware Attack on Surgical Residency Training. *Journal of Surgical Research*, 232:389–397, December 2018.
- [94] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User Perceptions of Smart Home IoT Privacy. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):200:1–200:20, November 2018.

Appendix A

Table 1: List of forums and subreddited

/r/homedefense	/r/talesfromtechsupport
/r/wyzecam	/r/cybersecurity
/r/Ring	/r/PS4
/r/HomeNetworking	/r/HayDay
/r/hacking	/r/privacy
/r/blinkcameras	/r/LegalAdviceUK
/r/raisedbynarcissists	/r/sonos
/r/nosleep	/r/homesecurity
/r/Hue	/r/ChoosingBeggars
/r/homeautomation	https://www.amazonforum.com
/r/galaxys10	https://community.norton.com/en/forums
/r/techsupport	https://forums.tomsguide.com
/r/smarthome	https://security.stackexchange.com/questions
/r/talesfromcallcenters	https://community.bt.com
/r/hometheater	https://forum.telus.com
/r/samsung	https://discussions.apple.com
/r/bravia	https://www.bleepingcomputer.com/forums
/r/dataisbeautiful	https://forum.level1techs.com/
/r/teslamotors	https://en.community.sonos.com
/r/Chromecast	https://answers.microsoft.com/en-us/xbox/forum
/r/googlehome	https://forums.wyzecam.com
/r/NoStupidQuestions	https://community.ring.com/
/r/funny	https://www.amazon.com/gp/customer-reviews/
/r/mildlyinteresting	https://forum.yitechnology.com/
/r/appletv	https://forums.wyzecam.com/
/r/PlayStationPlus	https://forums.tesla.com/discussion
/r/PewdiepieSubmissions	https://support.google.com/chromecast
/r/amazonecho	https://us.community.samsung.com
/r/alexa	https://community.blinkforhome.com
/r/googlehome	https://community.tp-link.com
/r/cybersecurity	https://lgcommunity.us.com/discussion
/r/techsupport	https://teslamotorsclub.com/

Appendix B

Table 2: Main categories of hacked devices with examples

Category	Example
Smart home devices	Amazon echo, Echo dot, Chromecast, Google home, Hue lights
Router and wifi	Archer C1200
Accounts, Game console and computers	Google account, Xbox, PS4
Smart locks	Ring doorbell
Phones and tablets	Apple, Samsung
Printer	Variety of models
Security camera	Wayz, Yi, Ring, Arlo
Smart speaker	Sonos
Smart tv	Sony, LG
Vehicle	Tesla

Table 3: Examples of high level codes. Note that each code can have multiple sub-codes and each post can be assigned multiple codes

Action taken	Hacker_family
Addressing the hacker	Hacker_neighbour
Addressing the manufacturer	Hacker_suspicious
Analysis of the hack	Hacker_unknown
Asking help from the forum	Harm
Comments on forum culture	Innovative tactics to solve the problem
Creepiness	Jokes
Paranoia	Lack of tech expertise with IoT
Cybersecurity	Manufacturer reply
Cybersecurity education	Hacker_expartner
Debating cybersecurity	Mental health
Description of the problem	Not a hack but
Distrust in police help	Other evidence
Evidence of the hack	Own expertise
Existing security practice	Paranoia
Forum reply_advice	Reasons to have the device
Forum reply_analysis of the hack	Shaming_questioning
Forum reply_comments on cybersecurity	Sharing own cybersecurity practices
Forum reply_debating other reply	Sharing own story
Forum reply_doubting the story	Type of post_asking for help
Forum reply_sharing own hacking story	Type of post_attention from manufacturer
Getting back at the hacker	Type of the device
Hacker_partner	What happened_the hack

Runtime Permissions for Privacy in Proactive Intelligent Assistants

Nathan Malkin[◦], David Wagner[◦], and Serge Egelman[†]

[◦]*University of California, Berkeley*

[†]*International Computer Science Institute*

Abstract

Intelligent voice assistants may soon become proactive, offering suggestions without being directly invoked. Such behavior increases privacy risks, since proactive operation requires continuous monitoring of conversations. To mitigate this problem, our study proposes and evaluates one potential privacy control, in which the assistant requests permission for the information it wishes to use immediately after hearing it.

To find out how people would react to runtime permission requests, we recruited 23 pairs of participants to hold conversations while receiving ambient suggestions from a proactive assistant, which we simulated in real time using the Wizard of Oz technique. The interactive sessions featured different modes and designs of runtime permission requests and were followed by in-depth interviews about people’s preferences and concerns. Most participants were excited about the devices despite their continuous listening, but wanted control over the assistant’s actions and their own data. They generally prioritized an interruption-free experience above more fine-grained control over what the device would hear.

1 Introduction

For many systems, privacy is an afterthought, with mitigations added after users have already adopted the product. This paper aims to reverse that trend by studying privacy solutions for a still-nascent technology: proactive intelligent assistants.

Smart speakers and other forms of voice assistants are highly popular, reaching hundreds of millions of people around the world [50]. Today, they are mostly invoked through

wake-words (e.g., “hey Siri”), but developers have deployed or are experimenting with more proactive features, such as reacting to sounds [99], identifying commands proactively [74], or removing wake-words altogether [55, 96]. Research prototypes have gone beyond this by offering contextually relevant information based on the content of conversations [12, 65, 82, 91]. In this project, we aim to prepare for the possibility that this technology becomes commonly available in the future.

Proactivity and contextual suggestions rely on the assistant continuously recording conversations, which is a clear privacy risk that will compound the many concerns people already have about smart speakers [2, 27, 42, 54, 60]. Nevertheless, consumers appear interested in this technology [85, 90], so we should not expect them to reject it outright. Instead, we need to find ways to improve the privacy of those who do adopt it.

One way to restrict what assistants hear can be through permissions, such as those used by smartphones to limit apps’ access to sensitive resources like location or camera. In fact, existing voice assistants already rely on permissions: Alexa, for example, shows them when installing “skills” (third-party add-ons) that access certain information, such as users’ names, addresses, or emails [8]. However, research has shown that install-time permissions are ineffective due to issues with attention and comprehension [36, 37, 46, 78]. As a result, in the mobile context, they have been largely supplanted by runtime permissions (i.e., asking at the time of data access) [23, 41].

Would runtime permissions be an effective privacy control for proactive assistants? Our study aims to investigate this question. To explore it, we simulated the experience of interacting with a proactive voice assistant for 23 pairs of participants. They tested several different permissions designs, triggered by different “apps” during the interactive session, and were interviewed about their preferences. This paper reports the themes that emerged. Our results help illuminate the design space of permissions for intelligent assistants and allow us to offer recommendations for this nascent technology.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

2 Related work

This section surveys existing work that our study builds on.

Proactive assistants Proactive assistants are a specific instance of ambient computing, which has seen considerable research in the field of human-computer interaction. We draw inspiration for the behavior of the assistant in our study from the following examples. The Ambient Spotlight [49] automatically searched for files relevant to a recorded meeting. Carrascal et al. [24] studied how to surface important details from transcribed phone calls. IdeaWall [82] ambiently displayed web search results relevant to conversations happening in real time. Similarly, Andolina et al. [12] developed a proactive search agent to assist people in natural conversations. Brown et al. [21] and McGregor et al. [65] focused specifically on meetings and automatically identifying action items that the computer could execute. Tabassum et al. [85] had participants propose proactive services based on real-life conversations. Wei et al. [91] prototyped a proactive smart speaker that used contextual awareness to pick opportune moments to engage with its users, in order to support medication reminders and other health and fitness interventions. Völkel et al. [90] prompted participants to imagine dialogues with a perfect voice assistant, finding that people want them to have detailed knowledge about the user and behave proactively. We modeled the assistant in our study on these examples, deciding that it would listen continuously to conversations, proactively perform web searches, and ambiently display their results to the user. Our work further contributes to this literature by reporting people’s experiences using a proactive assistant.

Privacy concerns Our goal of developing effective privacy controls for proactive voice assistants is motivated by the threats they pose and the privacy concerns even existing (i.e., *not* always-listening) devices elicit. Since permissions are meant to safeguard particularly-sensitive resources, we draw on the literature about privacy concerns to understand what people consider most worth protecting.

Privacy concerns are ubiquitous among smart device users, both administrators [97, 98] and especially secondary users [38, 51, 95]. Furthermore, researchers have found that people have heightened privacy expectations when it comes to voice interactions [27, 54], and voice assistants elicit special concerns [61]. Lau et al. [56] found that concerns are present, but distinct, among users and non-users. A common finding has been of gaps in users’ understanding of their devices. Abdi et al. [2] found incomplete threat models; Malkin et al. [60] discovered incorrect beliefs about data retention; Major et al. [58] identified confusion about third-party skills; and Huang et al. [42] observed suboptimal risk management strategies. This paper contributes to this literature by documenting privacy concerns about proactive assistants.

Install-time and runtime permissions A key motivation of this study’s focus on runtime permissions were findings

on limitations of install-time permissions. In smartphones, when users had to review permissions before installing apps, studies found low attention and comprehension rates, which were only slightly improved by redesigned interfaces [47] and nudges [6]. Interviews by Kelley et al. [46] showed that people did not understand permissions, a finding confirmed by Felt et al. [37], whose surveys and lab studies also found that only 17% of users paid attention to permissions.

As a result of the limitations of install-time permissions, smartphone platforms have largely moved to relying on runtime permissions [41], in which requests are issued when the app attempts to access data. While showing improved performance, runtime permissions have their own limitations. They are typically implemented as “ask on first use,” but studies have shown that people want to deny some requests even if they approved the initial one [92]. Users still misunderstand things, for example the scope of the requests [81], though this can be improved by better timing and explanations [34]. One of the main contributions of our study is testing such runtime permissions in a novel context—proactive assistants—and documenting users’ reactions and potential pitfalls.

3 Methods

Here, we describe our approach to investigating whether runtime permissions could provide effective privacy controls.

3.1 Assumptions

Proactive assistant devices do not exist yet, so, in order to have a concrete basis for our study, we needed to make a variety of assumptions and design choices. We note that these represent just one possible set of options in a large design space.

Threat model Modern voice assistant ecosystems encompass several layers of trust. In addition to their core first-party functionality, they feature tens of thousands of third-party apps [7] (also known as “skills” or “actions”), which have been the source of a number of privacy and security vulnerabilities [25, 53, 67]. In this study, we assumed that *platforms are trusted with all audio* and are responsible for administering permissions, and our permission system’s task is to mediate and *restrict third-party apps’ access to speech*. Specifically, the system should deny any attempt to access information not relevant to an app’s stated purpose.

A limitation of this threat model is that users may distrust the assistants’ manufacturers [2] and struggle to distinguish them from their apps [58]. However, the primary alternative is for privacy controls to be implemented by a trusted third party; but who might they be and why should users trust them? We therefore believed our simplification would lead to fewer hypotheticals for our participants. Moreover, any findings about permission systems with this model are likely to be applicable in settings where the assistant is also distrusted.

Architecture Runtime permission requests may feature in a variety of different assistant architectures, and the experiments in this paper could inform any of them. However, to make it clear why permissions in our study refer to specific information, we now describe a particular architecture, which is the basis of our study’s permission implementation.

Under our *network-restricted architecture*, third-party applications gain full access to all audio, but run completely sandboxed from the outside world.¹ Most apps will still require some online functionality (to get or receive data) and apps *are* allowed to make network requests, but any user content must be in the form of transcript snippets, and they must be reviewed and approved by the user.² The following is a sample sequence of events for a weather app:

1. The user says something. (“Is it warm in Hawaii?”)
2. The app decides this speech is relevant to it. (Per our architecture assumptions, this happens in a sandbox.)
3. The app identifies the information it wants to share over the network. (e.g., the location, Hawaii)
4. The user is then shown the permission request, if appropriate. (“May the weather app share ‘Hawaii?’”)
5. If the user approves, the requested information can be sent to the server.

3.2 Permission frequency

The user experience of runtime permissions has many parameters [34]. For us, one of the main ones is whether every data access attempt generates a user-visible permission request.

Ask every time One option is to ask the user every time an app wants access to a sensitive resource. This guarantees that a human reviews and assents to every permission request. However, frequent or repeated requests are likely to annoy users and result in fatigue [5] and habituation [88, 89].

Ask on first use (“Rules”) Smartphone permission systems, where asking every time is impossible [92], show a permission dialog once per resource, per app. The risk of this approach is that an app could make an appropriate permission request the first time around, but then later access the same resource at inappropriate times [81]. We felt that a higher degree of restriction would be appropriate for proactive services and therefore extended the ask-on-first-use design to scope an app’s access to a specific entity or type of speech. Examples of subjects for *Rules* include locations, date, numbers, types of speech, categories of physical objects, or emotions:

- Always allow the weather app access to locations
- Always allow the events app access to dates and times
- Always allow the supermarket app access to groceries

¹One implementation is for apps to run on the device itself. Current computational constraints make this challenging, but it may be less so in the future. Alternately, the sandbox could be on manufacturer-controlled servers.

²Side-channel attacks are possible, but are out of scope in this work.

Contextually relevant permissions (“Learning”) In different permission contexts, researchers have trained machine learning models to predict whether people would allow or deny a given permission request [17, 26, 29, 32, 57, 93]. We hypothesized that a similar system may be possible for proactive assistants. We leave the exact details of this *Learning* approach implementation unspecified, as we believe that it may not be feasible with today’s natural language processing capabilities. Instead, we study an idealized version of what might plausibly become possible at some point in the future.

We selected the above modes for our study because we considered them representative and easiest to explain to participants. Other possibilities include randomizing requests, asking for user involvement only on anomalous requests (e.g., weather app accessing food), or aggregating permissions and asking users to review all requests that happened during a given period (e.g., once a week).

3.3 Study design

At a high level, our study encompassed three activities—explanation, interaction, and interview—that repeated three times: once for each of the permission modes (§3.2). We chose a within-subjects design to allow participants to reflect on the differences between the modes and express their preferences.

Our introduction included a demonstration of the “features” of the assistant, including the runtime permissions. This was followed by an interactive session where participants engaged with the assistant. The first interactive session lasted five minutes and featured the ask-every-time permission design. The two subsequent sessions were each 10 minutes long, testing the *Rules* and *Learning* designs in randomized order.

Interactive simulation We simulated the experience of a proactive assistant for our participants, providing a realistic interface, but with a researcher performing the actions expected from the software. This “Wizard of Oz” technique is common in user experience research [31, 45, 62, 76]. The interface took the form of a smart display, such as Echo Show and Nest Hub and inspired by research prototypes from ambient computing [12, 82]. The “assistant” would passively listen to conversations and ambiently display relevant suggestions. To ensure more natural dialogue, we recruited participants in pairs of people who already knew each other.

Wizard of Oz implementation Our study was conducted remotely, over a video call. For the interactive portion of the study, the interviewer shared their screen, which contained a browser window showing the presentation view of a rapid prototyping tool;³ this represented the assistant’s display. The interviewer would update the screen, as quickly as possible, based on conversation content and commands.

The content on-screen would be either a permission request or (if permission had been granted) information relevant to

³<https://www.figma.com>

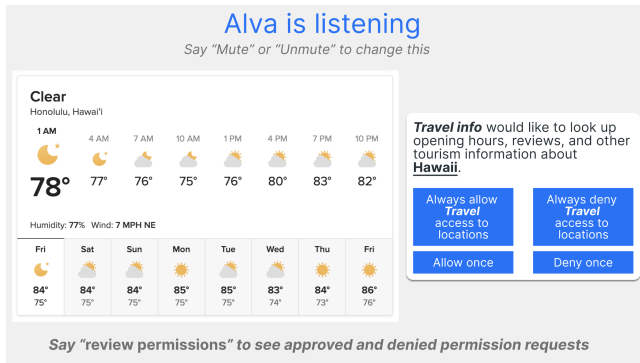


Figure 1: **Sample interface view**, as seen by participants, with the *Rules* design

the discussion topic (see Figure 1). Examples of the latter included weather, tourist information, ticket prices, etc. To accomplish this, the interviewer entered relevant keywords into a search engine, took screenshots of the summary boxes returned, and pasted the screenshots into the prototyping tool. For the permission requests, we had pre-made templates for each app, which the interviewer updated with speech from the participants, then brought into the viewport.

Due to the manual nature of the simulation, there was an average delay of approximately 5–25 seconds between when participants said something and when the corresponding visual appeared on screen. We warned participants about this delay upfront, and while many commented on it, others found it acceptable even for a real system.

Task selection To guide people’s conversations and ensure they covered topics for which the assistant could offer suggestions, we provided participants with prompts, one for each of the three interactive rounds: cooking dinner, arranging weekend plans, and planning a vacation. For each of these topics, we came up with a selection of proactive apps that would be listening, for example *Recipes* and *Shopping List* (for cooking) and *Flights* and *Weather* (for making plans). (See Appendix A for complete list.)

Permission designs A major design consideration was whether permission requests would be presented visually or using audio. We opted for a combination, with the request presented on-screen (to match the modality of the suggestions) but accompanied by an audible bell. We also included this design choice as one of the discussion topics in our interview.

We came up with a design and behavior pattern for each of the permission modes (§3.2). The default permission design was a dialog box with two “buttons,” *Allow* and *Deny* (Figure 2a). Participants were instructed to say one of these words out loud to signal their preference. The same dialogue was used for the *Learning* variant, but it was shown only once or twice for each app, as a simulation of the assistant having “learned” the user’s preferences. The *Rules* variant permission request featured two additional choices: *Always allow* and

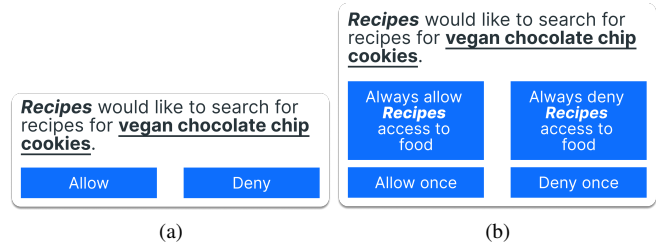


Figure 2: **Sample permission request for (a) ask-every-time and Learning designs and (b) Rules design**

Always deny (Figure 2b). These options were adjusted for each relevant app and data type (e.g., *Always allow* Calendar access to dates).

As part of our explanations, we told our participants that both the *Rules* and *Learning* designs had an extra feature: a “review mode” that allowed users to see what decisions were made automatically on their behalf and change them if necessary. Participants could invoke this mode during the simulation by asking to review their permissions. If they did so, we showed them a separate screen that contained copies of approved or denied permission requests. One of our research questions was whether participants would make use of this.

Misbehaving apps Most apps in the simulation were intended to perform correctly, only asking permission for pertinent information at relevant times. However, we also wanted to see how people would react to inappropriate permission requests. This would also serve as a basic test of the permission system’s effectiveness at preventing malicious apps. To that end, during each of the interactive sessions, participants encountered a permission request from a new, previously unseen app, which would request access to the last thing said, even though it had no relevance to the app’s actual functionality. The three misbehaving apps were *Celebrity gossip*, *Bedtime stories*, and *Smart lightbulb*. We chose them because they were plausible apps for an intelligent assistant generally, but unlikely to come up in conversations on the topics we provided to participants. To make this “attack” more random, we tried to vary when in the conversation it happened.

Interview questions After each interactive session, we interviewed the pair of participants about their experience. Our questions covered general impressions of the proactive assistant and specific feedback about the permission prompts. We also collected perceptions and preferences for the different permission modes. Finally, we asked directly about privacy with respect to the proactive assistant, including any concerns people had and controls they wished to see in a device. The complete interview guide can be found in Appendix B.

Analysis We analyzed the interviews in our study using an inductive approach to thematic analysis [20]. Two coders reviewed each interview and created a codebook with themes

identified across responses. After agreement was reached, both coders annotated passages with themes from the codebook. We did not compute interrater reliability, as it is not well-defined when the unit of analysis is an entire interview [15,64]. We also did not compute statistics, as the small scale of qualitative research does not lend itself to quantitative generalizations [68]; instead, we report the range and general prevalence of different attitudes.

3.4 Recruitment and demographics

We recruited participants for our study by advertising a “computer gig” on Craigslist in different locales in the United States. A screening survey asked for basic demographics and three free-response questions about the respondent’s use of smart home devices. When inviting people to the main study, we tried to balance different levels of experience with smart home technologies: low (limited or no usage of voice assistants), moderate (usage of smart speakers only), and high (multiple smart home devices besides smart speakers). Among those who completed the study, 65% used a smart speaker and 39% had other smart devices. We also aimed to balance our sample demographically. All procedures were IRB-approved.

Our screening survey was completed by 176 people, from whom we selected 23 pairs to participate in the study. The majority of the pairs (52%) consisted of spouses or partners, 30% were made up of family members, and the others were friends or roommates (9% each). Among the 46 participants, 57% were female; the mean age was 37; and 30%, 28%, 24%, and 18% self-identified, respectively, as White, Black, Asian, and of multiple or different ethnicities. The study session lasted 90 minutes, and participant pairs received \$60 in compensation (to be shared by the two people).

3.5 Limitations

Our work has a number of limitations, which are driven in large part by the hypothetical nature of our target devices. Wizard of Oz simulations may elicit different reactions compared with real-world deployments; the time delay in ours further reduces realism. Runtime permissions, the focus of this paper, are just one type of privacy control; future work may investigate others. Some of our assumptions about architecture as well as the *Learning* mode may currently be impractical; but this may change due to the rapid progress of machine learning and other computing fields. Also, this work’s threat model focuses on assistants and their apps and does not address the privacy threats posed by intra-household dynamics [28,38,51].

While smart displays (e.g., Echo Show) are becoming more widespread, most users currently interact with intelligent assistants through voice. Yet, a proactive assistant needs to provide suggestions ambiently, and we chose to deliver these on a screen, because this matched prototypes in literature [12,82], while audio-based ambient suggestions had not previously

been studied on their own. After deciding on this, we felt that having audio permission requests to go along with visual suggestions would be confusingly inconsistent, opting for permissions to also be requested visually (though accompanied by an audible bell). Since interaction modality can affect privacy perceptions [27], future work should investigate whether user reactions differ towards voice-based permissions.

Overall, our study required design choices that involve simplification and guesswork; nonetheless, we took care to control for and isolate privacy-relevant aspects of the system, so that our findings would be generalizable and could shed light on proactive assistants, even if the eventual products’ exact implementation details will differ.

4 Results

This section describes participants’ behavior during the interactive sessions and reports the major themes that resulted from analyzing the interview portions of our study.

General perceptions When making sense of the proactive assistant’s functionality, existing smart speakers were a baseline for feature comparison: “*It just seems like an enhanced Alexa*” (P16B). We found that our participants were, on the whole, receptive and even enthusiastic about the idea of a proactive assistant when it was first introduced to them. One of the closing questions in our interview was whether the participants would choose to adopt a proactive assistant. With only a few exceptions, our participants agreed that they would.

“*I think it’s nice that you don’t have to call out the name because it’s already picking up on the conversation*” (P16A)

Though participants perceived proactivity positively, they were aware of its privacy implications. For example, a number of participants relayed stories of existing devices listening at unexpected times, such as voice assistants interrupting a conversation to answer a question no one asked. Such accidental activations remain a regular occurrence [33,79].

4.1 Privacy perceptions

When we asked participants for their initial reactions, only a small fraction mentioned privacy, but the subsequent interviews revealed nuanced and situation-dependent viewpoints. This relative nature of privacy perceptions is consistent with other research [94] as well as the theory of contextual integrity [69], which argues that privacy expectations depend not only on data type, but also on contextual factors including the data subject, recipient, and transmission principle. In this way, our findings echo those of many other privacy studies. Despite the potential repetitiveness, we report these results to convey that context holds constant even with a new and potentially controversial technology like always-listening devices.

Privacy nihilism A very small number of people claimed that they do not care about privacy at all, repeating the common trope about having nothing to hide:

“I think we’re very average people, you know, and privacy is not an issue, at least for us.” (P4B)

“I guess I don’t have too much to hide.” (P22B)

Resignation A more common opinion, though still in the minority, was privacy resignation, a phenomenon that has been observed in other contexts as well [80]. While these people valued their information, they felt that attempts to protect it would, to a large extent, be futile because modern technology is designed to collect as much data as possible.

“In this day and age, everybody’s recording everything.” (P21B)

“We have technology everywhere, like that’s kind of beyond us at this point.” (P23A)

The other common reason for resignation was the belief in hackers’ ability to obtain almost any information:

“Anybody can hack into anything.” (P17A)

“There’s always third parties out there now. If they really want to hack in anything it’s easy—so easy—for them.” (P12B)

Worries about hackers were common even among those who did not express quite such an absolute conviction about attackers’ abilities. As evidence, participants cited recent high-profile cyberattacks that had been reported in the media. P1A, for example, felt that the government was powerless to stop these (“they can’t secure nothing”).

Privacy contradictions Even the people who claimed that they were not concerned about privacy actually demonstrated nuanced views. (This is consistent with much research on the so-called “privacy paradox” [84].) For instance, P9B described themselves, “I’m pretty much an open book. I mean, I think a lot of people worry too much about privacy.” Yet, shortly thereafter, they provided an explicit example of data types they did consider private: “If I ask [my partner] for a social security number, if I’m filling it out, you know, I may not want [the assistant] to do things like that.” P17A drew a clear distinction between two privacy-invasive behaviors, one that they did not mind and another they considered unacceptable:

“I don’t really care that they’re kind of tracking me in a way, but I don’t want someone to break into the system and find out where I am and stuff. That’s scary.” (P17A)

Consistent with the theory of contextual integrity [69], this example illustrates that while P17A finds some data flows acceptable, others would be considered norm violations.

General privacy concerns The majority of our participants articulated some privacy concerns about always-listening devices, either organically over the course of the interview, or directly, when prompted. Often, these concerns were attributed to “some people,” rather than themselves:

“I think this would be something that I feel like a lot of people would be concerned about.” (P18A)

Only a couple of interviewees expressed discomfort with the always-listening nature of the device more generally. (P1A, for example, referenced Orwell’s *Nineteen Eighty-Four* [71]). On the whole, though, always-listening did not bother people; instead, there was specific information and scenarios that they were concerned about.

Sensitive data types Consistent with popularly held notions about what is considered private information [22, 73] and research on voice assistants and their third-party apps [3], the most common data type participants worried about was financial information, such as bank accounts, credit card details, social security numbers, or account credentials.

“Anything that has to do with my banking information, anything about money.” (P16B)

“Like your address, your social security number.” (P15A)

“I’m talking to customer care and they ask me for my credit card details or my PIN.” (P20B)

Participants were also worried about the device overhearing conversations on subjects they considered sensitive, with several highlighting gossip as a specific example.

“Let’s say we’re gossiping.” (P19B)

“What if I’m talking to someone, you know? We’re planning a funeral or something? Maybe I don’t want Alva⁴ listening. And maybe that person is sharing stuff and they don’t want it listening.” (P8A)

The latter quote also demonstrates concerns about non-owners of the assistant whose voice might be captured against their will. Tensions between primary and secondary users are a common feature of smart homes [38, 51, 95].

While medical information is often considered sensitive in the United States [30, 73], only two participants brought it up in our interviews.

“I wouldn’t want the whole world to know my medical history.” (P1A)

“When it comes to financial and medical things, that should obviously be protected.” (P19A)

A few people referenced arguments or disputes as another example of a specific sensitive conversation subject.

“We got into an argument and we’re going, ‘he said, she said.’” (P7B)

“If we’re ever having, let’s say, an argument. Or we’re, you know, having a tough conversation or something.” (P4A)⁵

⁴Alva is the name we used for the intelligent assistant in our study.

⁵In this case, however, the interviewee felt that there actually could be a role for a (sufficiently smart) assistant to step in and mediate: “It would say, hey, take a break. You two should take some time apart right now.”

Other examples of sensitive conversations that participants came up with included “family matters” (P2A), relationships and cheating—“I’m having an affair with somebody” (P1A)—and business calls made while working from home.

“Now it’s work from home, or I might be just calling a colleague and talking. [...] That’s confidential.” (P20B)

While most concerns focused on specific data types, such as the ones above, one person brought up the issue of metadata leakage, pointing out that even innocuous conversations could reveal potentially sensitive details. They felt, therefore, that all data—not just “private” conversations—merited protection.

“Anything can be used. Like me making a dinner reservation for seven o’clock is not a problem, until the stalker breaks into my house and wants to find out what I’m doing at seven o’clock. So it could be information that’s not harmful. But in the wrong hands, it can become harmful.” (P19A)

Indeed, a variety of inferences can be made from voice even without considering content [52], and advertisers have sought to exploit all information available to them [66].

Data uses Some of the concerns voiced by participants focused on what would happen with their information—for example, who would get it, where it would be stored, and for how long—rather than the specifics of the data. Concretely, a number of people expressed discomfort with the possibility of their data being sold.

“If they were selling my information and then if I was wanting to plan a trip to Hawaii and then suddenly I received calls from my travel agent or something.” (P14A)

Intra-household data leakage Several participant pairs brought up the possibility that the assistant would overhear conversations and later reveal their contents, in one way or another, to other members of the household, leading the person to find out secrets others are keeping from them.

“Maybe something that you discussed—it was really really private—popped up on the screen and somebody else in the house saw it.” (P6A)

Secrets need not be a sign of malfeasance or problems in the household, but are instead benign everyday occurrences:

“Kids, they’re very nosy, so they don’t need to know everything. What if you’re planning a surprise party and they’re going to want to be, like, oh what were mom and dad talking about?” (P10A)

“Let’s say I’m throwing a surprise dinner for [partner]. [...] But then [the partner is at] home and [assistant] just starts blurting out next week’s plans, and I’m, like, did I freaking tell you to do that?” (P19A)

Impactful actions Overwhelmingly, concerns expressed by participants in our study focused on impactful action the assistant might take. These worries—that the assistant would do

something the user would disapprove of—were much more common than concerns about what would happen with data.

While different in kind, the *contexts* for these concerns were similar to the data types above. For example, the top concern was that the assistant would take actions with financial consequences, such as buying items or booking tickets.

“I want to make my own financial decisions.” (P1A)

People also worried about social consequences that might follow from the assistant performing actions without approval, for example messaging friends or creating invitations. (Communications are often a source of privacy concerns [13, 83].)

“I would always have it [ask me] only when it’s going to send something to someone else, like a person in my contacts or something else.” (P4A)

Even if the assistant’s actions affected no one but the user of the device, participants observed, they are still able to cause annoyance or inconvenience, for example through unwanted events being scheduled or alarms being set.

“If you’re having a discussion with someone and it comes up, hey, should we cancel dinner for tomorrow? [...] She might automatically do that without hearing the end result, or put random things on your calendar.” (P9B)

While the inconvenience stemming from such autonomous actions may be judged as relatively minor, participants often felt that it was these violations that permissions ought to be, or were, guarding against.

4.2 Runtime permissions effectiveness

A major goal of this study was to observe how runtime permissions would perform in a semi-realistic setting.

Concept comprehension Overall, we observed that nearly all participants understood how to use permissions right away. The majority of permission requests in our study were approved; when participants denied one, it was typically because they considered the service unnecessary, for example if the assistant offered driving directions to a familiar destination.

One area where there may have been a gap in participants’ understanding was in the role of third-party apps. As part of our overview, everyone heard that features—including the most basic ones—were implemented by apps. Nonetheless, participants never treated the apps as distinct from the assistant. It is possible that this was an artifact of our study, since we framed it as a test of the assistant in general. However, researchers have observed similar confusion with existing third-party skills [58], so the issue may be more universal.

Detecting inappropriate requests We found that our permissions system worked fairly well for preventing data capture by the “misbehaving” apps (§3.3). Participants denied a large majority of permission requests from these apps, whereas they allowed most requests from other apps. Many

also commented about the misbehaving apps, providing evidence that they were paying attention and that the observed behavior was anomalous and memorable.

“That made me very alert: why did they talk about bedtime stories right now? It’s got nothing to do with what we were talking about.” (P20A)

Some participants (less than a quarter of all cases) did allow permission requests from misbehaving apps. This was primarily due to lack of attention or some amount of habituation.

While some described the inappropriate permission requests as weird or even “spooky,” most were not concerned by them. Rather than evidence of an attempt at data capture, people saw them as in line with bugs they had experienced using current voice interfaces, for example due to speaking English with an accent. Consistent with our observation that participants did not clearly distinguish apps from the platform, those who commented on inappropriate requests attributed the mistakes to the assistant itself.

“It’s kind of like when Siri gets stuff wrong.” (P10A)
“Sometimes my accent makes me say the things or certain words with a different tone or something. And the program could misunderstand those types of things.” (P4B)

4.3 Runtime permissions perceptions

One of our main research goals was to collect first-hand feedback on the user experience of runtime permission requests.

Ask-every-time is annoying In the first session, the assistant asked for permission on every potential data access. As expected, everyone agreed that this resulted in too many permission requests, describing the experience as “annoying” and expressing a strong desire for fewer interruptions.

“That’s going to get on people’s nerves, okay?” (P3B)

Because they resulted in significantly fewer permission requests, the streamlined permission modes (*Rules* and *Learning*) were received much more positively. However, beyond that, there was not much consensus about the two modes and their distinctive properties.

Advantages of Learning Between the two permission modes, a slight majority preferred *Learning*. This group expressed trust in the automation to accurately learn their preferences and explained that they were not concerned about it making mistakes and granting inappropriate permissions.

“Well I don’t see any damage that it can do since it’s not giving out any demands or orders anywhere.” (P13B)

Weaknesses of Rules Another reason people cited for preferring *Learning* was the cognitive overhead of the four permission choices in the *Rules* variant. The extra options required more time to read and also made the decision more complicated, since users had to think about whether they

wanted to allow an app always or just once. While deliberation can help reduce the influence of heuristics and cognitive biases [16], too much may turn users away from the product.

“It creates a sense of paralysis by analysis.” (P17A)

Furthermore, nearly half of participants expressed some sort of confusion about this variant. Specifically, users were uncertain about whether “always allow” referred to the specific app being always allowed, or if it was the specific speech they uttered (for example, any app could always access the location they just mentioned).

“It kind of got me more distracted, because I’m having to stop to think about that.” (P14A)
“Is it that I don’t need to allow the music or is that allowing the music allows all of the music apps?” (P5B)

Another concern was that rules were active forever. Some assumed that was not the case, while others felt that it should not be. Research in other domains has identified users’ desire for more dynamic rules [63] as well as for automatic data deletion and other forms of longitudinal privacy management [18, 48, 60].

“Just as a regular consumer, I assume it was good for just that day and then it would probably reset again.” (P16A)
“I hesitate to do it once or because I might change next time. I’m not sure if next time I go I might change, so I debate on should I use always or should I just use it once?” (P12B)

Advantages of Rules Those who preferred the *Rules* variant expressed a desire for greater control over the assistant.

“Sounds really like therapist stuff, but I feel like I have more support with [Rules mode]. I felt like there was more hand-holding going on. I felt like I had guidance.” (P16A)

This variant was also popular among those who distrusted the assistant’s automation—or simply did not see it as beneficial—and did not want it to make decisions on their behalf, especially if they might have undesirable consequences.

“It’s like the AI would be the one controlling it. And I think, in that situation, it’s, like, why are you asking permission if you’re going to not ask for permission later?” (P23A)

Non-use of the review feature The review mode (in either condition) also received mixed feedback. Only a minority invoked it during the sessions, mostly out of curiosity. Many said afterwards that they forgot about it, but some critiqued its user experience or even the need for it.

“I find it difficult to use that feature, actually.” (P22B)
“I don’t need that. I trust [the assistant].” (P1A)

Most participants were not opposed to the idea of a review feature and many claimed they would use it, with varying frequency. The most common use case was if something suspicious happened, which is consistent with its use in existing

devices [56, 60]. Thus, the review feature’s relative unpopularity may be an artifact of our study, and it may prove to be more in demand with prolonged use of the assistant.

Similar to permissions, most saw the value of the review feature in being able to oversee the apps’ actions and the device’s understanding, rather than an audit mechanism to verify that the apps and automation were not behaving badly.

“I wanted to see if not only I could see what apps I’ve approved, but also what I asked them to do. [...] So that I wouldn’t have any duplicate actions or events.” (P4A)

4.4 Trust in permissions

We wanted to know whether the permissions helped people trust always-listening devices more.

Some see little value We found that a number of participants, especially those who were less concerned about their privacy, did not see a strong reason for permissions.

“I see [permissions] more of like a redundancy. [...] Buying it and having it in my house is almost like implicit consent as it is.” (P17B)

Others appreciate the control Nonetheless, when prompted, a little under half of participants commented that permissions enabled their trust in the assistant.

“It makes me feel like I have the control for what I am allowing and I’m not allowing. So that gives me a sense of trust. Just because I feel like I’m the one making the decision.” (P14A)

Supporters of permissions spoke about how they provided a greater degree of control, which they wanted.

“If it’s hearing everything, you know that it’s already not private, but you’re also wondering where this is going to. So that gives you a little bit more room to control it.” (P10A)

The fact that this preference was common but not universal could be a reflection of differences in the preferred level of control displayed by different people: while some people are interested in decision automation, others want only analysis automation and to make decisions themselves [72].

Many, including those who liked having the permissions, saw them as a way to control the suggestions, rather than a privacy feature.

“For me, the only time I would deny is if it was trying to help me too much. If it was something that I didn’t want to do just yet.” (P4A)

Permissions don’t address all concerns Even those who found permissions valuable did not see them as a comprehensive solution. When presented with a scenario in which they were reading a credit card number out loud near the assistant, only one person stated that the permission system on its own

would provide adequate protection; the rest explained that they would not feel comfortable relying on it alone.

“One of the main things that I think of is the app malfunctioning. What if the information did get through even despite the permission?” (P15A)

Instead, people described other protective behaviors they would engage in, such as leaving the room that had the smart speaker or unplugging the device.

“I would go to another room. I don’t trust the microphones. I’ve been told that microphones are never off.” (P16A)

Some pointed out that they were worried not only about the apps but also about the device itself compromising their privacy. This is an important reminder that the threat model our study adopted is not fully aligned with that of real users.

“That doesn’t have to do with the apps. All this has to do with Alva.” (P19A)

Retroactive auditing sufficient for those less privacy-conscious

In addition to the less-interrupting permission modes that we tested, we also surveyed our subjects about a design we refer to as “auditing,” in which an app’s permissions requests are always approved automatically, but can be reviewed at any time, using the same interface that was provided for the other conditions. When we described this design to our participants, many thought it was preferable to all of the approaches they experienced first-hand. However, we note that prior work suggests that, in practice, engagement with such a review feature may be low [60].

“I’m kind of a lazy individual. I mean, I still get to control at the end, that’s all that matters.” (P14B)

However, some had reservations about this approach, explaining that they felt that it took too much control out of their hands and that it could be abused by apps.

“I always want to know, because the companies sneak in those random ones [...] and they’re just looking for some free data for their pockets. I like to catch that.” (P21A)

4.5 Other desired privacy protections

Participants discussed a variety of additional controls they wanted to see implemented and general privacy demands.

Turning listening off The ability to turn off the device’s microphone was considered very important and helped our interviewees feel more comfortable with the device. However, studies of current smart speakers suggest that the mute button, present in all of them, is rarely used [56].

“Just having a simple on/off switch, or just saying verbally, ‘Alva, turn yourself off!’ ” (P21B)

Some wanted always-listening to only occur on demand, with the device *not* listening as its default behavior. User

studies have discovered analogous demands from users of existing smart speakers [56].

“Maybe there should be a feature where it doesn’t listen to you all the time, it’s an option when you want to start a conversation.” (P6B)

However, five different people admitted that, if the always-listening mode existed, they would forget to turn it off.

“The logical thing to do would be to turn it off, but if they’re always there, I think I would just forget that.” (P23A)

Voice identification More than half of participant pairs independently requested a voice identification feature, in which the device should only respond to recognized voices and potentially treat different people or voices differently. Similar features are available in existing voice assistants as Alexa’s Voice Profiles [11] and Google Assistant’s Voice Match [40]. Voice authentication is also offered by many banks [43].

“You can select that Alva should only detect some voices. Maybe it’s my voice. It can only do tasks after it hears my voice. And if it’s someone else’s voice, it just mutes.” (P11A)

Parental controls Many also independently suggested parental controls as an important feature. While such controls are used relatively infrequently by parents of teenagers [39], the participants in our study generally sought protection for much younger children [70].

“Does Alva have a way to block off a toddler? Because our son can talk now. If he figures this out, he can send reminders non-stop every day.” (P7B)

Parents had different views about how much access their children should have. Some felt that the device should ignore children’s voices altogether, while others simply wanted to get age-appropriate content.

“It would be me and my wife and then the kids would be excluded.” (P5A)

“If something was going to be kind of inappropriate or like 18+ type content, then a pop-up or a preference allowance or warning would come up.” (P21B)

Passwords and other prohibitions Other controls people came up with included limits on the times of day when the device would operate.

“If I could maybe set up some times when Alva should be muted, then I think that would be good. Like if it could only hear me in the morning or in the evening and not apart from that.” (P22A)

Another recurring suggestion was per-user passwords that would restrict access to data on the device.

“There could be an option of putting a password that could enable Alva to recognize yourself as the owner” (P2A)

Participants may have been inspired by a variety of current systems; most relevantly, Alexa already offers the option to set a 4-digit “voice code” which is used to confirm purchases and prevent accidental orders [9]. However, research has found that this approach does not meet everyone’s security needs, especially in higher-risk scenarios [75].

Other suggestions included “stop” words that would direct the device to stop recording, blocklists of specific words, and filtering if the conversation turns to certain topics. These approaches, while not available in present devices, appear practical based on techniques in published research [86].

“I would have a list of banned words. Financial, order, whatever. Social Security, tax, financial, money, cash.” (P1A)
“I would want some type of masking to automatically happen, if it’s possible.” (P19A)

Business practices Participants brought up other privacy expectations for always-listening platforms that focused on how the companies operated.

One requirement was a rigorous review process that all apps for the device would have to undergo, analogous to that used by smartphone app stores.

“The main security feature is I would want Alva to monitor anything that looks suspicious.” (P17B)

Today’s voice assistant platforms already require third-party skills to undergo “certification” [10]; however, this verification process may become more difficult for proactive assistants, if they allow their apps the same level of freedom and flexibility allowed by our architecture.

Multiple participants said that they wanted to be compensated in the event a data breach occurred. Some responses suggested a belief that there are existing policies or laws that provide for this. Such misunderstandings of privacy regulations are long-standing and well-documented [87].

“You get your money back and like a compensation type of thing. You know, like in the privacy article.” (P15A)

One respondent explained that they hoped developers would only collect the data they need, a strategy recognizable as data minimization, which is a requirement of regulations such as GDPR [35].

“If it’s not using it to work or to search for us, then it doesn’t need it and it shouldn’t sell it.” (P21A)

Participants also discussed other privacy factors that they found important. Among them was having a privacy policy that promised to respect their data, as well as providing security disclosures. These may be satisfied by requirements that arise from laws such as CCPA [1].

“I just want an assurance of my privacy and maybe its safety and reliability information.” (P2B)

Others brought up that their decision about adopting the device would be influenced by the manufacturer’s reputation and their business model.

“I would be concerned about the company collecting and selling data, so I would probably search about how they operate.” (P21A)

5 Discussion

This study collected people’s perceptions of proactive assistants, their privacy preferences, reactions to runtime permissions, and suggestions for other privacy controls.

5.1 Proactive assistant reactions

Many will welcome proactive assistants One basic observation from our study is that there was no wholesale rejection of proactive listening as creepy or excessive. Our participant sample *is* biased: we recruited people who were willing to be interviewed (and recorded) and many were already owners of smart speakers and other IoT devices. Still, we believe that smart speakers have paved the way for proactivity: our interviewees described it as a natural extension of present-day functionality. Even if our sample is not representative, there is evidently a market opportunity manufacturers may pursue.

Concerns center on actions and consequences While participants were open to proactive assistants, nearly all also expressed privacy concerns about them. Promisingly, the most common concerns seem plausible to overcome. With proactive assistants, people seem most worried about impactful activities: an assistant taking autonomous actions that carry financial, social, or personal consequences for the user. This result echoes recent findings about people’s hesitance towards solely automated decision making [44], and can also be seen, through the lens of contextual integrity [69], as concerns about unintended flows. On the other hand, looking up information for ambient suggestions was seen as safe. From a designer’s perspective, this appears straightforward to address by ensuring the assistant (or app) *confirms* with or *notifies* the user about any actions it is taking, such as making purchases or setting alarms. Allowing this feedback over multiple modalities may make it more convenient for the user in case, for example, they are too far away to see the display, or, conversely, the environment is too loud for the assistant to be heard.

Standard sensitive content should be excluded When it comes to the assistant simply hearing information (as opposed to taking actions), the concerns voiced by participants were similar for everyone. They centered primarily around a few sensitive data types, such as financial information or gossip,

which is consistent with findings about privacy concerns generally [22, 73] as well as documented concerns about smart homes [14] and voice assistants specifically [3]. An implication of this finding for system developers is that they can assuage users’ concerns, to a high degree, by blocking any app from hearing speech about financial, medical, or personal information. While these will vary in how easy they are to implement (detecting credit card numbers seems much more tractable compared with identifying gossip), this appears to be a promising research direction and likely an effective way of winning the trust of many potential users.

Intra-household controls needed Our interviews provided evidence for the well-known fact that people are concerned about protecting their privacy not just from apps, strangers, and other third parties, but also within the household [4, 19, 38, 51, 95]. As many participants suggested, voice identification could help: assistants could use it to limit access to interaction history, preferences, and other personal data.

5.2 Takeaways about runtime permissions

Our testing illuminated both positive and negative aspects of runtime permissions for proactive assistants.

Permissions, with architecture, help catch bad requests Permissions showed potential as a way of fending off inappropriate data access by apps, as most participants effectively identified and blocked the misbehaving apps in our study. For many, permissions also increased their trust in the device and gave them a sense of control, which they described as very important, especially for a device in such a sensitive setting.

Proposed permission designs show promise, face adoption challenges As a user experience for assistants, runtime permissions showed some promise, as participants understood them and were able to use them effectively. They were also quite successful methodologically, as an interactive and engaging way to elicit privacy attitudes and requirements. However, none of the permission modes we tested is likely to yield a user experience that would be acceptable for a real product. As predicted, no one—even those who were more privacy-conscious and wanted greater control—was happy being prompted every time an app wanted to access data. Reactions to the less-interrupting designs were much more positive, as participants appreciated their streamlined nature; still, they exhibited limitations of their own.

The **Rules design** provided the option to “always” allow or deny requests for specific combinations of apps and data types. People saw it as more usable than ask-every-time, while still leaving the user in control, which was especially welcome to those who were less trusting of the system. That sense of control may be misleading, however, as the relatively permanent nature of rules may lead people to forget about the permissions they granted. This is exacerbated by the fact that many were confused about what exactly they were allowing. Finally, a majority felt that having four options on every request

was too cognitively taxing. These pain points suggest that the *Rules* design, in its current form, would face challenges if adopted as a general-purpose permissions approach.

In contrast, the *Learning design* has the advantage of a simpler user experience. However, a sizeable minority of participants (even in our, potentially biased, sample) were unwilling to give up control over data access to a black-box algorithm. The development of an algorithm that can effectively learn people's preferences across a variety of contexts also remains an open research question, though it can build on existing work on predicting privacy preferences [3,17,26,29,32,57,93], which also show that a promising strategy may be to combine *Rules* and *Learning* approaches.

One interesting challenge for machine learning-based approaches to inferring people's preferences is the way participants used permission requests: they denied them not only when they considered the access inappropriate but also (and more commonly) when the provided service was not useful in that moment. Lacking a way to distinguish between these two reasons for denying requests, a model trained on this data may reach incorrect conclusions. This may be a fruitful avenue for future research, but for now, these challenges cast the practicality of the *Learning* approach into further doubt.

We also surveyed our subjects about “auditing,” in which permissions were approved automatically, but subject to review after the fact. For the more privacy-conscious, this was unacceptable, but the majority actually preferred it, since it did away entirely with irksome interruptions from the permission requests. Yet our findings suggest that adopting this variant would likely lead to poor privacy outcomes. People would be unlikely to make use of the review feature, as evidenced by this study and experience with other systems [60]. This would be exacerbated by the misunderstanding many users have about the distinction between the assistant itself and third-party apps for it.

5.3 Design recommendations

While better or more practical approaches may emerge in the future, what if someone were trying to build a proactive assistant today? The most effective tactic may be to combine the strategies that emerged as most promising from this study. Concretely, we would recommend that an assistant have some of the following features.

First, since so many participants were uncomfortable with the assistant making consequential decisions independently, any actions that trigger consequences beyond ambient information display would be subject to manual approval at run-time. Feedback to and from the assistant should be supported through multiple modalities (e.g., on-screen and using voice), as many pointed out that audio is better when they are not in front of the device, but that there are also times when background noise makes the screen a more effective medium.

While privacy is context-dependent, some data types are

universally seen as more sensitive and deserve special scrutiny. To account for this, the platform should, by default, identify and block access to any financial information and other known sensitive topics. Users might review a list of such topics during setup, and exceptions could be made on a case-by-case basis (e.g., for banking apps).

The majority of our participants were not comfortable with always-on continuous listening, despite acknowledging its convenience. As a result, we believe that a privacy-friendly default would be to allow users to opt in to “online” proactive listening only for specific conversations or short periods of time. The rest of the time, the assistant would operate on-demand, like current voice assistants. In this setup, since users would opt into the listening deliberately, there is a greater expectation for conversations to be analyzed and therefore a reduced need for interrupting permission requests; these could instead be automatically approved. However, they should still be auditable after the conversation has ended, since participants expressed a desire to be able to go back and review the assistant's behavior. Because most people express confusion between apps and the first-party assistant [59], during these listening sessions (as well as at other times), users should be made aware of which specific apps are accessing their conversation, as well as whether they are first- or third-party [77]. Inspired by recommendations from our participants, the device should feature voice identification (to restrict users' access to their own data) and parental controls.

While this proposed prototype may not procure perfect privacy, it would significantly enhance it compared with other approaches where apps might always be listening, and it would address many of the concerns and user experience pain points perceived as part of our probe. Future work could explore whether there are permission designs or approaches that were not part of our study, which would yield a more favorable user experience or stronger privacy guarantees.

As assistant platforms prosper and proceed in popularity, perhaps progressing into proactivity, pressure will persist to provide proper protections from their potential problems; while not perfectly practical, and plainly no panacea, permissions proffer promising performance, which plenty of people perspicuously prefer to the present predicament of pitifully poor privacy.

Acknowledgments

We would like to thank Alex Thomas for assistance with data analysis, Julia Bernd for guidance and feedback during study development, and Florian Schaub, Noel Warford, Wentao Guo, Alan Luo, and Julio Poveda for comments on draft versions of the paper. This work was supported by the NSA's Science of Security program, NSF grant CNS-1801501, Cisco, and the Center for Long-Term Cybersecurity at UC Berkeley.

References

- [1] California Consumer Privacy Act, 2018.
- [2] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, 2019.
- [3] Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2021.
- [4] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J. Lee. Tangible privacy: Towards user-centric sensor designs for bystander privacy. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), October 2020.
- [5] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Proceedings of the 22nd USENIX Security Symposium*, pages 257–272, 2013.
- [6] Hazim Almuhammedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorie Faith Cranor, and Yuvraj Agarwal. Your Location has been Shared 5,398 Times!: A Field Study on Mobile App Privacy Nudging. pages 787–796. ACM Press, 2015.
- [7] Amazon. Alexa Skills. <https://www.amazon.com/alexa-skills/b?ie=UTF8&node=13727921011>.
- [8] Amazon. Configure Permissions for Customer Information in Your Skill. <https://developer.amazon.com/en-US/docs/alexa/custom-skills/configure-permissions-for-customer-information-in-your-skill.html>.
- [9] Amazon. Require a Voice Code for Purchases with Alexa. <https://www.amazon.com/gp/help/customer/display.html?nodeId=GAA2RYUEDNT5ZSNK>.
- [10] Amazon. Skill Certification Requirements. <https://developer.amazon.com/en-US/docs/alexa/custom-skills/certification-requirements-for-custom-skills.html>.
- [11] Amazon. What Are Alexa Voice Profiles? <https://www.amazon.com/gp/help/customer/display.html?nodeId=GYCXKY2AB2QWZT2X>.
- [12] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. Investigating Proactive Search Support in Conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference, DIS '18*, pages 1295–1307. ACM, 2018.
- [13] Julio Angulo and Martin Ortlieb. “WTH..!?!” experiences, reactions, and expectations related to online privacy panic situations. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*, pages 19–38, Ottawa, July 2015. USENIX Association.
- [14] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2):59:1–59:23, July 2018.
- [15] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study. *Sociology*, 31(3):597–606, August 1997.
- [16] Paritosh Bahirat, Martijn Willemsen, Yangyang He, Qizhang Sun, and Bart Knijnenburg. Overlooking context: How do defaults and framing reduce deliberation in smart home privacy decision-making? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2021.
- [17] Natã M. Barbosa, Joon S. Park, Yaxing Yao, and Yang Wang. “What if?” Predicting Individual Users’ Smart Home Privacy Preferences and Their Changes. *Proceedings on Privacy Enhancing Technologies*, 2019(4):211–231, October 2019.
- [18] Lujo Bauer, Lorrie Faith Cranor, Saranga Komanduri, Michelle L Mazurek, Michael K Reiter, Manya Sleeper, and Blase Ur. The post anachronism: The temporal dimension of Facebook privacy. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, pages 1–12. ACM, 2013.
- [19] Julia Bernd, Ruba Abu-Salma, and Alisa Frik. Bystanders’ privacy: The perspectives of nannies on smart home surveillance. In *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*. USENIX Association, August 2020.
- [20] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, January 2006.
- [21] Barry Brown, Moira McGregor, and Donald McMillan. Searchable objects: Search in everyday conversation. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*,

- CSCW '15, pages 508–517, New York, NY, USA, 2015. Association for Computing Machinery.
- [22] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society, WPES '14*, pages 35–46, New York, NY, USA, 2014. Association for Computing Machinery.
- [23] Weicheng Cao, Chunqiu Xia, Sai Teja Peddinti, David Lie, Nina Taft, and Lisa M. Austin. A large scale study of user behavior, expectations and engagement with Android permissions. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 803–820. USENIX Association, August 2021.
- [24] Juan Pablo Carrascal, Rodrigo De Oliveira, and Mauro Cherubini. To call or to recall? That’s the research question. *ACM Transactions on Computer-Human Interaction*, 22(1), March 2015.
- [25] Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, and Hongxin Hu. Dangerous Skills Got Certified: Measuring the Trustworthiness of Amazon Alexa Platform. In *ACM Conference on Computer and Communications Security (CCS)*, 2020.
- [26] Saksham Chitkara, Nishad Gothoskar, Suhas Harish, Jason I. Hong, and Yuvraj Agarwal. Does This App Really Need My Location?: Context-Aware Privacy Management for Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):42:1–42:22, September 2017.
- [27] Eugene Cho. Hey Google, Can I Ask You Something in Private? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pages 258:1–258:9. ACM, 2019.
- [28] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. “I would have to evaluate their objections”: Privacy tensions between smart home device owners and incidental users. *Proceedings on Privacy Enhancing Technologies*, 2021(4):54–75, 2021.
- [29] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorie Faith Cranor, and Norman Sadeh. Informing the Design of a Personalized Privacy Assistant for the Internet of Things. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM.
- [30] Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule, Board on Health Sciences Policy, Board on Health Care Services, and Institute of Medicine. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press, Washington, D.C., February 2009.
- [31] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of Oz studies: Why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces, IUI '93*, pages 193–200, New York, NY, USA, 1993. Association for Computing Machinery.
- [32] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *IEEE Pervasive Computing*, 17(3):35–46, July 2018.
- [33] Daniel J. Dubois, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, and Hamed Haddadi. When Speakers Are All Ears: Characterizing Misactivations of IoT Smart Speakers. *Proceedings on Privacy Enhancing Technologies*, 2020(4):255–276, October 2020.
- [34] Yusra Elbitar, Michael Schilling, Trung Tin Nguyen, Michael Backes, and Sven Bugiel. Explanation beats context: The effect of timing & rationales on users’ runtime permission decisions. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 785–802. USENIX Association, August 2021.
- [35] European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016.
- [36] Adrienne Porter Felt, Serge Egelman, Matthew Finifter, Devdatta Akhawe, and David Wagner. How to Ask for Permission. In *HotSec*, 2012.
- [37] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android Permissions: User Attention, Comprehension, and Behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security, SOUPS '12*, pages 3:1–3:14, New York, NY, USA, 2012. ACM.
- [38] Christine Geeng and Franziska Roesner. Who’s In Control?: Interactions In Multi-User Smart Homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pages 268:1–268:13. ACM, 2019.

- [39] Arup Kumar Ghosh, Karla Badillo-Urquiola, Mary Beth Rosson, Heng Xu, John M. Carroll, and Pamela J. Wisniewski. A matter of control or safety? Examining parental use of technical monitoring apps on teens' mobile devices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Machinery, New York, NY, USA, 2018.
- [40] Google. Link your voice to your devices with Voice Match. <https://support.google.com/assistant/answer/9071681>.
- [41] Google. Android 6.0 Changes. <https://developer.android.com/about/versions/marshmallow/android-6.0-changes>, 2015.
- [42] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. Amazon vs. My brother: How users of shared smart speakers perceive and cope with privacy risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA, 2020. Association for Computing Machinery.
- [43] Rupert Jones. Voice recognition: Is it really as secure as it sounds? *The Guardian*, September 2018.
- [44] Smirity Kaushik, Yaxing Yao, Pierre Dewitte, and Yang Wang. "How I Know For Sure": People's perspectives on solely automated Decision-Making (SADM). In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 159–180. USENIX Association, August 2021.
- [45] J. F. Kelley. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '83, pages 193–196, New York, NY, USA, 1983. Association for Computing Machinery.
- [46] Patrick Gage Kelley, Sunny Consolvo, Lorrie Faith Cranor, Jaeyeon Jung, Norman Sadeh, and David Wetherall. A Conundrum of Permissions: Installing Applications on an Android Smartphone. In Jim Blyth, Sven Dietrich, and L. Jean Camp, editors, *Financial Cryptography and Data Security*, pages 68–79, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
- [47] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3393–3402. Association for Computing Machinery, New York, NY, USA, 2013.
- [48] Mohammad Taha Khan, Maria Hyun, Chris Kanich, and Blase Ur. Forgotten But Not Gone: Identifying the Need for Longitudinal Data Management in Cloud Storage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 543:1–543:12, New York, NY, USA, 2018. ACM.
- [49] Jonathan Kilgour, Jean Carletta, and Steve Renals. The Ambient Spotlight: Queryless desktop search from meeting speech. In *Proceedings of the 2010 International Workshop on Searching Spontaneous Conversational Speech*, SSCS '10, pages 49–52, New York, NY, USA, 2010. Association for Computing Machinery.
- [50] Ilker Koksall. The Sales Of Smart Speakers Skyrocketed. *Forbes*, March 2020.
- [51] Vinay Koshy, Joon Sung Sung Park, Ti-Chung Cheng, and Karrie Karahalios. "We just use what they give us": Understanding passenger user perspectives in smart homes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2021.
- [52] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference. In Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker, editors, *Privacy and Identity Management. Data for Better Living: AI and Privacy*, volume 576, pages 242–258. Springer International Publishing, Cham, 2020.
- [53] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. Skill Squatting Attacks on Amazon Alexa. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 33–47. USENIX Association, 2018.
- [54] Christoffer Lambertsson. Expectations of Privacy in Voice Interaction—A Look at Voice Controlled Bank Transactions. Technical report, 2017.
- [55] Frederic Lardinois. Google makes it easier to chat with its Assistant. *Techcrunch*, May 2022.
- [56] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):102:1–102:31, November 2018.
- [57] Bing Liu and Ian Lane. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Interspeech 2016*, pages 685–689, September 2016.

- [58] David Major, Danny Yuxing Huang, Marshini Chetty, and Nick Feamster. Alexa, who am I speaking to?: Understanding users' ability to identify third-party apps on Amazon Alexa. *ACM Transactions on Internet Technology*, 22(1), September 2021.
- [59] David J. Major, Danny Yuxing Huang, Marshini Chetty, and Nick Feamster. Alexa, Who Am I Speaking To? Understanding Users' Ability to Identify Third-Party Apps on Amazon Alexa. *arXiv:1910.14112 [cs]*, October 2019.
- [60] Nathan Malkin, Joe Deatrack, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy Attitudes of Smart Speaker Users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271, 2019.
- [61] Lydia Manikonda, Aditya Deotale, and Subbarao Kambhampati. What's Up with Privacy?: User Preferences and Privacy Concerns in Intelligent Personal Assistants. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 229–235, New York, NY, USA, 2018. ACM.
- [62] David Maulsby, Saul Greenberg, and Richard Mander. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pages 277–284, New York, NY, USA, 1993. Association for Computing Machinery.
- [63] Michelle L. Mazurek, Peter F. Klemperer, Richard Shay, Hassan Takabi, Lujo Bauer, and Lorrie Faith Cranor. Exploring reactive access control. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2085–2094, Vancouver BC Canada, May 2011. ACM.
- [64] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.
- [65] Moira McGregor and John C. Tang. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 2208–2220, New York, NY, USA, 2017. ACM.
- [66] Wei Meng, Ren Ding, Simon P. Chung, Steven Han, and Wenke Lee. The Price of Free: Privacy Leakage in Personalized Mobile In-App Ads. In *Proceedings 2016 Network and Distributed System Security Symposium*, San Diego, CA, 2016. Internet Society.
- [67] Richard Mitev, Markus Miettinen, and Ahmad-Reza Sadeghi. Alexa lied to me: Skill-based man-in-the-middle attacks on virtual assistants. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, Asia CCS '19, pages 465–478, New York, NY, USA, 2019. Association for Computing Machinery.
- [68] Kate Moran. Collecting Metrics During Qualitative Studies, June 2021.
- [69] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books, Stanford, Calif, 2009.
- [70] Marije Nouwen, Maarten Van Mechelen, and Bieke Zaman. A value sensitive design approach to parental software for young children. In *Proceedings of the 14th International Conference on Interaction Design and Children*, IDC '15, pages 363–366, New York, NY, USA, 2015. Association for Computing Machinery.
- [71] George Orwell. *Nineteen Eighty-Four*. Secker & Warburg, London, 1949.
- [72] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000.
- [73] Pew Research Center. Public Perceptions of Privacy and Security in the Post-Snowden Era. Technical report, Pew Research Center, November 2014.
- [74] Kurt Wesley Piersol and Gabriel Beddingfield. Pre-wakeword speech processing, 2020.
- [75] Alexander Ponticello, Matthias Fassl, and Katharina Krombholz. Exploring authentication for security-sensitive tasks on smart home voice assistants. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 475–492. USENIX Association, August 2021.
- [76] Laurel D. Riek. Wizard of oz studies in HRI: A systematic review and new reporting guidelines. *J. Hum.-Robot Interact.*, 1(1):119–136, July 2012.
- [77] Aafaq Sabir, Evan Lafontaine, and Anupam Das. Hey Alexa, Who Am I Talking to?: Analyzing Users' Perception and Awareness Regarding Third-party Alexa Skills. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, New Orleans LA USA, April 2022. ACM.
- [78] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. A Design Space for Effective

- Privacy Notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, Ottawa, 2015. USENIX Association.
- [79] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. Unacceptable, where is my privacy? Exploring Accidental Triggers of Smart Speakers. *arXiv:2008.00508 [cs]*, August 2020.
- [80] John S. Seberger, Marissel Llavore, Nicholas Nye Wyant, Irina Shklovski, and Sameer Patil. Empowering resignation: There’s an app for that. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2021.
- [81] Bingyu Shen, Lili Wei, Chengcheng Xiang, Yudong Wu, Mingyao Shen, Yuanyuan Zhou, and Xinxin Jin. Can systems explain permissions better? Understanding users’ misperceptions under smartphone runtime permission model. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 751–768. USENIX Association, August 2021.
- [82] Yang Shi, Yang Wang, Ye Qi, John Chen, Xiaoyao Xu, and Kwan-Liu Ma. IdeaWall: Improving creative collaboration through combinatorial visual stimuli. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW ’17*, pages 594–603, New York, NY, USA, 2017. Association for Computing Machinery.
- [83] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. "I read my Twitter the next morning and was astonished": A conversational perspective on Twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3277–3286. Association for Computing Machinery, New York, NY, USA, 2013.
- [84] Daniel J. Solove. The Myth of the Privacy Paradox. *George Washington Law Review*, 89, February 2020.
- [85] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. Investigating users’ preferences and expectations for always-listening voice assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(4), December 2019.
- [86] Welderufael B. Tesfay, Jetzabel Serna, and Kai Rannenberg. PrivacyBot: Detecting privacy sensitive information in unstructured texts. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 53–60, October 2019.
- [87] Joseph Turow, Michael Hennessy, and Nora Draper. Persistent misperceptions: Americans’ misplaced confidence in privacy policies, 2003–2015. *Journal of Broadcasting & Electronic Media*, 62(3):461–478, 2018.
- [88] Anthony Vance, Jeffrey L. Jenkins, Bonnie Brinton Anderson, Daniel K. Bjornn, and C. Brock Kirwan. Tuning Out Security Warnings: A Longitudinal Examination of Habituation Through fMRI, Eye Tracking, and Field Experiments. *MIS Quarterly*, 42(2):355–380, February 2018.
- [89] Anthony Vance, Brock Kirwan, Daniel Bjornn, Jeffrey Jenkins, and Bonnie Brinton Anderson. What do we really know about how habituation to warnings occurs over time? A longitudinal FMRI study of habituation and polymorphic warnings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2215–2227. Association for Computing Machinery, New York, NY, USA, 2017.
- [90] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. Eliciting and analysing users’ envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI ’21*, New York, NY, USA, 2021. Association for Computing Machinery.
- [91] Jing Wei, Tilman Dingler, and Vassilis Kostakos. Developing the proactive speaker prototype based on Google Home. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA ’21*, New York, NY, USA, 2021. Association for Computing Machinery.
- [92] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android Permissions Remystified: A Field Study on Contextual Integrity. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 499–514, Washington, D.C., August 2015. USENIX Association.
- [93] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 1077–1093, May 2017.
- [94] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Laura Brandimarte, and Alessandro Acquisti. Would a Privacy Fundamentalist Sell Their DNA for \$1000...If Nothing Bad Happened as a Result? The Westin Categories, Behavioral Intentions, and Consequences. In *Proceedings of the 2014 Symposium on Usable Privacy and Security*, pages 1–18. USENIX Association, 2014.

- [95] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata McDonough, and Yang Wang. Privacy perceptions and designs of bystanders in smart homes. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.
- [96] Bob Yirka. Google Nest hacker finds evidence of Google considering getting rid of 'Hey Google' hot words. *Tech Xplore*, October 2020.
- [97] Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security and Privacy Concerns with Smart Homes. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 65–80, Santa Clara, CA, 2017. USENIX Association.
- [98] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User Perceptions of Smart Home IoT Privacy. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):200:1–200:20, November 2018.
- [99] Marrian Zhou. Amazon's Alexa Guard can alert you if an Echo detects smoke alarm, breaking glass. *CNET News*, December 2018.

Appendices

A Conversation prompts

For each of the three rounds of the study (§3.3), participants were given a different prompt to guide their conversation with their partner. This section includes the specific directions provided to the participants, as well as the list of apps that was “active” for that conversation. In verbal instructions, we explained that these were suggestions, rather than a script to follow, and that participants were free to deviate from them, as long as they stayed with the main topic.

A.1 Task 1

Dinner + shopping Your task is to arrange to cook dinner with your partner. You can decide things like:

- which day you'll be cooking
- who will be doing the cooking
- what you will cook
- what recipe you will use (feel free to find one online!)
- whether you have the necessary ingredients for the recipe
- which ingredients you need to buy
- where you'll go to buy those ingredients
- when you'll do that shopping

As you work on this task, Alva's apps may try to offer helpful suggestions on its screen or out loud.

Installed apps Here are some of the apps installed on your device:

- Supermarket helper

- Recipe search
- Shopping list
- Reminders
- Maps
- Calendar
- Social network

A.2 Task 2

Booking a weekend trip Your task is to plan an outing for this weekend with your partner. As part of your conversation, you might:

- Discuss availability and other conflicting events
- Discuss budget
- Choose destination
- Look up things to do
- Choose activities
- Look up directions
- Decide on where to eat
- Talk about whom you want to invite along

Installed apps

- Maps
- Calendar
- Social network
- Travel info
- Weather
- Flights (and other tickets)
- Lodging
- Coupons

A.3 Task 3

Booking a vacation Your task is to plan a vacation together with your partner. As part of your conversation, you might:

- Choose travel dates
- Discuss budget
- Choose destination
- Look up things to do
- Choose activities
- Search for tickets
- Decide on where to stay

Installed apps

- Maps
- Calendar
- Social network
- Travel info
- Weather
- Flights (and other tickets)
- Lodging
- Coupons

B Interview guide

B.1 Round 1 (ask-every-time)

B.1.1 General impressions

- Please give us your general impressions of being an Alva device user. What did you like about it? What did you dislike?

B.1.2 Why do people deny requests?

- I noticed you denied (or didn't approve) app _'s permission request. Can you explain why?

B.1.3 General feedback about permission prompts

- What did you think of Alva's permission requests (in general)?
 - Understandability
 - * Were they clear or were they confusing?
 - * Did they provide enough information?
 - Modality
 - * Would you prefer to receive these requests in some other way?
 - * What did you think about receiving them on the device's screen? (instead of on your phone, etc.)
 - Attention
 - * Were the notifications effective at getting your attention?
 - * Do you think, in a real situation, you'd notice or interact with these requests?
 - * Would you want them to draw more attention to the notification? (e.g., louder noise) Or less?
 - Distractingness
 - * Were the requests too distracting?
 - * Do you think they should be more noticeable or less?

B.2 Round 2

B.2.1 Condition-specific UX questions

Learning

- Do you think Alva accurately learned your preferences? (Please explain.)
- Would you want your preferences learned in this way (if the learning were more accurate)?

B.2.2 General privacy questions about this *specific* condition

- Assuming you had an Alva, how willing would you be to install apps — either new ones or the ones from today
- Did you (want to) review the decisions made by the learning?
 - on it?
- Overall, how do you feel about your privacy with respect to Alva?
 - Do you feel that your privacy is adequately protected?
 - If not, why not? What scenario are you envisioning? What's missing?

B.3 Round 3 / exit interview

B.3.1 Condition-specific UX questions

Rules/heuristics

- Did you (want to) review the decisions made by the rule?
- Did you regret your decision to make it a rule? Are there choices the rule made that you would've preferred it didn't?
- Would you have wanted a more (or less) restrictive rule? “only allow locations when I said _”
- (if no rule ever used) Why didn't you make use of the “always allow/deny” option?

B.3.2 Comparing Alva 1 vs 2

- How did the experiences of Alva 1 and Alva 2 compare for you?
 - Which Alva version does *each of* you prefer? Why?
 - Did you find the differences between the two Alva versions meaningful? (Please explain.) How strong is this preference? Is it only because I'm asking? Would you only use one of them, or you prefer one but it's not that big a deal?
 - What are the pros and cons of each version?
 - Did you prefer the user experience one or the other?
 - Do you trust one or the other more?
- Would you be comfortable having a conversation that involves sensitive topics, if you knew the apps from today's session would be listening (but they'd still have to request permission before sharing any data)?

Normative and Non-Social Beliefs about Sensor Data: Implications for Collective Privacy Management

Emilee Rader
Michigan State University
emilee@msu.edu

Abstract

Sensors embedded in wearable and smart home devices collect data that can be used to infer sensitive, private details about people’s lives. Privacy norms have been proposed as a foundation upon which people might coordinate to set and enforce preferences for acceptable or unacceptable data practices. Through a qualitative study, this research explored whether normative beliefs influenced participants’ reactions to plausible but unexpected inferences that could be made from sensor data collected by everyday wearable and smart home devices. Some reactions were grounded in normative beliefs involving existing disclosure taboos, while others stigmatized the choice to limit one’s use of technologies to preserve one’s privacy. The visible nature of others’ technology use contradicts individual concern about sensor data privacy, which may lead to an incorrect assumption that privacy is not important to other people. Findings suggest that this is a barrier to collective privacy management, and that awareness interventions focused on information about the beliefs of other users may be helpful for collective action related to data privacy.

1 Introduction

Sensors in wearable and smart home devices collect intimate information about people’s bodies and activities in contexts that are usually considered to be very private. These data can be used to make new inferences about people that are difficult to anticipate and can be surprising, unsettling or harmful when used for unexpected purposes [18, 40, 41]. Privacy self-management, also called “notice and choice”, is the established framework for data sharing rights and permissions [43].

Under this framework, an organization providing a sensor-enabled device and associated service sets its terms, and potential users must make a one-time, up-front, take-it-or-leave-it decision to consent to the terms or not. But when sensor data collection is automated, always-on and invisible, it is difficult to imagine how people can be making informed decisions about their preferences [29]. The consent decisions people make before ever using a technology may not reflect their beliefs and preferences once they have experienced using it [48]. In addition, as sensors embedded in everyday wearable and household devices allow service providers to amass more and more data, new inferences may become possible that were not at the time the user initially gave their consent [26]. For these reasons, privacy self-management fails as a mechanism for people to exert meaningful control over sensor data.

Because privacy self-management is so widespread, it is difficult to imagine what alternatives might look like. However, scholars have begun to suggest that a collective privacy management model based on norms for acceptable data collection, use and sharing might be a more natural and effective way for people to set boundaries for how information about them should be used [52]. Privacy norms are often described as a contextual factor that affects whether disclosure happens in a particular situation [37]. However, norms can also be thought of as a mechanism by which groups of people coordinate about behavior that is considered appropriate or inappropriate for the situation [7].

People adhere to norms for offline privacy-related behaviors [17, 39]. But, existing norms about private information might or might not influence people’s beliefs and behaviors regarding the acceptability of sensor data collection and use. Anecdotally, it is possible to posit scenarios that support either position (that norms do or do not have an influence). For example, while people may believe that one should not physically sit outside someone else’s home for hours at a time observing their comings and goings, many people install technologies such as doorbell cameras that record data about the behavior of neighbors and passers-by. In this example, a norm against spying on one’s neighbors does not apply to adopting

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

a technology that effectively does the same thing.

The goal of this research was to investigate whether norms exist pertaining to the collection and use of sensor data. If normative beliefs play a role in determining acceptable and unacceptable sensor data practices, then it may be possible to design a method for collective data privacy management that relies on norm-based coordination among people. Sixty-five people who used activity trackers or voice assistants were interviewed about their own and others' reactions to hypothetical scenarios involving plausible but unexpected inferences made using sensor data collected by these technologies.

Participants' reactions to the scenarios demonstrated both normative beliefs and personal, non-social beliefs about the data and inferences presented in the scenarios. Normative beliefs involved existing disclosure taboos and also stigmatized the choice to limit one's use of technologies to preserve one's privacy. Personal beliefs focused on the desire to have control over data about oneself, the importance of awareness and consent, and the freedom of each individual to choose to use a technology or not according to their individual perceptions of how it could help them.

Norms arise where others' beliefs and behaviors are visible or known, and people can become aware of others' approval or disapproval. The choice to use a technology tends to be highly visible, whereas privacy-related concerns and motivations typically are not. The apparent contradiction between public behavior accepting data collection and private concern about it may lead people to an incorrect assumption that privacy is not important to others, and that engaging in privacy-preserving behavior is deviant. This may present a significant barrier to the development of collective privacy management strategies based on normative beliefs about sensor data. However, it suggests that awareness interventions focused on information about the beliefs of other users, rather than information about what sensor data are collected and shared, may be helpful for collective action related to data privacy.

2 Related Work

2.1 Social Norms

Social psychologists refer to two kinds of social norms: descriptive and injunctive. Descriptive norms are defined as “what is commonly done” [10]. They are beliefs about what others do, and arise through social comparison [21]. Injunctive norms are beliefs about “what is commonly approved/disapproved of” [10]. They are beliefs about what others believe, and are reinforced when specific feedback occurs in a given situation communicating to someone that their behavior violates the norm.

Social norms guide behaviors, but so do other types of beliefs, attitudes and values. This means that the presence of a norm cannot be determined by observing behavior alone—the same behavior might be caused by different kinds of beliefs.

For example, Bicchieri [7] makes a distinction between independent but similar behaviors among a group of people that emerge from the needs and circumstances of a given situation (e.g., it is cold outside so everybody is wearing a heavy coat), and interdependent behaviors that arise through social influence. Interdependent behaviors can be caused by social imitation (e.g., everybody is wearing bow ties because they see everyone else doing it) which would be considered a descriptive norm. Or, interdependent behaviors can be caused by beliefs about the approval or disapproval of others (e.g., one should not ask someone else about how much money they make), which would be considered an injunctive norm because of the evaluative aspect.

To find out if there is a norm influencing behavior in a given situation, one must identify social beliefs and expectations that cause the behavior in question. If there's a correlation among the behaviors of a group of individuals, like everybody wearing a bow tie, the objective would be to find out whether this behavior serves some need or function and everyone is just coincidentally doing it, or if beliefs about others' beliefs or behaviors are causing the behavior to happen. Observations of actual behavior are important for identifying patterns, but not enough to tell what caused the behavior. One way to try to identify whether a behavior is norm-based is to identify factors that might have caused the correlation, and then ask questions about hypothetical situations that may or may not have occurred, to find out what people would do in those situations [7]. This makes it possible to discover whether social beliefs and expectations are associated with the behaviors, and thereby understand whether norms are at work.

2.2 Individual vs. Collective Privacy Management

Many conceptualizations of privacy treat it as an individual right, which means that individuals are responsible for controlling their own information according to their concerns and preferences [43]. Privacy is contextual, so it is difficult for people to know what their preferences for future contexts or inferences will be, based on the context in which they are making a privacy decision [37]. People can't make informed decisions when they're unaware of the consequences or don't have the expertise to figure them out [27]. And, there are too many different entities involved in collecting data about users for people to reason about them all individually. Solove [48] argues that the existing consent framework for *privacy self-management* is not working; while asking for consent makes data collection legally legitimate, it does not provide “meaningful control”.

Collective privacy management is based on the idea that groups of people working together can coordinate to form and manage disclosure rules and boundaries [22, 46]. Much previous work on collective privacy management has focused on coordination among individuals about disclosure boundaries in social media. Multiple people may have different prefer-

ences about how a photo or other content should be seen and shared by others [6], and contextual factors like the nature of the relationship and network distance can play a role in negotiating and managing interpersonal boundaries [53]. Researchers have developed prototypes which, after individuals specify their privacy preferences, automatically merge preferences from multiple people to identify conflicts and propose or enforce boundary management solutions [4, 49, 50].

In addition to being used in research on interpersonal privacy, the phrase “collective privacy management” has also been used refer to policy and governance oriented approaches to managing data and information privacy. Sloan and Warner [46] argue that information privacy is a collective action problem, in which people have a common goal: to use technologies to meet their needs without disclosing information they don’t want to disclose. However, the consent framework of privacy self-management does not support coordination between individuals or groups and the organizations collecting and using data about them.

Traditional grassroots organizing and activism may be one way for groups to argue for policies that would allow them to have more agency when choosing how their data may be collected and used [12]. Other research has explored ways to support people in coordinating with each other on privacy decisions. The coordination in these studies took the form of seeking advice from the community on privacy decisions [9], delegating consent for disclosure to trusted others [36], and presenting information to people faced with a privacy decision about others’ privacy choices in similar situations [34].

In interpersonal privacy, disclosure rules and boundaries are often norm-based. People learn about appropriate and inappropriate disclosure behavior from others in their family or organizations they belong to, and form beliefs about what private information looks like and how it should be managed [39]. Those norms form part of the basis for collective interpersonal privacy management. It is difficult to envision what norms for data privacy look like, though, because agreeing upon conditions for the collection and use of digital data is typically treated as an invisible exchange between individuals and institutions. The goal of this study was to investigate whether norms exist pertaining to the collection and use of sensor data, and what specific normative beliefs might be present that could lay the groundwork for collective data privacy management.

2.3 Activity Trackers and Voice Assistants

Data privacy for sensor-based technologies is especially challenging because most sensors are by design invisible, embedded in everyday objects [19]. Data can be combined from multiple sensors, across points in time, and across multiple individuals or households to create inferences: new data points that cannot be directly collected from the environment by the sensors themselves, and are used to identify past patterns and

predict future behavior [30].

When people initially purchase smart home technologies, often they are more focused on how they’ll be able to use the features of the devices, and privacy concerns develop later [35]. Voice assistants, like Amazon’s Alexa or the Google Home, include always-on microphones that can feel intrusive to some users [8]. Across multiple research studies, participants voiced concerns related to being unsure about what data was actually being collected about them, and worried about audio data being shared with third parties and either used for targeted advertising without their permission or used for unknown purposes [2, 20, 32]. Many participants in these studies talked about feeling powerless and unable to control the data that was collected about them [23, 32].

In contrast to voice assistant users, users of activity trackers in previous research tended to be unconcerned about privacy, because they believed fitness data (step counts, calories burned) are not sensitive. Step counts are perceived to be anonymous, and users are more concerned about looking good to others when sharing their fitness data than privacy threats they perceive to be unlikely [5]. In one study, the lack of privacy concern was attributed to the belief that it is not possible to accurately infer personal characteristics beyond fitness-related information from activity tracker data [51]. The only exception was if the activity tracker collected location data—this was seen as a potential privacy risk by some users [55].

This study builds on existing work about privacy concerns in these sensor-based devices, by focusing on the sensor data itself rather than perceptions and use of the technologies as a whole. In contrast to research focused on privacy risks and concerns related to expected uses of activity trackers (fitness tracking) and voice assistants (receiving and executing spoken requests in limited domain areas), this study involves hypothetical uses of the sensor data for inferences that could enable functionality beyond the intended purpose of the devices.

3 Method

3.1 Approach

Sixty-five semi-structured interviews were conducted in which participants were presented with hypothetical scenarios involving data collected by sensor-enabled technologies. The first round of 30 interviews focused on activity trackers, which are wearable devices equipped with accelerometers and other sensors that record data about the wearer’s physical characteristics, like movements and heart rate. A second round of 35 interviews focused on voice assistants, such as smart speakers or integrated smartphone apps, which use microphones and speech recognition to accept questions and voice commands and respond by taking actions or providing information.

The hypothetical scenarios involved types of data that are typically collected by these technologies as part of their normal operation, so they would seem plausible to end users.

They described the data being used to make inferences that are not directly related to the typical usage scenarios for the two technologies. The scenarios were intended to prompt existing users of the technologies to imagine uses of the data they likely had not considered before, to elicit their initial reactions to these uses. Current or former users of these technologies were recruited to participate, so that participants could ground their reactions to the scenarios in their own experiences with the technologies, rather than relying on the interviewer’s implicit or explicit framing of the technologies.

The interview questions asked about each scenario were based on a framework developed by Bicchieri [7] to help with identifying types of beliefs that guide people’s choices and behaviors. For example, people may choose to act a certain way based on beliefs about what the outcome might be for them personally; beliefs about what they observe other people doing; or, they may hold beliefs about what one should do in a given situation. The difference between these is subtle but important. Consider the behavior of posting one’s current salary on an online profile. Beliefs related to whether someone will do this or not might focus on possible retaliation from one’s employer (non-social), seeing that others are/are not posting this information online (social), or anticipating that others will disapprove of posting one’s salary online (normative) [11]. An intervention to encourage more people to be transparent about how much money they make would only be successful if it were tailored towards the type of beliefs preventing the behavior from occurring in the first place. The interview questions were designed using this framework in order to understand the types beliefs that underlie reactions to unfamiliar uses of sensor data, and to inform the analysis:

- *non-social beliefs* are based on one’s own knowledge and experiences, and do not depend on others’ beliefs and/or behavior
- *social beliefs* are based on one’s expectations about how most others will behave in similar situations, and depend on observing others’ behavior (descriptive norms)
- *normative beliefs* are based on one’s beliefs about what others approve/disapprove of in similar situations (injunctive norms)

3.2 Interviews

Each interview began with background questions about the participant’s use of the technology that was the focus of the interview, activity trackers or voice assistants. Most of each interview was spent presenting six hypothetical scenarios to the participant, one at a time, and asking questions to probe for reactions to each scenario. The scenario descriptions were brief, only a few sentences long. Each scenario mentioned both a type of data the device might collect (e.g., movements and location, content of recipes read aloud to the user by the device) and something the data might be used to infer (e.g., when the user went to the bathroom, how healthy the user’s

eating habits are). The scenarios did not present a rationale or motivation for the platform to do what the scenario described, nor for why the user would want to use the technology in the given scenario, so that participants were not biased or primed to understand the technology in the scenarios as serving a particular purpose. They also did not mention sharing the information in the scenario with third parties or other people.

The six scenarios were very different from each other, because participants’ reactions were expected to vary according to their own beliefs and past usage of the technologies, and the interviews aimed to elicit a range of reactions from each participant. They were presented in the same order in each interview, and were designed to progress from more plausible inferences (closer to the intended purpose of the technology), to less plausible inferences. In pilot interviews, it was more difficult to gain participants’ trust and build rapport when the scenarios with the least plausible inferences came first. Trust and rapport are necessary when asking about potential norm violations. This seemed an acceptable tradeoff for order effects for this investigation, which does not intend to make causal claims. See Appendix C for the text of the scenarios.

After introducing the first scenario, the interviewer began probing for participants’ reactions by asking, “What are some different kinds of reactions people might have if [technology] could do this?” where [technology] was either activity trackers or voice assistants, referred to by the term the participant had used for the technology in the introductory part of the interview. The interviewer probed for specific examples and asked participants to explain terms and colloquialisms, and also used general prompts like “tell me more about that” to encourage participants to elaborate on their initial reactions. By asking about “different kinds of reactions people might have” the interviewer was encouraging the participant to consider not just their own reactions, but different ways they thought other people might react as well.

When the interviewer felt that the participant had nothing new to add about reactions to the scenario, they asked, “Do you feel like most people would think it is ok or not ok to use [technology] if it can know [information from scenario]?” and followed a similar strategy for probing for more detail. The third interview question asked about each scenario was, “How would you personally feel about using [technology] if it could know [information from scenario]?” This question was only asked if the participant had not already spoken about what they thought about the scenario. Once the participant had answered the three questions, the interviewer moved on to the next scenario.

The interview questions and follow-up prompts elicited reactions to each scenario in a neutral way, rather than framing the focus of the research as being about concern or privacy. The interviewer did not mention privacy or related ideas (e.g., surveillance, consent) unless the participant did first, which all participants did at some point during the interview. Likewise, the questions asked about the technology “knowing” the

information in the scenario instead of more precise terms like “infer”, “calculate” or “detect” in order to avoid providing clues about how a system might do what was described in the scenario. See Appendix B for the interview questions.

3.3 Participants

Participants were recruited using a subject pool composed of volunteers from the community surrounding a large university in the midwest region of the United States, and by snowball sampling on social media to obtain greater geographic diversity in the sample. Close contacts of the researchers were ineligible, as were undergraduate students and people who reported having received formal training in computer science or IT (information technology).

In the first round of 30 interviews, participants were current or former users of wearable activity tracker devices (19 participants) or smartphone apps that tracked physical activity (11 participants). Eighty percent of participants in this round were women, and 60% came from snowball sampling. Participants ranged in age from 23 to 48 ($M=33$). Their self-reported occupations included stay at home mom, administrative assistant, graduate student, personal trainer, state government worker, sales associate, writer.

The second round of 35 interviews¹ focused on current or former users of voice assistants, described to potential participants as technologies similar to “Alexa, Hey Siri, or OK Google.” Seventeen participants reported that they used Apple’s Siri; the remaining used Google’s voice assistant (13), Amazon Echo (6), Microsoft Cortana (3), and HTC Assistant (1). All of these except the Amazon Echo were apps on smartphones. About 30% of participants in the second round came from snowball sampling, and 46% of participants were women. Participants in the second round ranged in age from 20 to 72 ($M=39$), and their self-reported occupations included sports radio producer, chef, retired, small business owner, restaurant server, call center specialist, homemaker.

Recruiting for each round of data collection was conducted separately. At the conclusion of each interview, the interviewer created detailed memos describing emerging themes and similarities and differences across interviews. Recruiting continued until the majority of the reactions to the scenarios showed similar high-level themes to previous participants in that round. Overall descriptive statistics for both samples are presented in Appendix A. The interviews were conducted by telephone prior to the start of the COVID-19 pandemic, and ranged in length from 28 to 85 minutes ($M=51$). Each participant received a \$25 Amazon.com gift card by email after the interview ended. This study was approved by the Michigan State University IRB.

¹The voice assistant interview protocol initially used a scenario about inferring stress based on vocal pitch and speech patterns. However, the first several participants did not find this scenario plausible. A different scenario was used for the rest of the interviews, and five additional voice assistant interviews were conducted. The stress detector scenarios were not analyzed.

3.4 Analysis

Iterative qualitative analysis proceeded in several rounds [44]. First, the transcripts were coded for participant attributes like demographics, the type of activity tracker or voice assistant they used, etc. This round of coding also involved structural coding for which scenario was being discussed (Scenario 1-6) and which round of interview the transcript was from (activity tracker or voice assistant). This made it easier in later rounds of coding to identify which technology and scenario was the context for participant reactions.

Then, the transcripts were coded for inductive themes, focusing on statements indicating participants’ beliefs and reasoning related to whether the data collection in the scenario was acceptable or unacceptable and why. Beliefs were loosely defined as thoughts and perceptions about what is true, based on personal knowledge and experiences [44]. For example:

- *acceptable because it doesn’t seem harmful*: “And so if someone out there is tracking that about me, because I can’t see what the harm is ultimately, maybe it’s a little spooky, I don’t know, but I feel like in this day and age, it’s not even spooky anymore.” (AT10, woman, 38, S4)²
- *unacceptable because being monitored is uncomfortable*: “That’ll be kind of creepy. I don’t know if I would like that. ’Cause it’ll be almost like you were being watched, but through the microphone basically. I don’t know if that’s something that I would enjoy Siri knowing. I just don’t think that’s something that Siri needs to know about.” (VA13, man, 39, S4)

Participants typically spoke about multiple beliefs related to the same scenario, even conflicting beliefs, as they considered the aspects of the scenario that came up while they thought about it and how others might react to it. In other words, participants could and often did make statements about both acceptable and unacceptable aspects of the scenarios, and not all of the beliefs they talked about were their own. The codes evolved through coding an initial set of about 10 transcripts across both rounds of interviews, and once the codes had stabilized the initial set was re-coded.

Then, another coding pass focused on just the segments of the transcripts coded with belief codes, and additional codes were applied that differentiated whether participants were talking about their own beliefs versus their beliefs about what other people believe. This coding pass also identified whether the beliefs evident in the transcript segments had either non-social, social or normative characteristics.

In the final stage of the analysis, the belief codes were grouped into several higher-level themes. Codes were combined that focused on similar reasons and explanations for

²Participants are referred to by ID number, gender, age and the scenario they were speaking about in the transcript excerpt. ‘S4’ stands for Scenario 4. ‘AT’ before the number indicates a participant in the activity tracker round of interviews; ‘VA’ indicates a participant in the voice assistant round. The full text of all scenarios can be found in Appendix C.

why the data collection and use in the scenario would be acceptable or not acceptable. These codes differentiated between statements focusing on privacy-relevant beliefs such as awareness, consent and control and those that did not.

3.5 Limitations

Participants' reactions to the hypothetical scenarios, and their beliefs about how others would react, should not be interpreted as accurate predictions about how they or others would behave if the scenarios were real. Privacy choices are context-dependent, and platforms and technologies often do not provide the options people would need to make choices according to their privacy beliefs and preferences. However, beliefs about privacy are important in their own right, because they are another factor that guides and constrains behavior. The goal of eliciting participants' reactions was to better understand normative influences on beliefs about appropriate versus inappropriate sensor data collection and use, to identify new opportunities for design and policy interventions that might help people better manage the privacy of their data.

The six scenarios used in this study were designed to seem plausible to participants and also to have potential privacy implications. The technologies involved, activity trackers and voice assistants, are both discretionary use technologies. This means that unlike smartphones or cars, these technologies are not necessary to support basic needs and activities. There may be beliefs and reactions related to non-discretionary technologies, or other uses of data from activity trackers and voice assistants not present in the scenarios, that were not elicited in this study due to the nature of the scenarios. In addition, if the scenarios had been presented in a different order, the specifics of participants' reactions may have varied. However, scenario order should not affect underlying beliefs.

Finally, this research used an opt-in convenience sample consisting of mostly white, highly educated people in the United States. The sample size, at 65 participants, is larger than many qualitative studies [31]. However, these findings should not be generalized to a more diverse population without being validated in a representative sample.

4 Findings

4.1 Norms about Private Information

Normative beliefs were present in many participants' reactions to multiple hypothetical scenarios. These beliefs focused on data collection and use about information and behaviors that participants said should be private or nobody else's business. Overall, 53 of the 65 participants (82% overall; 90% AT, 74% VA) across both rounds of the study had a reaction to at least one scenario that involved normative beliefs.

References to normative beliefs demonstrated an awareness of what others believe, like the following reactions from two participants to Scenario 3, about an activity tracker that could

count how many times a person had used the bathroom. Here, AT17 (woman, 24, S3) described her expectation that nearly all other people would disapprove of the data collection and use in the scenario: "I think ninety-nine percent of people would say absolutely not. For no reason." Reactions involving normative beliefs also often had an evaluative component, like this belief described by AT26 (man, 28, S3): "Going to the bathroom's a personal thing, so it might just be a bit of a taboo subject."

In contrast, personal beliefs were typically spoken about in first person, as the participant's own belief rather than something everyone believes, e.g., "It just seems a little creepy to me, I don't know why, the phone knowing how often you oversleep" (VA16, woman, 56, S1). There were also instances where participants said they were unsure about what others would think, like the following from AT12 (woman, 39, S6): "I don't know. I would think that most people wouldn't care but I can also see why it would bother some people, but I guess I don't know about that." Statements like this were not considered to be examples of normative beliefs.

The most common reactions involving normative beliefs were about the hypothetical scenarios focused on bodily functions, like bathroom behavior and sleeping (24 participants, S2 and S3), about data collection in the home (24 participants, S4), and about inferring information about children (23 participants, S5). Forty-eight out of 65 participants (74% overall; 87% AT, 63% VA) described normative beliefs related to the use of the information in at least one of these three scenarios.

Many of the participants' reactions focused on how people in general feel that information about bathroom behavior is "personal" or "intimate" and is something one does not talk about with other people. Some spoke about how they felt like collecting this information would violate a taboo or be invasive of private space. For example, AT03 (woman, 32, S3) said, "People feel very personal about that [going to the bathroom], I don't think people would want anyone knowing that business." Participants had very little doubt or hesitation when they spoke about what others' reactions would be. They didn't equivocate—they were certain others would not like this. AT10 (woman, 31, S3) described it this way: "Oh, I think it would be outrageous. People would be outraged. Again, that's something that's very intimate, very personal." There was also an expectation that people would be angry if they found out this was being tracked without their knowledge. For example, "I would think people would just be, maybe, upset or angry that there would be information being kept on how many times you're going to the bathroom..." (AT21, woman, 40, S3).

Scenario 4 in both rounds of interviews involved the technology collecting data and making inferences about some aspect of the user's home environment. In the activity tracker interviews, the scenario involved the device making a map of the inside of the user's home while they wore it, and in the voice assistant interviews it involved doing voice detection

and counting the number of guests in the user's home. Participants' reactions to these scenarios centered on the idea that nobody would approve of this, because things that happen in one's home should be private. These participants talked about how if these inferences were being made, to most people it would feel like they were being "spied on" (VA02, man, 71, S4). Participant AT15 (woman, 36, S4) talked about how if her activity tracker did this, it would feel like being monitored—if the GPS and accelerometer data collected by the activity tracker were used for mapping rather than step counting, it would violate a norm about the home being private space: "I mean, if people wanted to know I could tell them, but personally people I don't think like to be monitored in their homes."

One scenario that was the same across both interviews involved data collected by the technology being used to infer whether the user had young children or not. Most of the reactions to this scenario invoked normative beliefs concerning protecting children from harm in general, and information about children more specifically. For example:

"Oh man, I think that having young kids at home is such a huge personal line for people, that they... that would just probably be considered a huge, huge overreach, very intrusive, and posing a lot of security and personal safety issues." (AT16, woman, 29, S5)

Participants spoke with great confidence about this, even the participants who had no children themselves. For example, VA24, who did not have children, had this to say:

"I think parents are bothered by everything involving people knowing things about their children they don't offer." (VA24, man, 27, S5)

A smaller number of participants (13 overall; 9 AT, 4 VA) talked about normative beliefs in response to other scenarios, particularly where it related to being healthy and hard-working as something people are supposed to do in order to be considered a good person. Most of these comments focused on the discomfort that comes with being evaluated negatively by others, and an expectation that the information in the scenario is something that people are often judged on. For example, in the following two examples a voice assistant participant and an activity tracker participant both spoke about beliefs about how people are supposed to behave in order to appear healthy:

"Because especially for a woman, everybody thinks you're too fat or you're too thin. You're never perfect, and that's... If it's going to automatically evaluate you based on what you're cooking... Can we have one more thing not judging us?" (VA22, woman, 29, S3)

"I think the majority of people would be afraid of being judged based on how many steps they do, or oversleeping an alarm... And we all accept that there's this basis of health that we're all supposed to maintain. There's this line that we all kind of say,

ok, this is healthy living. Were you doing it or not? If we're not, we always feel guilty, and we always feel judged." (AT05, woman, 34, S1)

In the above excerpt, participant VA22 was reacting to Scenario 3 in the voice assistant round of interviews, which was about how a device with access to the user's recipes could read them aloud and assist them while they were cooking, but also make inferences about how healthy the user is based on characteristics of the meals they prepare. Her statement illustrates normative beliefs about women's physical appearance as being related to her reaction to the scenario. Participant AT05 was reacting to Scenario 1 which was about an activity tracker that is worn to bed and counts how many times the user has overslept, and she felt that information could be used to categorize someone as lazy. These examples both illustrate very clearly the strong normative beliefs about how others approve or disapprove of people based on these characteristics.

Privacy theory considers norms to be part of the contextual factors that are important for people choosing whether or not to disclose private information [37, 39]. The findings in this section show that normative beliefs about the use of information about certain behaviors and contexts were part of participants' reactions to the scenarios. Because these norms (intimate behavior, home as private space, protecting children) are not specific to the digital context, it may seem obvious that normative beliefs about private information would apply to situations where technology is the observer of the information, not a person. However, it is also reasonable to hypothesize that people might feel like it is acceptable for their wearable devices or voice assistants to collect this information if it were not visible to other people or stayed on the device, or if the data were anonymous. The scenarios said nothing about whether the inferences would be shared or de-identified, so it is somewhat surprising that normative beliefs were present.

4.2 Norms about Privacy-Preserving Behaviors

In addition to norms about private information, there is also evidence in the data that norms exist regarding privacy-preserving behaviors, such as limiting one's use of technologies to restrict or prevent data collection about oneself. However, this evidence came in the form of normative beliefs that stigmatize concerns about data collection, and behaviors such as using less modern and sophisticated technologies (e.g., a flip phone; AT20, VA07, VA09, VA32) due to privacy concern.

A stigma is a strong sense of disapproval [38]. Stigmas often come about as punishment from a group for violating norms or deviating from accepted practices. To understand whether a stigma against privacy-preserving behaviors exists, first it is necessary to understand what people believe about normal, accepted practices related to the collection and use of digital data about themselves.

4.2.1 Data Collection is an Unavoidable Fact of Life

Thirty-three participants (51% overall; 43% AT, 57% VA) in this study believed, and also thought that other people believed, that digital data collection is an unavoidable fact of life. This is similar to the phenomenon of *digital resignation* described by Draper and Turow [14] and Seberger et al. [45]. These participants spoke about how it's not actually possible to choose not to have data collected about you—choosing to use technology is choosing to allow data collection. Participants said things like “This is the way the world is” (VA08); “I don't even know if I've agreed for them to pull out my information” (AT05); “by using the internet you're somewhat passively agreeing to be tracked” (VA02); “that's something that I feel is probably out of my control” (AT02). These participants did not seem happy about this, but rather unhappy and resigned.

VA05 (genderqueer, 24, S1) talked about it like a physical, physiological connection to their smartphone: “I mean, we're already so connected to our phones and now we have them monitoring our sleep and wake cycles like, plug me into my phone, we're the same being now.” While this seemed uncomfortable for participants, there was also a sense of futility that made it difficult for them to rationalize objecting to it. As VA07 (man, 34, D5) said, “At some point it just becomes Google knows everything, and I have to deal with that if I... Either I use Google or I don't but they're gonna find out everything if I do.”

Comments like these illustrate participants' beliefs about the data that has already been collected about them, and their reasoning about what data the technologies they use are capable of collecting about them. It was something they felt would be pointless to get upset over or do anything about, because it's already happened and is currently happening. For example, VA22 (woman, 39, S1) summed this up well: “I think people who use [voice assistants] are probably okay with it, 'cause they're already doing it. They're already doing things that are collected.”

In addition, 22 participants (34% overall; 27% AT, 40% VA) talked about how they believed that using these systems indicates one must have consented, and that consent means people must be aware of the data collection and use practices (they “knew what they were getting into,” AT15; or “knew about it going in,” AT26). This was despite the fact that the participants themselves admitted not reading terms of use and privacy policies. These participants talked about data collection as an inevitable part of using technology, and used language that had connotations of defeat (“give up [information],” AT02), coercion (“forced to go along,” VA09), and surrender/loss (“sacrifice,” AT19) to describe it. They talked about unwanted data collection as common knowledge—something everyone knows is part of using technology and cannot be avoided (“you accept certain types of information being tracked,” AT10). They rationalized uses of the data for

purposes that were separate and unrelated to providing the service that was their reason for using the technology. But, at the same time, they said it was something that most people are not concerned about because, after all, they chose to use the technology (“Well you either buy the iPhone or you don't buy the iPhone,” VA09).

These comments voice a belief that if a person chooses to use a technology or service, they are implicitly agreeing to everything that it does. For example, VA17 (man, 28, S4) said, “But in this theoretical scenario, I'm sure that probably the user has accepted via the application, the [voice assistant] or whatever, to do this sort of thing.” This belief, consistent with the notice and choice framework, places the responsibility squarely on the user to know everything the technology is collecting and using. Participant VA09 described this well:

“If you wanna use technology, I think that you have to accept the fact that you're gonna have data collected on you that you might not want to be collected on you.” (VA09, man, 23, S3)

Two things are important about this for understanding whether a stigma against privacy-preserving behaviors exists. First, participants expressed personal beliefs that data collection is commonplace and inevitable, and they use these technologies anyway. And second, they believe that other people believe this as well. In other words, participants talked about social beliefs that most people accept and are fine with sensor data collection, and have consented to it. They know that choice or consent is required to use these technologies, and believe that the choosing to use them makes the individual responsible for what comes after.

4.2.2 Objecting to Data Collection Sounds “Crazy”

Twenty-three participants (35% overall; 17% AT, 51% VA) commented that people who are concerned enough about privacy that they believe technology is harmful, feel surveilled all the time, or are focused on other harms due to lack of privacy are crazy and/or paranoid. These comments arose when participants were asked about how they thought others would react to one of the scenarios, and frequently followed immediately after a statement about the participant's own desire to protect some aspect of the information about themselves in which they distanced themselves from that desire.

Paranoia is a delusional state in which a person is excessively suspicious about being targeted for harm by others without evidence that this is happening [42]. Previous research has also found that people perceive others who might use encryption tools as ‘paranoid’ [15, 54]. By associating this state with people who want to preserve their privacy, these participants indicated that they believe being concerned about privacy is at some level irrational and deviant. For example, AT25 described herself as “a little paranoid” because she “[doesn't] think that people need to know exactly what I'm doing every minute of every day.” VA05 said that not

wanting Google to “know all these things about me” sounds “really paranoid.” VA22 said, “I don’t think that you can do anything electronically without it possibly coming back to you at some point, other people finding out about it.” But, then she distanced herself from that belief by subsequently saying, “I don’t want to sound like a paranoid person.” VA33 talked about turning off location services on his mobile device, but then also said about himself, “I’m not super paranoid”—twice. This indicates that he believes that turning off location services could be viewed by others as paranoid, and he wanted to make sure that the interviewer knew he wasn’t one of those paranoid people.

Participants also described that an unfounded, unreasonable anticipation of harmful outcomes is something that paranoid people do. For example, when asked about how she thought people would react to the scenario of an activity tracker collecting data about the inside of one’s home, AT08 (woman, 42, S4) said, “I guess it depends on how paranoid you are and [the crime rate] where you live.” VA35 (man, 51, S1) described worrying about “somebody finding out that they’ve hit the alarm so many times” as paranoid. And, VA24 (man, 27, S5) said, “Apparently, parents as a demographic seem like a paranoid group of people to me” after thinking about how parents would view a system that could automatically infer whether or not a person has children, and the possible harmful uses of that information. This indicates participants believed that paranoia is related to thinking about the likelihood and severity of privacy-related harms, and that paranoid people believe negative outcomes are more likely than is reasonable.

Believing in “conspiracy theories” was also often discussed as something that people who take steps to preserve their privacy do. Participants described that people with these beliefs feel like the government and companies are watching them and scrutinizing their activities, and that this feeling is extreme and unreasonable, even crazy (e.g., “crazy conspiracy nuts,” VA07). So whereas the non-conspiracist beliefs held by the participants conveyed understanding that the data is being collected, it was considered to be a conspiracy theory to believe that the government and/or companies are paying attention and using that information for surveillance. AT28 (woman, 24, S1) described “the conspiracy theories people” as “the people who refuse to own smart phones because they believe the government is tracking their every move and that if you have a smart phone, you’re signing away your right to all privacy ever.” VA04 (woman, 32, S2) talked about how people should be “more suspicious” of data collection, “because there are people and programs that do want that kind of information, maybe the government.” But then she immediately distanced herself from those beliefs by saying “I’m not a conspiracy theorist,” like the people who “are rebelling against technology” because they are suspicious of it. And, VA26 (woman, 32, S1) gave an example of a coworker who doesn’t want to share fitness tracker step counts with anyone, and referred to beliefs like that as, “a little more conspiracy

theorist.”

The findings in this section show that, in addition to norms about not disclosing some types of information and social beliefs about how everyone uses technologies that collect data about users, participants also held normative beliefs related to what they saw as deviant privacy-preserving behavior. They described that the common, accepted practice they engage in and see everyone around them also engaging in is to use technologies that everybody knows are collecting data about them. And, they talked in the abstract about how once a person chooses to use a particular technology, they’ve agreed to whatever data collection and use will take place. They also labeled people who object to this as “crazy” or “paranoid.” This indicates that a norm was evident in their reactions to the scenarios, supporting acceptance of data collection as an unpleasant consequence of using technology, and labeling those who visibly object as deviant.

4.3 Non-Social Beliefs about Control over Data

In addition to the normative beliefs already described, participants also expressed private, non-social beliefs that indicated they do personally care about privacy, and that being able to keep some information private is important to them. These comments from participants emphasized the idea that they want to be aware of any changes to how the technologies and services they use are handling their data. They don’t want the technologies to start doing something different with the data behind the scenes, like using it for some of the things described in the scenarios, without letting them know about it. Overall, 52 participants (80% overall; 73% AT, 91% VA) made statements like this in response to at least one scenario. For example,

“But if that’s a possibility we do need to be made aware of that, it can’t just start happening.” (VA04, woman, 32, S3)

“I think it’s important because I guess I want to know what information is being shared and being gathered. Even if it’s just the totality of what is being tracked.” (AT16, woman, 29, S4)

As part of speaking about a desire for control over how data about them is collected and used, participants talked about how they would not like it if the technology in the scenario were to start doing something they did not expect with the data. The participants’ expectations were based on what they used the system for and what they believed it was doing. As VA18 said (woman, 30, S4), “it could be a little bit more off-putting that it could be just collecting more information about how many other people are around.” In essence, these participants were saying if they’re not aware of it and can’t anticipate that it would need to be collecting that data, then it would not be acceptable to them. For example,

“If they have that just hidden in there, like what Facebook does with a whole bunch of stuff, then

no, I don't think it would be okay, and I think most people would be opposed it, if there was something that they were just sneaking in there.” (VA23, man, 64, S4)

Twenty-seven participants (42% overall; 40% AT, 43% VA) emphasized that it was important to them personally to feel like they have a choice about opting in to any functionality that involves doing something they perceive to be new with the data that is collected by the technology in the scenario. There was an expectation articulated by these participants that if the technology wanted to do what the scenario described, then it would need to seek the user's permission first. VA29 (man, 45, S3) said, “Well, I imagine that it would be my choice to turn on [voice assistant] for this particular purpose, not that it would randomly come on.” And, AT15 (woman, 36, S6) said, “So, that consent would at least, if somebody wanted to use the information for something, that they need to be very clear what they want to use it for, how it's gonna be used.”

Thirty-nine participants (60% overall; 50% AT, 69% VA) made comments focused on the idea that participants want to have some control over aspects of the data collection and use described in the scenarios. A lot of these comments had to do with being able to keep the data on the device and not send it elsewhere or make it visible to others. They wanted to be able to create boundaries such that the data would only be used for the purpose the participant wants to use the technology for. To describe the types of data collection and inferences participants wanted to prevent, they used words indicating a boundary being crossed or violated, like “overreach” (AT16), “intrusion” (VA03, VA10), and “invasive” (AT21, VA10, VA17). VA07 talked about wanting to make sure that the voice assistant was not able to “hold onto that information any longer than it needs to”, and AT18 (woman, 25, S6) talked about how the scenario would be more acceptable to her if she was able to turn off parts of that functionality: “I would be very confident if that's something I could be in charge of.”

Participants' private, non-social beliefs about privacy are an interesting contrast with the normative beliefs, described in the previous section, about the acceptance of data collection and inferences as being just an inevitable (if unpleasant) part of using digital technology. Participants' private beliefs show that they do value the ability to have control over data about themselves, and are unhappy that they cannot.

4.4 Non-Social Beliefs about Usefulness

Participants' first thoughts immediately after hearing each scenario were nearly always focused on how they personally might use the functionality described in the scenario, or how it could help other people—as long as the scenario did not violate an existing norm. Overall, 64 participants (98% overall; 97% AT, 100% VA) made a statement about how important the usefulness/helpfulness of the scenario was for determining whether the data collection and inferences described were

acceptable or not. This echoes the findings of research such as Dinev and Hart [13] and the recent literature review by Gerber et al. [16] about tradeoffs between the potential benefits of disclosing information and foreseeable harms.

4.4.1 Usefulness as Necessary Condition

Thirty-three participants (51% overall; 50% AT, 49% VA) believed that having access to the information in one of the scenarios over time would help them identify a pattern in their lives and behavior or in the world around them. Knowing about the pattern would then allow them to make better decisions, to change their behavior, or it would allow the system to make predictions or suggestions that would help them with their specific situation (e.g., changing the alarm time if you overslept a lot, making food substitution suggestions if you were eating too much salt/sugar, etc.). Participant VA09 (man, 23, S3) described his idea about how the scenario could help him: “you could have [voice assistant] suggest certain changes to your diet that she's been tracking for however long and you can be like, wow, I haven't eaten a fruit in two weeks, I should add an apple in or something.” Similarly, AT15 felt that a greater awareness of oversleeping would be beneficial:

“I guess that would at least give me a heads up like, ‘Okay, maybe I need to do something different,’ or, ‘What can I do different so that I don't oversleep in the future?’ So, I think it would be a positive thing.” (AT15, woman, 36, S1)

In contrast, if the participant couldn't imagine a way that the information would be useful, then they felt the scenario would not be ok with other people, and the participant would not like it either. Twenty-four participants (37% overall; 17% AT, 54% VA) talked about how a particular scenario would not be useful because they thought it was not possible for the technology to make accurate inferences of the kind described in the scenario. A majority of these comments came from voice assistant participants who did not believe that microphones in one's home or smartphone could be used to accurately detect potential crimes being committed. For example, VA04 (woman, 32, S6) said, “maybe I'm having an argument with my boyfriend and it thinks, oh, there's domestic violence here. But really we're just having an argument. I think the data might be a little corrupted or just not accurate.”

4.4.2 Useful or Not? It Depends...

Forty-four participants (68% overall; 67% AT, 69% VA) said that whether other people would find the functionality in the scenario useful would vary based on their beliefs, desires, characteristics or circumstances. These participants had difficulty even speculating about others' reactions to particular scenarios (that didn't violate norms or taboos) without knowing more about the other person's personality, preferences or life circumstances. This was most common in relation to

Scenario 6, which in the activity tracker interviews was about inferring one’s carbon footprint from movement data, and in the voice assistant interviews was about inferring how safe one’s neighborhood is from ambient sounds. In the carbon footprint scenario, participants talked about how others’ reactions would depend on their beliefs about climate change, and in the crime monitoring scenario it would depend on how safe or unsafe one’s neighborhood is. For example:

“I think it depends on the individual person. If there’s somebody who wants to reduce their carbon footprint, if they’re looking to kind of get an idea, like a snapshot, of what their activities that impact on the environment around them.” (AT21, woman, 40, S6)

“So I guess if you’re in a safe neighborhood, you’d probably say, great. It’s giving my neighborhood a positive ranking. But if you’re in one of the bad neighborhoods, you probably wouldn’t like it.” (VA23, man, 64, S6)

Two-thirds of participants (40, 62% overall; 73% AT, 51% VA) said that whether a given scenario would be useful or not would depend on additional information about the situation or context of use that were not provided as part of the scenario. The scenarios described sensor data being used by the technology to make inferences, but didn’t provide much background or motivation for those inferences to be made or how a person would use the inferences in their lives. As such, the scenarios didn’t contain the information participants felt like they needed to understand how the inferences could help someone, and this meant they could not conclusively say what their own or others’ reactions would be. So when asked, participants talked about the kinds of things they believed would affect people’s assessment of the scenario. These comments often focused on characteristics of possible harms in the scenarios that people would prefer to avoid, or whether or not the information would be shared. For example, 15 participants spoke about how people would react badly if the information in the scenario were shared with others when the user didn’t want it to be, and another 15 participants talked about harms in the form of data breaches, higher insurance rates, or loss of physical safety due to the information being known.

Finally, over one third of participants (37% overall; 27% AT, 46% VA) said they didn’t know or couldn’t say what other people would think about at least one scenario. This was explained as just not having any idea (16 participants, e.g., “I don’t know. I don’t know what other people think,” VA03), or not being able to say whether more people would be ok with it or not ok with it (9 participants, e.g., “I think it’d be pretty mixed... So 50-50 really,” VA05).

Speculation about whether a scenario would be useful or not was a universal reaction to the scenarios, and an important perspective for participants’ own evaluations of whether the scenario would be acceptable to them or not. In addition, participants believed that other people would also find use-

fulness to be an important factor, so much so that for most participants, more details about the situation and context were required to make a reasonable guess about others’ reactions (again, for scenarios that did not violate existing norms). This indicates that there is no collective set of social or normative beliefs about usefulness related to whether or not one should or should not use a technology that collects a certain kind of data. Rather, usefulness is left up to the individual to determine for themselves.

5 Discussion

Norms are a form of collective action, in that they represent the convergence of beliefs among a group of people regarding behaviors that are acceptable and unacceptable. Notice and choice (privacy self-management) is the opposite of collective action—it makes the individual solely responsible for understanding the data practices and consequences of using a technology before they’ve even tried it, and once they’ve consented, makes it their fault if something happens that they don’t like [48]. A collective approach to data privacy management would provide a framework for coordination among technology users so that they can take action as a group to set rules and policies for the data collection and use practices of organizations and platforms [12].

This research investigated whether normative beliefs play a role in people’s reactions to plausible but unexpected inferences based on sensor data from common wearable and smart home devices. If norms do influence whether a particular inference is judged to be acceptable or unacceptable, then it is possible that collective privacy management strategies could be designed based on that foundation.

Normative beliefs were evident in participants’ reactions to the hypothetical scenarios presented to them in this study. Common norms about disclosure of intimate information and protecting children were part of participants’ reasoning for deciding that some scenarios would be unacceptable to them, and to most people. They were also uncomfortable with the idea that their voice assistants and activity trackers could use data collected as part of the technologies’ normal operation to generate new inferences without informing them about what the inferences were and how they would be used. The existence of normative beliefs about unacceptable uses of sensor data is encouraging for the prospect of collective data privacy management. However, the findings of this study also identified three significant barriers that stand in the way of governance approaches or group collective action in support of better sensor data privacy solutions.

5.1 Barriers to Collective Data Privacy Management

The first barrier arises due to non-social beliefs about usefulness, and individual choice. The only universal rubric for deciding whether a scenario would be acceptable or unacceptable was how useful the data and inferences in the scenarios

might be. Participants initially considered each scenario from this perspective, and believed it would be of the utmost importance to other people as well. They also believed that it is up to each person to decide whether to use a technology or not according to their own individual circumstances, and everyone who uses technology accepts data collection as part of this and knows what they are getting into. This echoes the logic of privacy self-management, and supports the interpretation of others' continued use of potentially invasive technologies as an endorsement of the data practices those technologies employ. It is a highly individualistic approach that does not provide much common ground across people for collective data privacy management.

The second barrier is a result of social beliefs about technology use, and the inevitability of data collection. Participants believed data collection is an unavoidable fact of life if one chooses to use technology. The choice to use a technology tends to produce visible results, but privacy concern tends to have less visible outcomes. Knowing that others are using these technologies, participants assumed that most people approve of the company or service provider's data collection and use practices, because if they didn't they wouldn't consent and would not be using it. This makes it seem like nobody else cares about privacy, and perpetuates the belief that others must be comfortable with the status quo. This is a barrier to collective privacy management because it creates a descriptive norm supporting the use of potentially invasive technologies, no matter what their data practices are.

The third barrier stems from normative beliefs disapproving of privacy-preserving behaviors. Taking steps to limit data collected about oneself was viewed by participants as deviant, and individuals who do so were labeled as crazy or paranoid. During the interviews, some participants were even concerned themselves about being labeled in this way due to their speculation about possible harms from loss of privacy. But, participants' own non-social beliefs about the importance of controlling data and inferences about them contradicted this norm. In other words, privately, they valued privacy, but publicly they saw everyone not valuing it and negatively judging those who take steps to protect it. This contradiction is strikingly similar to a phenomenon called *pluralistic ignorance* [33], which occurs when people engage in behaviors they privately do not believe in or approve of, but they do it anyway because they believe that everyone else approves of it and they don't want to appear deviant.

Under conditions of pluralistic ignorance, normative beliefs about others' behaviors related to data collection and use conflict with private discomfort about the status quo. And in fact, stigmatization of people who violate the norm is a common component of pluralistic ignorance situations, and is especially difficult to combat when trying to change a prevailing norm [7]. Since there is little visible evidence that others value privacy and disapprove of privacy-violating data collection, people feel isolated in their private beliefs and are unlikely

to speak up or take action. Pluralistic ignorance makes it extremely risky for individuals to speak up and advocate for better data privacy options and solutions. This would make meaningful reform of the existing data privacy governance structure (notice and choice) quite difficult.

5.2 Towards Collective Data Privacy Management

Effective governance of data collection and use practices based on collective data privacy management seems unlikely, given the barriers described above. Non-social beliefs echo the logic of privacy self-management, and both social and normative beliefs exist that are essentially anti-privacy. However, participants still wanted control over their data and disapproved of some types of data collection and use. The normative beliefs supporting privacy identified in this study all apply to any type of information, not just sensor data and inferences. In other words, they were unrelated to externalities created by massive datasets and machine learning. People's concerns about the lack of control over the data collected about them are generally invisible to others, making it nearly impossible for new norms related to sensor data and inferences to form. For example, imagine a norm similar to the existing norm about protecting children, but instead disapproving of providing data to a platform that could be used to harm someone else. For such a norm to form, information about others' beliefs about this would need to be more widely available.

Most approaches to helping people gain more control over their data focus on ways to make platforms' data practices more transparent to end users. But, awareness interventions focused on information about the beliefs of other users and their privacy choices, rather than information about what sensor data are collected and shared, may be helpful for collective action related to data privacy. People who use sensor-based technologies need to know they are not alone in their privacy concerns. Even small changes to the current notice and choice framework may create an opportunity to weaken the perception that others do not value privacy. For example, in April 2021, Apple provided a new feature in iOS 14.5 called App Tracking Transparency. This feature allows iPhone users to opt out of app data tracking. According to tech news sources, 50-60% of iPhone owners have chosen to opt out as of February 2022 [1, 24]. However, platforms do not routinely disclose this type of information to end users.

Ultimately, privacy itself seems to be at odds with collective action. Behaviors like not disclosing information or opting out of using certain technologies are less visible than disclosing or opting in. In addition, choice—individual refusal—is the only option people believe they, and others, have for exercising control. But often, choosing not to allow data collection isn't really an option at all. Without more visible evidence of others' privacy-preserving beliefs, choices and behaviors, collective privacy management is unlikely to succeed.

Acknowledgments

The activity tracker interviews and pilots for the voice assistant interviews were conducted by Janine Slaker. In addition, the BITLab research group at Michigan State University provided feedback on earlier stages of this research. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1524296.

References

- [1] How Apple’s privacy push cost Meta \$10bn. *The Economist*, February 3 2022. <https://www.economist.com/the-economist-explains/2022/02/03/how-apples-privacy-push-cost-meta-10bn>.
- [2] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. Privacy Norms for Smart Home Personal Assistants. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.
- [3] Phil Adams, Mashfiqui Rabbi, Tauhidur Rahman, Mark Matthews, Amy Volda, Geri Gay, Tanzeem Choudhury, and Stephen Volda. Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild. *PervasiveHealth*, pages 72 – 79, 2014.
- [4] Khaled Alanezi and Shivakant Mishra. Incorporating individual and group privacy preferences in the internet of things. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–16, 2021.
- [5] Abdulmajeed Alqhatani and Heather Richter Lipford. “There is nothing that I need to keep secret”: Sharing Practices and Concerns of Wearable Fitness Data. In *Symposium on Usable Privacy and Security (SOUPS)*, 2019.
- [6] Andrew Besmer and Heather Richter Lipford. Moving beyond untagging: Photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1563–1572, 2010.
- [7] Cristina Bicchieri. *Norms in the Wild*. Oxford University Press, 2016.
- [8] George Chalhoub, Martin J Kraemer, Norbert Nthala, and Ivan Flechais. “It did not give me an option to decline”: A Longitudinal Analysis of the User Experience of Security and Privacy in Smart Home Products. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.
- [9] Chhaya Chouhan, Christy M LaPerriere, Zaina Aljalalad, Jess Kropczynski, Heather Lipford, and Pamela J Wisniewski. Co-designing for Community Oversight: Helping People Make Privacy and Security Decisions Together. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–31, 2019.
- [10] Robert B Cialdini, Linda J Demaine, Brad J Sagarin, Daniel W Barrett, Kelton Rhoads, and Patricia L Winter. Managing social norms for persuasive impact. *Social Influence*, 1(1):3–15, March 2006.
- [11] Zoë B. Cullen and Ricardo Perez-Truglia. The Salary Taboo: Privacy Norms and the Diffusion of Information. Technical Report NBER Working Paper No. 25145, 2020.
- [12] Sauvik Das, W. Keith Edwards, DeBrae Kennedy-Mayo, Peter Swire, and Yuxi Wu. Privacy for the People? Exploring Collective Action as a Mechanism to Shift Power to Consumers in End-User Privacy. *IEEE Security & Privacy*, 19(5):66–70, 2021.
- [13] Tamara Dinev and Paul Hart. An extended privacy calculus model for e-commerce transactions. *Information Systems Research*, 17(1):61–80, 2006.
- [14] Nora A Draper and Joseph Turow. The corporate cultivation of digital resignation. *New Media & Society*, 21(8):1824–1839, 2019.
- [15] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. Secrecy, flagging, and paranoia: Adoption criteria in encrypted email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 591–600, 2006.
- [16] Nina Gerber, Paul Gerber, and Melanie Volkamer. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, 77:226–261, 2018.
- [17] Erving Goffman. *Behavior in Public Places: Notes on the Social Organization of Gatherings*. The Free Press, New York, NY, 1966.
- [18] Samantha Hautea, Anjali Munasinghe, and Emilee Rader. That’s not me: Surprising algorithmic inferences. In *Poster presented at the 2020 Symposium on Usable Privacy and Security*, 2020.
- [19] Jong-yi Hong, Eui-ho Suh, and Sung-Jin Kim. Context-aware systems: A literature review and classification. *Expert Systems With Applications*, 36(4):8509–8522, 2009.
- [20] Yue Huang, Borke Obada-Obieh, and Konstantin Beznosov. Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy

- Risks. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, pages 1–13, 2020.
- [21] Kyle Irwin and Brent Simpson. Do Descriptive Norms Solve Social Dilemmas? Conformity and Contributions in Collective Action Groups. *Social Forces*, 91(3):1057–1084, February 2013.
- [22] Haiyan Jia and Eric P.S. Baumer. Birds of a feather: Collective privacy of online social activist groups. *Computers & Security*, 115:102614, 2022.
- [23] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 2018.
- [24] Kif Leswing. Apple’s ad privacy change impact shows the power it wields over other industries. *CNBC*, November 13 2021. <https://www.cnbc.com/2021/11/13/apples-privacy-changes-show-the-power-it-holds-over-other-industries.html>.
- [25] Ying Li, Jose D Contreras, and Luis J Salazar. Predicting Voice Elicited Emotions. The 21th ACM SIGKDD International Conference, pages 1969 – 1978, 2015.
- [26] Heather Richter Lipford, Madiha Tabassum, Paritosh Bahirat, Yaxing Yao, and Bart P. Knijnenburg. Modern Socio-Technical Perspectives on Privacy. pages 233–264, 2021.
- [27] Eden Litt and Eszter Hargittai. A bumpy ride on the information superhighway: Exploring turbulence online. *Computers in Human Behavior*, 36:520–529, 2014.
- [28] Hong Lu, Mashfiqui Rabbi, Gokul Chittaranjan, Denise Frauendorfer, Marianne Schmid Mast, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. Stresssense: detecting stress in unconstrained acoustic environments using smartphones. *Proceedings of the 2012 ACM international joint conference on Pervasive and ubiquitous computing*, 2012.
- [29] Ewa Luger and Tom Rodden. An informed view on consent for UbiComp. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 529–538, 2013.
- [30] David Lyon. *Surveillance Studies: An Overview*. Polity Press, Cambridge, UK, 2007.
- [31] Bryan Marshall, Peter Cardon, Amit Poddar, and Renee Fontenot. Does sample size matter in qualitative research?: A review of qualitative interviews in is research. *Journal of Computer Information Systems*, 54(1):11–22, 2013.
- [32] Nicole Meng, Dilara Keküllüoğlu, and Kami Vaniea. Owing and Sharing: Privacy Perceptions of Smart Speaker Users. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–29, 2021.
- [33] Dale T. Miller and Cathy McFarland. Pluralistic Ignorance: When Similarity is Interpreted as Dissimilarity. *Journal of Personality and Social Psychology*, 53(2):298–305, 1987.
- [34] Pardis Emami Naeini, Martin Degeling, Lujo Bauer, Richard Chow, Lorrie Faith Cranor, Mohammad Reza Haghghat, and Heather Patterson. The Influence of Friends and Experts on Privacy Decision Making in IoT Scenarios. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1 – 26, 11 2018.
- [35] Pardis Emami Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring How Privacy and Security Factor into IoT Device Purchase Behavior. Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1 – 12, 2019.
- [36] Bettina Nissen, Victoria Neumann, Mateusz Mikusz, Rory Gianni, Sarah Clinch, Chris Speed, and Nigel Davies. Should I Agree? Delegating Consent Decisions Beyond the Individual. Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1 – 13, 2019.
- [37] Helen Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79:119–158, 2004.
- [38] Bernice A. Pescosolido and Jack K. Martin. The stigma complex. *Annual Review of Sociology*, 41(1):87–116, 2015.
- [39] Sandra Petronio. *Boundaries of Privacy: Dialectics of Disclosure*. State University of New York Press, Albany, NY, 2002.
- [40] President’s Council of Advisors on Science and Technology. Big data and privacy: a technological perspective. Technical report, May 2014.
- [41] Emilee Rader, Samantha Hautea, and Anjali Munasinghe. I have a narrow thought process: Constraints on explanations connecting inferences and self-perceptions. In *Symposium on Usable Privacy and Security*, 2020.
- [42] Nichola J. Raihani and Vaughan Bell. An evolutionary perspective on paranoia. *Nature Human Behaviour*, 3(2):114–121, 2019.
- [43] Priscilla M Regan. Privacy as a Common Good in the Digital World. *Information, Communication & Society*, 5(3):382–405, 2002.

- [44] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. Third edition edition, 2016.
- [45] John S Seberger, Marissel Llavore, Nicholas Nye Wyant, Irina Shklovski, and Sameer Patil. Empowering Resignation: There’s an App for That. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–18, 2021.
- [46] Robert H. Sloan and Richard Warner. Why are Norms Ignored? Collective Action and the Privacy Commons. Available at SSRN: <https://ssrn.com/abstract=3125832>, February 18, 2018.
- [47] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke. Information Privacy: Measuring Individuals’ Concerns about Organizational Practices. *MISQ*, 20(2):167 – 196, 1996.
- [48] Daniel J Solove. Introduction: Privacy self-management and the consent dilemma. *126 Harvard Law Review*, pages 1880–1903, 2013.
- [49] Jose M. Such and Natalia Criado. Resolving multi-party privacy conflicts in social media. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1851–1863, 2016.
- [50] Kurt Thomas, Chris Grier, and David M. Nicol. un-Friendly: Multi-party Privacy Risks in Social Networks. In Mikhail J. Atallah and Nicholas J. Hopper, editors, *Privacy Enhancing Technologies*, pages 236–252, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- [51] Lev Velykoivanenko, Kavous Salehzadeh Niksirat, Noé Zufferey, Mathias Humbert, Kévin Huguenin, and Mauro Cherubini. Are those steps worth your privacy? fitness-tracker users’ perceptions of privacy and utility. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(4), 2022.
- [52] Richard Warner. Notice and choice must go: The collective control alternative. *SMU Science and Technology Law Review*, 23(2):173–198, 2020.
- [53] Pamela Wisniewski, A.K.M Namul Islam, Heather Richter Lipford, and David C Wilson. Framing and measuring multi-dimensional interpersonal privacy preferences of social networking site users. In *Communications of the Association for Information Systems*, volume 38, 2016.
- [54] Justin Wu and Daniel Zappala. When is a Tree Really a Truck? Exploring Mental Models of Encryption. In *Symposium on Usable Privacy and Security*, pages 1–16, 2018.
- [55] Michael Zimmer, Priya Kumar, Jessica Vitak, Yuting Liao, and Katie Chamberlain Kritikos. ‘There’s nothing really they can do with this information’: unpacking how users manage privacy boundaries for personal fitness information. *Information, Communication & Society*, 23(7):1020–1037, 2020.

Appendix

A Participant Descriptives

Two rounds of interviews were conducted. The first focused on activity trackers (e.g., Fitbit), and the second focused on voice assistants (e.g., Amazon Alexa, Google Assistant, Apple Siri). Recruiting for the second round of interviews commenced after data collection of the first round was completed. Participants were recruiting using a subject pool composed of volunteers from the community surrounding a large university in the midwest region of the United States, and by snowball sampling on social media to obtain greater geographic diversity in the sample. This appendix presents overall descriptive statistics for both samples.

Note that at the end of each interview, participants were asked to fill out a brief demographic questionnaire which included questions from the Collection and Unauthorized Secondary Use subscales of the concern for information privacy (CFIP) instrument by Smith, Millberg and Burke [47]. These data were not analyzed for this paper; descriptive statistics are presented here as background. The privacy concern questions are listed below. The instructions were: “Here are some statements about personal information. From the standpoint of personal privacy, please indicate the extent to which you, as an individual, agree or disagree with each statement.” The 8 items were presented in random order to each participant, and were answered on a 5-point Likert scale where 1 was Strongly Disagree and 5 was Strongly Agree.

Collection Subscale

- It usually bothers me when companies ask me for personal information.
- When companies ask me for personal information, I sometimes think twice before providing it.
- It bothers me to give personal information to so many companies.
- I’m concerned that companies are collecting too much personal information about me.

Unauthorized Secondary Use Subscale

- Companies should not use personal information for any purpose unless it has been authorized by the individuals who provided the information.
- When people give personal information to a company for some reason, the company should never use the information for any other reason.
- Companies should never sell the personal information in their computer databases to other companies.
- Companies should never share personal information with other companies unless it has been authorized by the individuals who provided the information.

Age	
Mean	36
Min	20
Max	72
SD	12
Gender	
Man	24
Woman	40
Other	1
Collection Subscale	
Mean	3.72
Min	1.25
Max	5
SD	0.81
Secondary Use Subscale	
Mean	4.43
Min	2.25
Max	5
SD	4.75

Table 1: Demographics of the 65 participants (30 from round 1 and 35 from round 2). Two participants (one from each sample) did not complete all CFIP [47] subscale items, and were excluded from those descriptives.

B Semi-Structured Interview Questions

Warm-Up Both rounds of interviews began with about 10-15 minutes in the interviewer introduced the study and asked some warm-up questions. These questions focused on learning more about the specific technology the participant used, the language and terminology they used to refer to it, the situations in which they used it, and the kinds of things they used it for. The interviewer did not mention privacy in the introduction to the study, and only asked follow up questions about privacy if the participant spoke about it first.

Transition to Scenarios After the warm-up questions, the interviewer transitioned to the hypothetical scenarios. At this point, the interviewer said something resembling:

“Now, we’re going shift our focus a bit from how you use your [technology], to thinking about some hypothetical scenarios about [technology]. The scenarios are different things that all might be possible in the near future, using the different kinds of information that [technology] can collect. We’d like you to imagine that the scenario is something that can really happen.

The scenarios are designed to stretch your imagination and get you to think about ways of using a [technology] that you may not be used to, and how information generated by using a [technology] might be collected and used in the future.

What I’m most interested in is your impressions

and ideas about different ways people might react to each scenario. So I have some follow-up questions for each scenario related to that. Do you have any questions?”

Questions about Scenarios The following questions were asked about each scenario presented to the participants:

- What are some different kinds of reactions people might have if [technology] could do this?
- A few participants had a hard time getting started talking about reactions to the scenario. Questions like these were used to prompt them to begin speaking about what they were thinking.
 - Can you imagine someone who would or wouldn't mind [technology] knowing this kind of information?
 - Tell me more about what makes it hard to imagine the scenario.
 - Who were you thinking of that might react like that?
 - Why do you think they would react that way?
- Do you feel like most people would think it is ok or not ok to use [technology] if it can know [information from scenario]?
- (If the participant hadn't answered this question yet...) How would you personally feel about using [technology] if it could know [information from scenario]?

C Text of Hypothetical Scenarios

Each round of interviews included six hypothetical scenarios involving possible future uses of data that could be collected

by the technology (activity trackers or voice assistants). The scenarios are purely speculative, designed to seem plausible, but for the most part probably not something these technologies were actually doing at the time the interviews were conducted. Table 2 on the next page presents the text of each scenario used in the study.

Scenario 1 (S1) and Scenario 5 (S5) are very similar. The other scenarios were necessarily somewhat different, as the two technologies were quite different from each other and collected different kinds of data. However, even where the scenarios were different there are still some parallels:

- Scenario 2 (S2): both versions involve something that very closely resembles an existing use case for the two technologies.
- Scenario 3 (S3): both versions involve information that could be used to infer something about the user's health.
- Scenario 4 (S4): both versions involve always-on monitoring some type of information about the user's home environment
- Scenario 6 (S6): both versions involve information that some users might perceive as being in the public interest, that must be collected about the user and then aggregated across a group to produce a ranking

Scenario 6 in both rounds of interviews was a little bit different in that for both technologies it was about a societal issue (carbon footprint, crime) that has interdependent consequences beyond individual users. In other words, one person's carbon footprint or criminal activity in or near their home affects other people in the community (i.e., the environment or property values) in ways that oversleeping or the data and inferences in the other scenarios do not.

-
- S1** *Activity Tracker:* Imagine a wearable sensor device that a user wears to bed. Some people use their activity trackers this way already. This hypothetical device can use information about the user’s movements and alarm settings on their smartphone to know how many times the user overslept last week.
Voice Assistant: Imagine a person that uses [voice assistant] as an alarm clock, to set an alarm to wake them up in the morning. Some people already do this, actually. Asking [voice assistant] to set an alarm means that it could use information about how many times the user hit snooze in the morning, or how long the alarm goes off before the user shuts it off, to know how often the user overslept last week.
Sensor: Accelerometer (activity tracker), alarm time, interactions with device to snooze or stop the alarm
- S2** *Activity Tracker:* Imagine an app that can use information from a user’s wearable sensor device to make a graph or chart of when the user was sitting down at his or her desk at work each day last week.
Voice Assistant: Imagine that [voice assistant] can be activated accidentally based on hearing the wake word when the user didn’t actually intend to issue a command. This might happen if a user says the wake word when talking to someone else, or even when an actor in a TV commercial says it. This could allow [voice assistant] to know the content of some of the user’s conversations when they don’t mean to talk to the device.
Sensor: Accelerometer, GPS (activity tracker); Microphone (voice assistant)
- S3** *Activity Tracker:* Imagine that instead of time spent sitting down in a location, a wearable sensor device could use information about a user’s movements and location to count how many times [he or she] went to the bathroom yesterday.
Voice Assistant: Imagine that it’s possible to use [voice assistant] while preparing meals, to read recipes and provide cooking instructions. This means that it would have access to information about ingredients, cooking methods, and meals the user prepares, and could determine how healthy a person is based on his or her eating habits.
Sensor: Accelerometer, GPS (activity tracker); Microphone (voice assistant)
- S4** *Activity Tracker:* What if an app were able to use information from a wearable sensor device to observe something about the user’s environment based on their movements and altitude, like how many levels/floors there are in the user’s home? What are some reactions you think other people might have to a device that could know that?
Voice Assistant: What if [voice assistant] were able to use information from past voice commands to observe something about the user’s home environment, like how many different guests or visitors the user has over? This could happen based on analyzing things like vocal pitch and speaking patterns, or the number of different voices in the background when a command is spoken.
Sensor: Altimeter (activity tracker); Microphone (voice assistant)
- S5** *Activity Tracker:* Imagine that it is possible for a system that uses wearable sensors to know whether a user has young children at home or not. This could be possible based on information about the user’s movements, and GPS locations of places they visit, like playgrounds and parks.
Voice Assistant: Imagine that it’s possible for [voice assistant] to figure out whether the user is a parent who has a baby or toddler at home? This could be possible based on the content of the commands issued to the system, or the vocal pitch of the user, especially if a child asks [voice assistant] questions or directs it to play music.
Sensor: Accelerometer, GPS (activity tracker); Microphone (voice assistant)
- S6** *Activity Tracker:* Imagine that a system could estimate a user’s weekly carbon footprint, and rank it against the carbon footprint of other users in their area. A wearable sensor device that can detect a user’s movements and identify [his or her] GPS location can use this information to figure out when the user is in a moving vehicle, and estimate the carbon footprint based on that.
Voice Assistant: Imagine that [voice assistant] could estimate how safe or unsafe the user’s neighborhood is, and rank it against the safety level of the neighborhoods of other users in their region. This could be possible if at the same time as it is listening for the wake word, it is also listening for the sound of gunshots inside the home or nearby. Information about whether there has been gunfire at a location could be used to make a ranked list of each property and average that across a neighborhood.
Sensor: Accelerometer, GPS (activity tracker); Microphone (voice assistant)
-

Table 2: Text of the hypothetical scenarios read to participants. The text in [brackets] was replaced by the terminology the participant used during the interview to refer to the technology they had experience with. The first two voice assistant interviews (DA01 and DA02) used a different Scenario 3, about detecting stress based on vocal pitch and speech patterns. However, both participants strongly felt this scenario was not believable, so the scenario was revised for the remaining interviews. (Note that detecting stress levels from audio data is actually feasible [3, 25, 28].)

Sharing without Scaring: Enabling Smartphones to Become Aware of Temporary Sharing

Jiayi Chen
University of Waterloo

Urs Hengartner
University of Waterloo

Hassan Khan
University of Guelph

Abstract

Smartphone owners often hand over their device to another person for temporary sharing, such as for showing pictures to a friend or entertaining a child with games. This device sharing can result in privacy concerns since the owner’s personal data may become vulnerable to unauthorized access. Existing solutions have usability problems and neglect human factors of sharing practices. For example, since device sharing implies trust between people, explicitly hiding data may signal mistrust. Besides, an owner may fail to enable a sharing-protection mechanism due to forgetfulness or lack of risk perception. Therefore, we propose device sharing awareness (DSA), a new sharing-protection approach for temporarily shared devices, which detects a sharing event proactively and enables sharing protection *subtly*. DSA exploits natural handover gestures and behavioral biometrics for proactive sharing detection to transparently enable and disable a device’s sharing mode without requiring explicit input. It also supports various access control strategies to fulfill sharing requirements imposed by an app. Our user study evaluates handover detection over 3,700 data clips (n=18) and comprehensive device sharing processing over 50 sessions (n=10). The evaluation results show that DSA can accurately detect handover gestures and automatically process sharing events to provide a secure sharing environment.

1 Introduction

Prior research shows that it is common for smartphone users to temporarily share their devices with another person for trust

and convenience [20, 31]. For example, a smartphone user may show pictures stored on the phone to a friend or hand over the device to a child to play games. This device sharing can result in negative experiences due to all-or-nothing access control [12]. Liu et al. [27] report that 86% of the participants in their user study always kept their phone in sight when sharing, which puts an extra burden on the owner and may make the sharee feel mistrusted. (We use the terms “owner” to refer to a smartphone owner sharing their device and “sharee” to refer to people a device is shared with.) Hang et al. [12] report that the majority of participants in their user study wanted the ability to share only specific apps and features.

Existing solutions for temporary device sharing emphasize *how to* impose access restrictions on sensitive apps and data during sharing. They allow the owner to add a guest account [17], pin an app [15, 16], or launch apps with limited features (e.g., a camera app without a view of existing pictures) when the device is locked. However, most solutions require an explicit action from the owner before sharing the device. Vulnerabilities arise when humans are forgetful [36] or lack risk perception of certain situations [4]. Owners may forget to switch to the guest account or to pin an app before sharing. Furthermore, sharing behavior is closely related to trust [38]. An owner explicitly enabling these solutions signals mistrust for device sharing between the owner and sharee [1, 13, 20, 30, 32]. Besides, these solutions are inadequate for some sharing scenarios. A guest account works well to entertain children with a game but not for temporary sharing with spouses. Pinning an app grants access only to the current foreground app. It is insufficient when a sharee needs access to multiple shareable apps.

We introduce device sharing awareness (DSA) to address *when to* enable device sharing solutions. DSA should: 1) *proactively detect device sharing* instead of requiring an owner to remember performing a predefined action, 2) *continuously identify the owner* to prevent unauthorized access and ensure that only the owner has full access, 3) *be exception-resistant* to automatically handle possible false detection or exceptions and mitigate the exposure of sensitive resources.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022,
August 7–9, 2022, Boston, MA, United States.

For an outcome of device sharing awareness, DSA should provide *flexible access control* to choose an appropriate strategy based on the current app type.

To fulfill the above requirements, DSA automatically deals with all aspects of a device sharing event with little to no input from the device owner. For subtle and fast sharing detection, DSA continuously senses for device handover gestures using motion sensors and verifies the owner's identity using behavioral biometrics with high accuracy and low power consumption. Behavioral biometrics alone may make it hard to distinguish a sharing event from unauthorized access and rapidly react to the sharing event. When detecting a sharing event (or upon manual activation by an owner), DSA can enable app-level access control using allowlisting or blocklisting. Besides, it allows the shared app to adopt its sharing-specific access control strategies (if available). Other apps are also notified of the device sharing by DSA and can adopt their own sharing reaction (e.g., de-authenticating a user).

We conducted a user study and collected data from 18 participants to evaluate DSA. Our evaluation over 3,700 motion data clips shows that DSA can detect handover gestures accurately for 95% of the sharing events. On a public dataset containing 81-hour phone usage data from 100 users [40], DSA generated only 0.9 false positives per hour of continuous device use. For an average daily smartphone use of about three hours [11], DSA will generate about three false positives a day. We also tested the device sharing processing ability of DSA with a popular touch-based implicit authentication (IA) solution [10], which includes 50 complete device sharing sessions. DSA succeeded in detecting handover gestures in 48 sessions, and automatically handled 41 sessions without exceptions while its exception processing additionally recovered six sessions. DSA adopts adaptive sensing, and consumes 0.11% of battery per half hour at high-frequency sensing when there is significant movement, and consumes only 0.06% of battery per half hour at low-frequency sensing.

Our main contributions include: 1) A demonstration of the ability of low-cost proactive sharing detection using smartphone built-in sensors. 2) An open-source solution for Android¹ that secures sensitive data during device sharing, while mitigating human factors of forgetfulness and mistrust. 3) An extensive evaluation to demonstrate its practicality in terms of accuracy to detect sharing and battery consumption. 4) A public, labelled motion sensor dataset with over 3,700 sharing gesture clips for the research community.

2 Related work

Device sharing surveys. According to recent surveys, mobile device sharing is common in people's daily life [31] and even a systemic practice in some regions (e.g., South

¹The source code and the dataset are available at <https://github.com/cryspuwaterloo/DSA-Framework>

Asia [1, 4, 35]). Reasons for device sharing are not limited to economic consideration, help, convenience, or access to specific features. As a social and cultural practice, it is also driven by the need for maintaining social relationships and signaling trust among people [1, 4, 30]. However, as revealed by extensive qualitative studies [1, 20, 31], people still have privacy concerns over sharing their mobile devices given possible device misuse and exposure of sensitive or private data. For example, a social networking app may keep a user logged in due to its single-user design. A sharee can move to that app during sharing and access restricted data or functionality.

Access control for device sharing. Researchers have studied smartphone owners' security and privacy concerns with sharing different apps and called for access control mechanisms for device sharing [12, 20, 24, 31, 35]. Studies [12, 13, 20] show that all-or-nothing access control cannot meet the need for device sharing from both security and convenience aspects. Koushki et al. [24] show that app- or task-level access control can significantly reduce unnecessary or missed interventions compared to all-or-nothing access control. xShare [27] enables the owner to specify the resources to share and offers a restricted mode for the sharee but requires modification to the operating system. DiffUser [33] establishes a multi-user security model for Android smartphones, but it requires creating different accounts to apply different access control rules. SnapApp [6] adopts a time-constrained access control model where a short sliding gesture can activate a 30-second usage session. This scheme reduces the authentication overhead and enables quick device sharing, but the attacker can still launch an attack within the session. TreasurePhone [37] considers both environmental and user contexts to realize context-dependent access control to groups of apps. Overall, most existing systems need manual activation and lack interaction with third-party apps to secure sensitive resources. In comparison, DSA enables smartphones to proactively detect device sharing without being manually activated by an owner and provides flexible access control.

Trust. As trust is an important motivation of device sharing, we also need to take trust into account when designing device sharing solutions. A guest account for socially close sharees is deemed inappropriate since it signals mistrust [20, 30]. Explicitly hiding certain apps may also imply a lack of trust [1, 4]. Recent device sharing proposals have explored how to protect sensitive resources while not compromising trust. Seyed et al. [38] propose a modular smartphone comprising of multiple access-controlled hardware components to address the trust and convenience issues of device sharing. PrivacyShield [34] provides a subtle just-in-time privacy provisioning system, which enables the owner to quickly enable an access control rule by entering pre-defined touch gestures. Ahmed et al. [2] adopt two accounts for shared use and secret space, respectively, which can be accessed via the same interface but with different passwords. To address the trust issue, DSA takes control of the entire sharing process *proactively* and *auto-*

atically so that smartphone users do not need to specify or enable access control rules in front of a sharee. Note that DSA emphasizes the subtlety of *enabling* a device sharing solution. It does not try to hide from a sharee that the device is currently in a restricted environment, which is a design problem [1]. Achieving this goal reliably requires tremendous efforts of app developers to redesign their apps [2].

Activity detection. DSA uses smartphone motion sensors to detect a sharing event based on hand movements. Existing work [3, 28, 39] focuses on using motion sensors to detect specific hand gestures. DSA detects natural device handover gestures continuously so that owners do not need to remember to perform a pre-defined gesture for device sharing. Vaizman et al. [41, 42] propose a multi-modal system that uses various sensors on smartphones and smartwatches to recognize a person's behavioral context in natural environments, which is close to our purpose of detecting a sharing gesture in the wild. However, unlike behavioral contexts, such as walking and running, a sharing gesture lasts only one to two seconds and is not a repetitive or periodical activity. Nevertheless, we follow the feature selection from existing work [25, 41] to train our gesture detection model.

3 Device Sharing Awareness

3.1 Modeling temporary device sharing

Temporary device sharing is a social activity where a device owner shares certain resources on the phone with one or several sharees. The device sharing scenario targeted by our work is: The device is initially with its owner, and the owner directly hands it over to a sharee as a signal of granting temporary access. During sharing, a sharee should not have access to sensitive resources, including personal data (e.g., messages, photos) and critical operations (e.g., deleting files). We do not study device sharing where the device is initially not with its owner, or where a sharee can access the device indirectly (e.g., the owner puts the unlocked device on a table to pass the device) or without the owner's presence. Traditionally, this kind of sharing is enabled with separate user accounts or PIN sharing [1, 30]. We discuss this case in § 6.

We describe a sharing event with the following three-stage device sharing model:

1. **Pre-sharing.** The owner initially holds the phone. The owner unlocks the device and opens the app that contains the resources to be shared. Then, the owner passes the device to the sharee.
2. **Sharing.** The sharee holds the device and starts using the opened app. During sharing, the sharee should be able to access only the specified resources for sharing. For the multi-sharee scenario, sharees may pass around the device, but we still regard it as a single sharing event.

3. **Returning.** The (last) sharee finishes using the device and returns it to the owner. A sharing event ends only when the current user is confirmed to be the owner.

We define the *shared app* as the foreground app at the moment when sharing is initiated. Based on the owner's preferences, a sharee may be allowed to access further apps during sharing. The term *shared app* always refers to the original one.

3.2 Limitations of existing sharing solutions

Many technical solutions have been proposed to protect sensitive information from unauthorized access on a shared device. We classify these solutions into four categories based on their scopes and methods: 1) *Guest accounts* create an independent environment for sharees without access to the personal data of a device owner. However, it prevents sharees from accessing non-sensitive resources only available on the owner's account (e.g., non-sensitive photos, a public post on the owner's social networking app). 2) *App locks* (e.g., Samsung S Secure [7], Norton App Lock [26]) make an app require credentials (e.g., a PIN) for launching the app. App locks provide all-or-nothing access control: a device owner can only choose from sharing the entire app or nothing. It introduces unnecessary authentication overhead and does not apply to many common apps with personal data. For example, a browser app provides the essential web browsing function and may store the owner's passwords for auto-filling. 3) *App pinning* (e.g., Android Screen Pinning [16], iOS Guided Access [15]) restricts a sharee's access to the current foreground app only. While it is handy for single app sharing, it fully blocks access to other apps but imposes no restrictions on accessing in-app content of the foreground app. 4) *Vaults* (e.g., App Vault [18], Private Space [14]) allow owners to hide apps and files from sharees. A common practice is to provide two interfaces for shared access and private access, respectively. It provides finer-grained control over the shared resources compared to the other methods. However, vault solutions have been found to provide limited stealth functionality [2]: 1) Most vault apps on the market still provide an entry point that reveals the existence of a hidden vault. 2) They may only apply to specific file types (e.g., photos, text, videos).

There are still several gaps between the current practices and a desired device sharing solution:

1. **Lack of subtlety.** Ahmed et al. [1] have found that the act of locking or pinning an app or data may incur social challenges and raise suspicion, especially when it comes to device sharing with family members. Thus, a device sharing solution should be activated subtly and automatically by the device.
2. **Relying on a user's explicit input.** Many device sharing solutions require a user to manually trigger them. A user's forgetfulness or lack of risk perceptions [4] can

cause inaction to device sharing, resulting in the exposure of sensitive data. Besides, relying on a user's input can also result in poor usability since an owner may need to take additional steps (e.g., enter a PIN for app locks) to access certain resources during regular phone use.

3. **Coarse-grained access control.** Many solutions follow a simple access control model to grant all or nothing access to each app. However, it is preferable to give apps the option to adapt their own fine-grained access control strategies during sharing. For the browser app example, a sharee should be allowed to browse the web without accessing the owner's data.

As existing device sharing solutions [2, 6, 34] mainly address *how* to protect sensitive resources from unauthorized access during sharing, we focus on a novel perspective, device sharing awareness (DSA), to address *when* to (de-)activate such solutions. The device should be able to detect device sharing and identify the owner proactively and transparently. We present the following example to illustrate how DSA is expected to handle device sharing automatically:

In a coffee shop, Owen shares with his friend, Shannon, a bunch of travel photos stored on his smartphone. When he hands over the phone to Shannon, DSA automatically detects the sharing activity and notifies the gallery app so that the app can hide all photos labeled as private. At the same time, all notifications from messaging apps are silenced. During sharing, DSA allows Shannon to be redirected to the Map app by the location metadata of photos but not to move to any social networking apps to post photos. After Shannon finishes browsing the photos and returns the device, the system recovers as before the sharing activity.

3.3 Sharing detection

For minimizing the restrictions on an owner, a device sharing solution is supposed to take effect only when there is an ongoing sharing event. Therefore, an important requirement of device sharing awareness is to *proactively* determine the beginning and the end of a device sharing event. According to the device sharing model, we emphasize two factors for sharing detection: *sharing gestures* and *owner detection*. A sharing gesture is an indicator of a sharing event and implies that the owner authorizes the sharee to access the phone. We regard manual activation methods adopted by existing sharing solutions as *explicit* sharing gestures (e.g., buttons, touch-screen swipe gestures, and shortcut keys [15, 16, 34]). They explicitly indicate the beginning of a sharing event and trigger sharing solutions immediately. However, for subtlety and less reliance on explicit input, we also exploit an *implicit* sharing gesture, *the device handover gesture*, which can be directly sensed from the natural hand movements when the device is handed from one person to another. In this paper, we mainly investigate the detection of the device handover gesture.

While a handover gesture indicates the beginning of a sharing event, we cannot use it to determine the end of a sharing event since there may be multiple sharees passing around the device. Here, verifying the user's identity is essential: While a non-owner user is temporarily allowed to access the device during sharing, the device should ensure that the current user changes back to the owner at the end of a sharing event. A common practice for de-activating the sharing mode is to ask for explicit authentication (e.g., a PIN) to ensure the device has been returned to the owner. DSA should be able to determine if the current user is the owner proactively and transparently. It can be achieved by continuous and implicit authentication: A device can distinguish the device owner from other people based on biometrics, including continuous facial recognition [8, 29], voice recognition [44], or implicit authentication (IA) based on behavioral biometrics [19, 22]. In addition to determine the end of device sharing, owner detection can complement a device sharing solution in cases where a handover gesture is detected erroneously and can avoid false activation of the sharing mode.

Detecting sharing events based on owner detection alone, ignoring handover gestures, is insufficient. It is hard to distinguish a sharing event from unauthorized access of a stranger (e.g., a stranger using an unattended, unlocked phone without permission) since a non-owner user can be detected in both cases. Besides, continuous facial recognition may cause significant power consumption; voice recognition and behavioral biometrics require sufficient input data for identification, making the device slow to react to a sharing event. Thus, a crucial problem is how to combine handover detection and owner detection for detecting sharing events.

3.4 App types

Most sharing solutions impose access control on sharees to avoid access to sensitive resources. We name this restricted environment for device sharing as the *sharing mode*. Many apps contain both shareable and non-shareable content, while some apps may involve redirection to other apps to process specific requests. Thus, existing solutions that only restrict the sharee to the current foreground app cannot fulfill these requirements. Based on whether resources in an app are shareable and existing taxonomies [13, 27], we classify apps into the following three categories:

- **Shareable apps.** Apps that are completely shareable without any sensitive resources, such as games or weather apps. A sharee has full access to such apps.
- **Semi-shareable apps.** Apps that contain both shareable and non-shareable resources, such as social networking or photo gallery apps. A sharee can access the shareable resources during sharing without access to personal data and sensitive operations in such apps.
- **Non-shareable apps.** Apps that contain no shareable content, such as system settings, banking, or corporate

apps. During sharing, a sharee should have no access to such apps. Specifically, corporate apps have higher security requirements and need to react to the sharing event even when running in the background (e.g., terminate the session, disconnect from a remote service).

Our goal is to design a device sharing access control strategy that meets the requirements of different kinds of apps. Moreover, we need to consider some special apps or components such as the home screen and the notification bar, most of which are provided by the system launcher in Android.

3.5 Threat model

Device sharing involves two kinds of roles: an owner and one or more sharees. A malicious owner is out of the scope of DSA since the owner can disable DSA and launch attacks on a sharee (e.g., accessing a sharee’s account that is not properly logged out after sharing, or sniffing passwords the sharee enters on a website). Instead, we focus on attacks from sharees. We classify sharees into two categories: A *benign sharee* uses only the specified resources without any intention of accessing sensitive information or other apps during sharing. However, a benign sharee can do accidental mis-operations that expose private data (e.g., switch to other apps). It is also possible that some apps may push notifications that contain sensitive information to a benign sharee (e.g., an email notification with a preview). A *malicious sharee* targets other apps than the shared app and intends to access private information during sharing. They may try to leave the current app and access unauthorized resources. A malicious sharee may be aware of the existence of the protection mechanisms, such as screen lock and implicit authentication, and attempts to bypass them. A malicious sharee may also know of the existence of our proposed solution and launches attacks accordingly.

4 System design

We now introduce the design of our framework. We present how DSA works based on different states, its main modules, complete workflow, and exception handling strategies.

4.1 State transition

We define three states of a device: *normal*, *sharing*, and *locked*. In state “normal”, the user has full access to the device. In state “sharing”, the user has limited access to the device and cannot access sensitive resources. In state “locked”, the user has no access to the device and needs to explicitly authenticate. Fig. 1 shows the state transition among the three states. Existing app pinning solutions fully rely on manual operations to switch among the three states (see Fig. 1 Loop ①): 1) pin an app to start sharing (i.e., limit access only to the current foreground app), 2) unpin an app to end sharing and

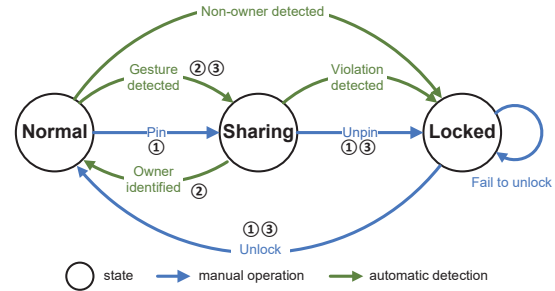


Figure 1: State transition of device sharing. Three sharing loops: ① explicit sharing loop (manual option), ② implicit sharing loop (handover gesture + owner detection), ③ hybrid sharing loop (handover gesture + manual unlock).

lock the device, 3) authenticate the user to return to normal state. DSA keeps this loop to allow users to start or end the device sharing manually.

Following § 3.3, we introduce an implicit sharing loop (see Fig. 1 Loop ②) as a new trigger mechanism: 1) Sharing: If DSA detects a handover gesture, the state changes from “normal” to “sharing”. 2) Returning: If DSA confirms the owner’s identity, the state changes back to “normal”. Note that detecting a handover gesture, which may occur when a sharee returns the device, cannot be used to end a sharing event given possible multi-sharee cases or gesture spoofing attacks (i.e., the sharee fakes a handover gesture). In the implicit sharing loop, DSA can handle device sharing and secure sensitive resources without locking the device or asking for manual actions by the owner. However, state “locked” is still useful for processing violations (see § 4.6). DSA allows a hybrid sharing loop (see Fig. 1 Loop ③) where DSA detects a handover gesture to move into sharing state while the owner or sharee have to manually end sharing, and explicit authentication is required to move back into state “normal”.

4.2 Handover detection

We use the device handover gesture as a trigger of an implicit sharing loop. A handover gesture lasts only a few seconds and does not occur frequently. Compared to the typical gesture recognition problem, a handover gesture is performed in a natural manner rather than a specified motion (e.g., drawing a circle). The key to handover gesture detection is to study the common patterns of handover gestures and distinguish them from similar motions (e.g., switch hand).

Pilot experiments. We conducted a pilot experiment to investigate possible handover gesture patterns for feature selection: one experimenter, acting as a device owner, handed over a Google Pixel phone to another person (i.e., sharee) with two different position settings: 1) the owner handed over the phone from their right hand to the sharee sitting in front of them; 2) the owner handed over the phone from their right hand to the sharee sitting next to them. Each setting was repeated ten times. We collected data from the accelerometer

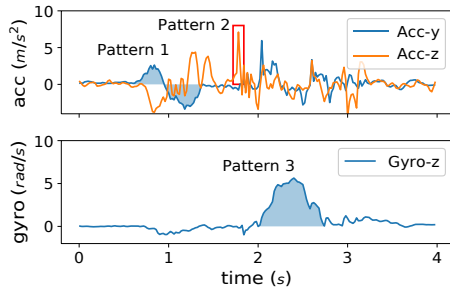


Figure 2: Handover patterns. 1. (horizontal movement): the device travels a distance in the xy-plane, where acceleration follows a sine curve like pattern; 2. (spike): When the sharee catches the device, a spike appears on the z-axis of acceleration; 3. (rotation): the device is rotated either by the owner or by the sharee to adjust the orientation.

and gyroscope at a sampling rate (denoted as f_s) of 50Hz. We use a software linear acceleration sensor provided by Android, which isolates gravity from raw acceleration data with the help of the gyroscope. The collected data includes linear acceleration and rotation speed on the three axes.

According to the collected sensor data, we observed that the length of a handover gesture is about two to four seconds. We also observed three patterns and exemplify them in Fig. 2. The observation shows the possibility to detect a handover gesture with motion sensors. It also helps us in determining features and targeting possible misleading activities that share similar patterns with handover gestures. For example, the acceleration readings of a horizontal hand movement follow a sine curve like pattern, which can be described by time-domain waveform features. A spike on the z-axis of acceleration resulting from a slight fluctuation of catching a device can be captured by entropy-based features. A misleading activity with similar patterns is a user’s passing the device from their left hand to their right hand (i.e., switching hand).

Feature extraction. To proactively detect handover gestures, the device continuously collects motion data from the accelerometer and the gyroscope. We first divide the collected time series data into fixed-size, overlapping segments, where the sampling rate is f_s , the segment length is d seconds (equal to $f_s \cdot d$ samples), and the interval between the start of two consecutive segments is p seconds. (Fig. 7 shows an example of segmentation.) The choices of d and p affect the detection performance. If d is too small, it is hard to capture the handover patterns from a data segment; if d and p are too large, it takes more time to capture handover. We investigate the impact of different settings on detection performance in § 5.

After segmenting the raw data, we extract the following features for each segment: We first calculate the magnitude of linear acceleration, $m = \sqrt{a_x^2 + a_y^2 + a_z^2}$. Together with each axis of linear acceleration and gyroscope data, we use common statistics widely adopted in gesture detection [3] and activity recognition [25] including: average, standard deviation, max-

imum, 25th percentile, median, 75th percentile, sum, double integration, and range. Also, we measure root mean square (RMS) [9] of the readings to capture time-domain wavelet patterns: $\text{RMS}(\mathbf{v}) = \sqrt{(v_0^2 + v_1^2 + \dots + v_{n-1}^2)/n}$, where $\mathbf{v} = \{v_0, v_1, \dots, v_{n-1}\}$ is a series of n sensor readings. Value and time entropy [41] measure sudden changes in a signal. We calculate the value entropy by quantizing all magnitude values to a 20-bin histogram for a moderate granularity. For time entropy, we normalize all sensor readings to form a probability distribution and calculate $H(|\mathbf{v}|) = -\sum_{i=0}^n \frac{|v_i|}{\sum |\mathbf{v}|} \log \frac{|v_i|}{\sum |\mathbf{v}|}$. Furthermore, to include correlations between different axes, we calculate the correlation coefficient between every two axes. In total, there are 87 features.

Classification. Based on the extracted features, DSA uses a pre-trained classifier to determine if the current segment belongs to a handover gesture. We adopt an offline learning strategy and train a generic classifier before deploying the system. For an online strategy, data labelling is challenging since it may need a user’s feedback to position a sharing event. Besides, our evaluation results in § 5 show the feasibility of applying a generic classifier to different users. To reduce false positives, we use a sliding window strategy that makes decisions based on several consecutive segments: if two consecutive segments are classified as positive, the system concludes that a sharing event is happening.

Adaptive sensing. Given that proactive handover detection is always running in the background, its power consumption is a concern. We adopt an adaptive sensing strategy to reduce battery consumption. The accelerometer and gyroscope initially collect raw motion data at a sampling rate at 10 Hz. When significant movement is detected (i.e., the acceleration magnitude exceeds a pre-defined *activation threshold*), it switches to a high sampling rate of 50 Hz and conducts handover detection. When the device is stationary for a period of time, the sampling rate is lowered to 10Hz. This strategy reduces unnecessary computations when the device is stationary.

4.3 Owner detection

Owner detection is provided by continuous and implicit authentication (IA) mechanisms. DSA relies on IA results to determine if the current user is the owner or not. Biometric mismatch results in a negative IA result, indicating that the current user is not the owner. In state “normal”, IA mechanisms are running continuously to prevent unauthorized access from non-owner users. They will lock out the current user upon negative IA results. In state “sharing”, IA mechanisms are automatically configured not to block users upon negative results as they indicate an ongoing sharing event. Once the IA results change from negative to positive in this state, DSA regards it as the end of a sharing event.

We incorporate existing IA mechanisms for owner detection and do not design a new IA mechanism. The selection of

IA mechanisms should take accuracy, availability, detection latency, and battery consumption into consideration. Ideally, IA mechanisms with low false rejection rate and low battery consumption are preferred in state “normal” since a device is not under sharing most of the time. In contrast, IA mechanisms with low false acceptance rate and short detection latency are preferred in state “sharing” to determine if the device has been properly returned to the owner. Owner detection can adopt multiple modalities to ensure accuracy and availability. For example, if touch-based IA produces a positive result, the device can automatically conduct face recognition to determine if the current user has changed back to the owner. It helps to ensure high accuracy with avoiding battery consumption of continuous facial recognition.

Considering the availability of various behavioral biometrics on smartphones, we use the touchscreen input biometric and adopt Touchalytics [10], whose reported equal error rate is below 4%, as the default scheme in our evaluation. Touchalytics extracts 29 features from touch events to capture behavior related to acceleration, velocity, duration, orientation, width, pressure, and trajectory length. It performs classification for each touch event and authenticates the user based on the results of several consecutive touch events. According to Khan et al. [22], the battery consumption of touch-based IA is low enough for continuous owner detection.

4.4 Access control for device sharing

For improved usability, DSA adopts different strategies for enforcing app-level access control based on the shared app type. It also notifies apps of sharing status changes so that a shared app can change its behaviors to a shared mode. The common app-level access control strategies involve: 1) **Blocklist**. A device owner can determine a list of non-shareable apps that cannot be accessed by a sharee. In state “sharing”, the system rejects all access attempts to the apps on the blocklist. Besides, hiding non-shareable apps is also applicable to block a sharee’s access in a subtle way. 2) **Allowlist**. A sharee is only allowed to access a list of shareable or semi-shareable apps. App pinning methods can be regarded as a kind of allowlist that makes only the current foreground app available to a sharee. 3) **Profile switching**. Mobile operating systems (e.g., Android) organize user data in profiles and allow the programmatic switching of an app’s profile [5]. It enables a sharee to use a semi-shareable app without accessing the owner’s data. If this feature is not available, an app can switch a profile as part of in-app sharing control (see below).

While an owner can configure access control strategies for different apps, DSA can infer what access control strategy to adopt: In most cases, DSA sets the current foreground app as a shared app and automatically adopts an allowlist-based strategy to restrict a sharee’s access to the shared app and any shareable or semi-shareable apps redirected to from the shared app. If there is “no app” running in the foreground

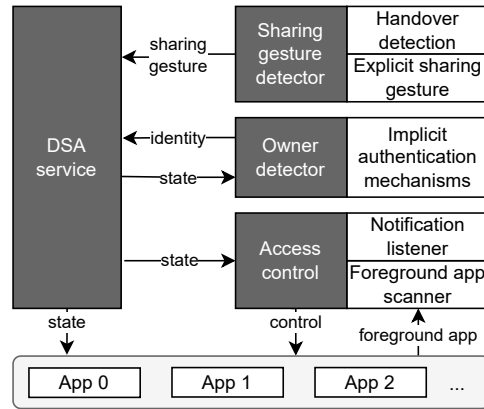


Figure 3: DSA architecture.

(e.g., the current foreground app is a launcher or home screen), DSA enables blocklist-based access control.

In addition to app-level access control, an app may have its own device sharing control strategies. Possible options include switching to guest mode, disabling user-specific content, logging out the owner’s account, etc. For example, a camera app can provide only the camera function without revealing any local photos. As suggested by existing studies [1, 2], it is important for apps to incorporate the “shared use” paradigm into their current design to provide more fine-grained in-app sharing control. In this case, DSA can provide important device sharing notifications to these apps to help them decide whether to enable such a shared use design.

4.5 DSA workflow

Fig. 3 shows the architecture of DSA. The high-level workflow of DSA follows the three stages introduced in the device sharing model in § 3.1.

Pre-sharing. The sharing gesture detector and the owner detector run continuously. At state “normal”, the owner detector is performing continuous authentication to reject non-owner users. Once a handover gesture is detected or the owner explicitly starts the sharing mode, the DSA service updates the current state to “sharing” and adopts an access control strategy according to the current foreground app. It also broadcasts the device sharing signal to other apps so that they can enable their own sharing mode or other reactions such as requesting re-authentication for the next access.

Sharing. The device is in state “sharing” and the sharing mode is enabled. The foreground app scanner continuously checks if the sharee is authorized to access the current foreground app. It rejects any unauthorized access to sensitive resources based on the access control strategy by redirecting a sharee to the shared app given possible mis-operations. If the mis-operations reach a pre-defined threshold, the DSA service locks the device. The notification listener intercepts incoming notifications to filter out the ones from non-shareable and semi-shareable apps. The blocked notifications are temporarily stored during sharing. The owner detector keeps verifying

if the current user is the owner and stops blocking non-owner users (i.e., negative IA results).

Post-sharing. Once the current user is identified as the owner or the owner manually ends sharing and passes explicit authentication, the DSA service updates to state “normal”. The DSA service notifies the foreground app scanner and the notification listener for lifting the access restrictions. The notification listener shows the owner all cached notifications that were missed during sharing. The owner detector resumes to defend against unauthorized access from non-owner users. The DSA service then broadcasts the state change to other apps so that they can revoke the changes made for device sharing.

4.6 Exception processing

As the implicit sharing loop allows DSA to handle device sharing automatically without a user’s explicit input, exceptions may occur, resulting from false detection, mis-operations, or attacks, and cause security or usability issues. It is critical to have an exception processing mechanism to recover from exceptions and mitigate their negative impact. Specifically, it needs to minimize the chance of sensitive resources exposed to a sharee. We classify exceptions into four types and provide solutions accordingly. In addition, in our user study (see § 5.3), we investigated the exceptions that DSA may encounter and how efficiently it handled these exceptions.

Non-owner user detected in state “normal”. When the owner detector detects a non-owner user, it locks the device and asks for re-authentication, such as a PIN code or password. There are three situations: 1) the current user is an attacker, and the owner detector successfully prevents unauthorized access, which is not an exception of device sharing; 2) the current user is the owner, and the owner detector falsely rejects the owner, which is a failure of the adopted IA mechanism; 3) the current user is the sharee, and the owner detector makes a correct detection, but the sharing gesture detector failed to capture the sharing event. Therefore, we need to distinguish the second and third situations. If the user passes re-authentication, the DSA service prompts a dialog to confirm if a sharing event was initiated. If so, it updates the sharing state and starts the sharing mode.

App exception. An app exception happens when a sharing event is detected but the current foreground app is invalid. It can be one of the following invalid apps: 1) a non-shareable app: DSA blocks the access and re-authenticates the owner. If the non-shareable app is logged in with the owner’s account, the current session of the app will be immediately ended. 2) system launcher: it provides entry points to all apps on the smartphone. Since no app is specified for sharing, DSA applies a blocklist-based access control strategy. The sharee is prevented from accessing non-shareable apps, and the notifications of non-shareable apps are hidden.

False positives of the handover detector. If the handover detector falsely detects a handover gesture when there is no

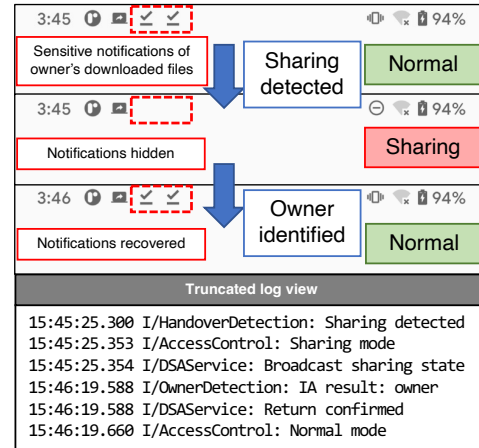


Figure 4: DSA Service Example: 1) At 3:45pm, DSA detected a sharing event and enabled the sharing mode; 2) During sharing, sensitive notifications were hidden, and DSA broadcast the sharing signal; 3) At 3:46pm, DSA identified the owner and ended the sharing mode with recovering the hidden notifications. Note: the first icon in the notification bar means the device is connected to a computer (for logging purposes); the second icon indicates that DSA Service is running; the third and fourth icons are the sensitive notifications.

sharing event, the DSA service still moves to state “sharing”, which causes inconvenience to owners. However, the owner detector can help correct false positives. If the owner continues using this app, the owner detector can identify the owner, and the system moves back to state “normal”. Even if the owner detector also happens to make a false detection and mistakenly regards the owner as a sharee, the owner can still explicitly end the sharing mode and re-authenticate. In § 5.2, we evaluate how the owner detector addresses false positives.

App redirection. A shared app may involve resources that redirect to other apps, such as a URL to be opened in a browser. DSA allows redirection to shareable and semi-shareable apps. Note that an app can activate its sharing mode by acquiring the sharing state from the DSA service at startup.

4.7 Implementation

We implement our demo DSA solution on Android as a service and release the source code¹. Developers and researchers can incorporate DSA into their device sharing solutions for automatic (de-)activation. Developers of third-party apps can set up a broadcast receiver to obtain the sharing notifications from the DSA service for enabling their in-app sharing control. Fig. 4 illustrates the log view and the changes of the notification bar at different states of an implicit sharing loop to reflect how DSA automatically handles device sharing. We can see that DSA automatically hid the sensitive notifications after detecting a handover gesture and recovered them once the user changed back to the owner. During this process, the owner did not need to manually enable and disable the sharing mode.

5 Evaluation

We first evaluate handover detection as it plays an important role in starting the implicit sharing loop. Then, we test how DSA coordinates handover detection and owner detection to automatically handle sharing events. We received approval from our IRB for the user study reported in this work.

5.1 Evaluation setup

Study description. We conducted a user study advertised as “the evaluation of context detection techniques for smartphone sharing”, and recruited 18 participants (5 females and 13 males) through word-of-mouth advertising. Eleven participants were between 18–29 years, five were between 30–39 years, and two were above 40 years of age. 13 participants were related to the field of Computer Science and the rest were in non-related fields. The study consisted of two parts: handover detection and device sharing. Participants chose to complete the first part only or both parts. 10 of 18 participants completed both parts. Participants received \$25 remuneration for completing the whole study (\$15 for completing the first part only). Due to the pandemic, most experiments happened remotely, and participants were instructed and supervised using a videoconferencing platform. Participants could choose to use a provided experiment smartphone or to install a data collection app on their devices. The phones recorded in the evaluation include Google Pixel, Google Pixel 3, Samsung S8, Xiaomi Redmi 5, and Huawei P9.

Model setup. For the detection of handover gestures, we used Support Vector Machines (SVM) and Neural Networks (NN) to train the gesture detection model and use it for classification. Considering the NN model’s superior performance and the increasingly mature support for NN on today’s smartphones, we adopted NN in our evaluation. The input layer of the NN is the feature vector (size=87) of each segment. The model includes two hidden fully-connected layers using ReLU as the activation function: one 64-neuron layer and one 48-neuron layer. We apply 10% dropout in between two hidden layers to reduce overfitting. The output layer uses Sigmoid as the activation function since our gesture detection is a binary classification task. We use the cross entropy loss function and Adam optimizer for model training. We set the number of epochs as 120 and the batch size as 128. For the training set, the positive instances were from handover gestures, and the negative instances were from movements sharing similar patterns with handovers. Given the low frequency of sharing events in practice, an imbalanced training set reflecting the actual distribution may make the model focus on detecting non-handover gestures [21]. Thus, we adopt a balanced training set where positive and negative instances are evenly distributed, and use 10% of the data for validation.

Metrics. Handover detection involves segmenting motion data, classifying each segment, and making decisions based

User#	1	2	3	4	5	6	7	8	9	10	11	12
AUC	0.98	0.98	0.98	0.98	0.99	0.99	0.97	0.97	0.97	0.96	0.97	0.96
EER	0.07	0.05	0.06	0.07	0.03	0.02	0.07	0.08	0.03	0.07	0.04	0.09

Table 1: Segment-level experiments: Per-user models.

User#	1	2	3	4	5	6	7	8	9	10	11	12
AUC	0.94	0.96	0.98	0.97	0.99	0.95	0.90	0.90	0.97	0.93	0.94	0.97
EER	0.11	0.10	0.09	0.10	0.04	0.12	0.16	0.15	0.07	0.15	0.11	0.09

Table 2: Segment-level experiments: Cross-user models.

on a number of consecutive segments. For segment-level classification, we evaluate the classifier performance based on its receiver operating characteristic (ROC) curve and use area under curve (AUC) and equal error rate (EER). For event-level detection, we use precision, recall, and f1-score to evaluate the overall detection performance under different settings. To measure the reaction time of each positive gesture detection, we use its elapsed time after the moment when the participant receives the instruction to hand over the device.

5.2 Evaluation of handover detection

The first part of our user study evaluated the accuracy of handover detection. Training a gesture detection model requires both handover (i.e., positive) and non-handover (i.e., negative) data to distinguish handover gestures from other movements. Thus, the user study involved two-participant handover tasks and single-participant non-handover tasks. In the handover tasks, participants were asked to hand over a smartphone from either their left or right hand in two directions: 1) Face-to-face: the owner was in front of the sharee. 2) Side-by-side: the owner was next to the sharee. Participants also performed the handover tasks with random directions to provide diverse handover data, where they randomly adjusted their relative positions each time. For each pair of participants, one participant handed over the device to the other at least 20 times per direction. Then, they swapped roles and repeated. In the non-handover tasks, we recorded motion data for activities having similar patterns with handover, such as switching hand, putting the device down, rotating the device, and random movements with combining device rotations and movements in different directions. All participants completed each single-participant task (e.g., switching hand) 20 times. Each data clip of handover or non-handover events lasts 5s to 10s.

In total, we collected 2044 positive and 1737 negative clips.

5.2.1 Cross-user experiments

A pre-trained gesture detection model should work on a new user (i.e., low user dependence) without retraining. Thus, we evaluate the cross-user performance of gesture detection from the perspectives of AUC and EER. A high AUC and a low EER indicate that a model can distinguish handover gestures from these activities better. We split each data clip into segments ($d = 2s, p = 1s$). For positive events, we focus on data segments that have 50% overlap with this time interval

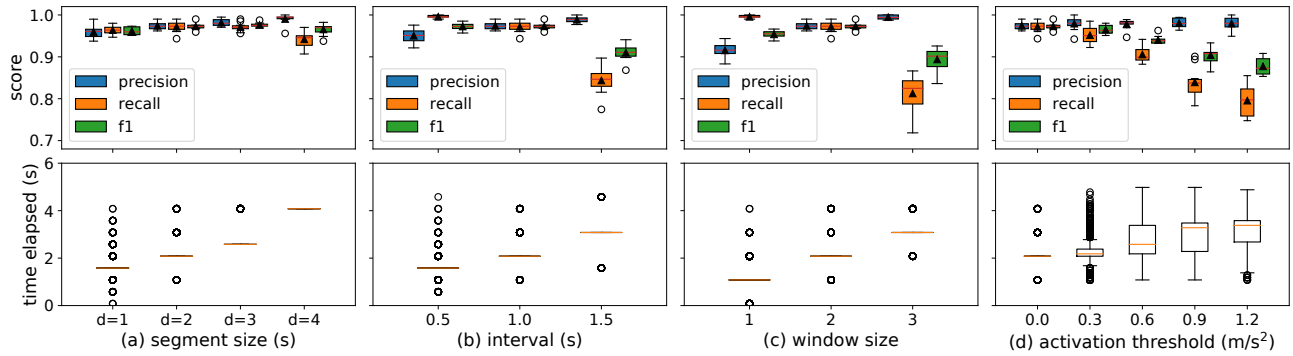


Figure 5: Event-level experiments: impact of different parameters on the detection performance and latency.

and label them as positive segments. We label all the segments from negative events as negative segments.

First, a cross-user model, which is pre-trained with hybrid data from multiple users, should have comparable accuracy as a per-user model. As a reference, we run a 10-fold cross validation to test the performance of per-user models for 12 participants using the same Google Pixel phone. We split each user’s data in 10 subsets and use one subset as the testing set and the other nine subsets as the training set for each fold. As shown in Table 1, the AUCs of all models are above 0.96 while the EERs are below 10%. Then, the cross-user model is trained with multiple users’ motion data. We adopt the following protocol: for each participant, we train a model with 11 other participants’ data and then test it on the chosen participant’s data. Table 2 shows that the cross-user model can still provide a high AUC when it is applied to a new user, where the worst AUC is 0.90 and the worst EER is 16%. We can observe an increase in EER by comparing cross-user models to per-user models, which implies a weaker ability to distinguish a new user’s handover gesture from their other movements. Nevertheless, we apply a sliding window-based strategy for sharing event detection (see § 5.2.2) and additionally apply IA based owner detection (see § 5.3) to further mitigate false detection. In summary, the results shows that the gesture detection model can recognize handover gestures across different users, which imply its low user dependence.

In the appendix, we present cross-device evaluation results to show model transferability to different devices.

5.2.2 Impact of different settings

As introduced in § 4.2, the choice of segment size d and interval p affects the detection accuracy. Besides, we adopt a sliding window-based strategy, where handover detection is performed over w segments to balance accuracy and detection delay. Moreover, we use adaptive sensing to save battery (see § 4.2), and the choice of its activation threshold θ may affect the detection performance. We divide all events into 10 subsets and adopt 10-fold cross validation. We enable adaptive sensing only for the adaptive sensing experiments.

Segment size and interval. Intuitively, a larger segment size d and a smaller interval p provide better ability to cover a sharing gesture. We tested four d ’s (1s, 2s, 3s, 4s) and set p to achieve a 50% segment overlap. We set $w = 2$. In Fig. 5(a), $d = 3s$ provides higher precision compared to $d = 2s$, but it takes longer to make a detection. To balance latency and accuracy, we set $d = 2s$. Then, we test three different p ’s. Fig. 5(b) shows that a shorter interval has higher recall, but lower precision. A larger overlap allows more classifications in the same period to improve recall and capture a gesture earlier.

Window size. Considering the length of a sharing gesture, we change the window size from one to three segments at $d = 2s$ and $p = 1s$. Fig. 5(c) shows f1-score reaches the highest (median: 98%) and the average elapsed time is only 2s when $w = 2$. When $w = 3$, the average recall decreases to 81%. When $w = 1$, the average precision drops to 92%. This result shows the necessity of a window-based strategy to avoid false positives instead of directly using event-level results.

Adaptive sensing. We set up the evaluation environment as follows: 1) low frequency mode: $f_s = 10Hz$ without classification task. 2) mode switch: if $m > \theta$, high frequency mode is activated; if $m \leq 0.1m/s^2$, low frequency mode is activated; there is a 90ms latency when mode switch happens, which is the maximum of 50 measurements on Google Pixel. 3) high frequency mode: $f_s = 50Hz$ with feature extraction and classification. We test five different thresholds: 0, $0.3m/s^2$, $0.6m/s^2$, $0.9m/s^2$, $1.2m/s^2$. Fig. 5(d) shows that recall drops with higher θ . Due to the mode switch delay, it is likely to miss data at the beginning of a gesture. Nevertheless, when $\theta = 0.3m/s^2$, recall is still acceptable (mean: 95%).

Given our results, we use the default settings $d = 2s$, $p = 1s$, $w = 2$, and $\theta = 0.3m/s^2$ to balance precision (mean: 0.98), recall (mean: 0.95), and reaction time (mean: 2.33s).

5.2.3 False positive evaluation

We train the gesture detection model using the training data from all participants in § 5.2.1 and adopt the settings summarized in § 5.2.2. We evaluate the long-term false positive rate of handover gesture detection using the HMOG

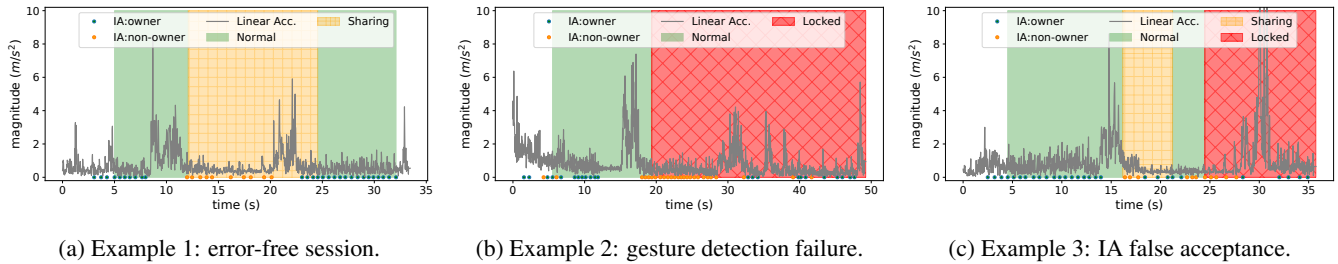


Figure 6: A user study session consists of three stages: 1. owner uses the device and then hands it to sharee; 2. sharee uses the device and then returns it; 3. owner receives the device. The grey plot shows the intensity of the movements measured by the accelerometer. Green area: the device is in state “normal”. Yellow area: the device is in state “sharing”. Red area: the device is locked. Blue and orange points are the per-swipe results of touch-based IA, representing owner and non-owner, respectively.

dataset [40,43]. This testing data involves 493 sessions (about 81 hours) of 100 smartphone users’ reading and writing activities, but no sharing activities, while standing or sitting. For each session, we keep detection running even after a false positive is detected. The result shows that the hourly false positive rate for continuous device use is 0.9 per hour. Since handover detection runs only when the screen is on, the number of false positives is correlated to the daily screen time of a user. For an average daily screen time of three hours [11], handover detection may produce two to three false positives in a day. Even if a false positive makes DSA move to state “sharing” falsely, DSA switches back to state “normal” once the owner’s identity is confirmed, which mitigates the false positive. Thus, the false positive rate of handover detection is acceptable for daily usage. For future work, we will conduct a longitudinal study on the impact of false positives on usability.

5.3 Device sharing processing

The second part of our user study tested if DSA is able to automatically detect the sharing and the returning of a smartphone with the help of both handover detection and owner detection. Besides, we captured potential exceptions.

Task description. We adopted a touchscreen input based IA scheme [10] (i.e., touch-based IA) and used a (m,n) -sliding-window-based strategy: If m out of n swipes are accepted as the owner’s, IA will accept the current user as the owner. Here, we set $m = 4, n = 7$ for balancing false rejection rate and false acceptance rate of touch-based IA. For IA enrollment, we collected 200 swipes from each participant to train the per-participant IA models. For handover detection, we trained a model with the training data from the controlled experiments in § 5.2 and use the default settings. In each session, a group of two participants was asked to perform a web page sharing task: the owner shared a web page and handed the phone to the sharee; after reading the page, the sharee returned the phone to the owner. Each participant was required to swipe at least 10 times during reading when the phone is in their possession. Once they completed the reading task, they swapped their

roles and repeated the above process. Each group contributed to 10 sessions. Given that the amount of time for temporary device sharing is usually limited [31], most sharing events in our study lasted from 30 seconds to one minute. Short sharing events require DSA to detect the starting and end of a sharing event rapidly. We did not specify the position of each participant and how they handed over the device so that participants could hand over the device in their natural manner. In total, we collected 50 device sharing sessions from five groups of participants for analysis.

Results. We counted the sessions with exceptions of either handover detection or IA among the 50 sessions. We observed three exception types: 1) failure in detecting handover gestures to enable the sharing mode: 2 sessions, 2) IA falsely accepting a sharee as an owner: 6 sessions, and 3) failure in detecting the end of a sharing event: 1 session. Therefore, DSA completed the implicit sharing loop in 41 sessions without explicit inputs from the owner. We note that the results were related to the performance of the selected IA scheme, which can be improved by using IA schemes with higher accuracy. Fig. 6(a) shows an example of a session without exceptions: The owner was using the phone during the first 9 seconds and then handed it over to the sharee; DSA detected a handover gesture and switched to state “sharing” at 12s; after the sharee finished using the phone and returned the phone to the owner, the owner detector detected the owner at 24.5s and switched back to the normal state.

Exception processing. We recorded all sessions with exceptions and analyzed how DSA processed them. In two sessions, handover detection failed to detect handover gestures but DSA blocked the sharee according to the negative IA results (see Fig. 6(b)). In six sessions, DSA initially falsely identified the sharee as the owner. However, in four of these sessions, it correctly identified the sharee as non-owner within several seconds after obtaining more touch events from the sharee. For example, in Fig. 6(c), DSA falsely identified the sharee as the owner, and consequently, state “sharing” was left at 21.2s. However, after DSA detected several non-owner touch events, it locked the sharee out at 24.4s to prevent potential unauthorized access. DSA failed to recognize the owner after

the device had been returned in only one session. In this case, the owner could still manually exit from state “sharing” by passing re-authentication. A possible solution to mitigating potential security threats brought by IA false detection is to set up stricter detection criteria for identifying the owner (e.g., requiring more positive swipes in a window size) in state “sharing”. Note that these errors or exceptions may be specific to touch-based IA owner detection. Using or combining different biometrics may improve accuracy. Furthermore, the training data was collected from only brief reading tasks, which lacks diversity and may result in more false detections.

6 Discussion

Battery consumption. We run the DSA service on Google Pixel in airplane-mode for 30 minutes without other running apps and repeat 5 times for both high-frequency sensing and low-frequency sensing. We use Battery Historian to estimate the battery consumption of DSA Service. As a reference, we leave the phone with screen always on, and the phone discharges 3% of battery in 30 minutes. The results show that the average estimated battery consumption of DSA Service alone for high-frequency sensing is 0.11% per half hour; the rate for low-frequency sensing is 0.06% per half hour. Therefore, the battery consumption is very small while adaptive sensing can further reduce battery consumption. Our battery consumption evaluation is preliminary. Since DSA performs sharing detection only when the screen is on, we chose a small time frame. In the future, we will evaluate battery consumption with longer time frames and its impact on device usage.

Defending against unauthorized access. Given the observed latency of handover detection, we conclude that DSA can swiftly activate the sharing mode, and a sharee can hardly conduct effective attacks during such a short interval. Even if handover detection fails, owner detection can block a sharee upon negative IA results. As observed in § 5.3, owner detection was limited by the performance of its IA scheme. False acceptance may temporarily deactivate the sharing mode so that a sharee can move to sensitive apps at this moment. The exception processing of DSA will reject a sharee if the IA result is negative again. However, similar situations may result from a malicious sharee launching specific attacks on the adopted IA scheme (e.g., mimicry attacks [23]). A promising countermeasure is to adopt multiple modalities (i.e., multi-modal IA) so that the failure of one modality is not likely to make owner detection fail. Thus, how to incorporate multi-modal IA into DSA will be our future work.

PIN sharing. For DSA, we assume that a device owner initially holds the device and performs a sharing gesture, indicating a device sharing event. However, PIN sharing, another way of device sharing, breaks the assumption. An owner shares their PIN/password with a sharee in advance so that the sharee can unlock the device without the owner’s pres-

ence. DSA cannot distinguish PIN sharing from unauthorized access since it only captures non-owner access for both cases. However, a device sharing solution can be made aware of PIN sharing through two ways: 1) The shared PIN can reveal a user’s identity. A device owner can set up two different PINs [2] for themselves (i.e., private use) and sharees (i.e., shared use), respectively. If a user is using the PIN for sharees, it implies a sharing event. 2) A sharee can register their biometrics (e.g., fingerprint, face, touch) in the system so that they can be identified. The device sharing solution can activate the sharing mode once the current user’s biometrics match any registered sharee’s record. Otherwise, it identifies the current user as illegitimate and locks the device.

Evaluation limitations. We list the following limitations in the evaluation of DSA: First, the evaluation of handover detection did not cover some conditions (e.g., from standing to sitting, in a vehicle) and edge cases (e.g., non-handover sharing actions via a table). For edge cases, as DSA’s sharing detection is extensible, a feasible solution is to add models for these sharing actions. For better security, sharing mode can be enabled for these cases only if a non-owner is detected under certain contexts (e.g., at home). Second, to collect sufficient device sharing events in a short period, we asked participants to execute tasks in the second part of our user study. Some handovers in the first part of the user study required participants to follow specific position and direction instructions. These may have influenced their device sharing behavior during the tasks in the second part. Third, our analysis focused on how DSA handles sharing a device from a system’s perspective. A potential avenue is to conduct a field study with our prototype DSA implementation so that we can investigate how DSA handles sharing events in the wild and collect smartphone users’ perceptions about DSA.

7 Conclusion

We present DSA, a device sharing awareness solution for temporary smartphone sharing. DSA enables smartphones to conduct continuous and proactive device sharing detection with low latency and low power requirements. It provides flexible access control strategies to protect sensitive apps and resources from unauthorized access during sharing. Extensive experiments show that DSA can detect device sharing with high recall and low false positive rates.

Acknowledgements

This research has been supported by the Waterloo-Huawei Joint Innovation Laboratory. The authors would like to thank the anonymous reviewers and shepherd for providing insightful comments and feedback to improve the paper.

References

- [1] Syed Ishtiaque Ahmed, Md Romael Haque, Jay Chen, and Nicola Dell. Digital privacy challenges with shared mobile phone use in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–20, 2017.
- [2] Syed Ishtiaque Ahmed, Md Romael Haque, Irtaza Haider, Jay Chen, and Nicola Dell. “Everyone has some personal stuff”: Designing to support digital privacy with shared mobile phone use in Bangladesh. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.
- [3] Ahmad Akl, Chen Feng, and Shahrokh Valaee. A novel accelerometer-based gesture recognition system. *IEEE Transactions on Signal Processing*, 59(12):6197–6205, 2011.
- [4] Mahdi Nasrullah Al-Ameen, Huzeyfe Kocabas, Swapnil Nandy, and Tanjina Tamanna. “We, three brothers have always known everything of each other”: A cross-cultural study of sharing digital devices and online accounts. *Proceedings on Privacy Enhancing Technologies*, 2021(4):203–224, 2021.
- [5] Android. Supporting multiple users. <https://source.android.com/devices/tech/admin/multi-user>.
- [6] Daniel Buschek, Fabian Hartmann, Emanuel Von Zezschwitz, Alexander De Luca, and Florian Alt. SnapApp: Reducing authentication overhead with a time-constrained fast unlock option. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3736–3747, 2016.
- [7] Samsung Electronics. S Secure. <https://galaxystore.samsung.com/prepost/000004637448>.
- [8] Mohammed E Fathy, Vishal M Patel, and Rama Chelappa. Face-based active authentication on mobile devices. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1687–1691. IEEE, 2015.
- [9] Davide Figo, Pedro C Diniz, Diogo R Ferreira, and Joao MP Cardoso. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, 14(7):645–662, 2010.
- [10] Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE transactions on information forensics and security*, 8(1):136–148, 2012.
- [11] The Guardian. Shock! Horror! Do you know how much time you spend on your phone? <https://www.theguardian.com/lifeandstyle/2019/aug/21/cellphone-screen-time-average-habits>.
- [12] Alina Hang, Emanuel Von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. Too much information! User attitudes towards smartphone sharing. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pages 284–287, 2012.
- [13] Eiji Hayashi, Oriana Riva, Karin Strauss, AJ Bernheim Brush, and Stuart Schechter. Goldilocks and the two mobile devices: Going beyond all-or-nothing access to a device’s applications. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pages 1–11, 2012.
- [14] Ltd. Huawei Device Co. Create a privatespace for your private data. <https://consumer.huawei.com/en/support/content/en-us15834600/>.
- [15] Apple Inc. Guided access. <https://support.apple.com/en-us/HT202612>.
- [16] Google Inc. Android pin & unpin screens. <https://support.google.com/android/answer/9455138?hl=en>.
- [17] Google Inc. Manage guests and users. https://support.google.com/nexus/topic/6126546?hl=en&ref_topic=3416294.
- [18] Xiaomi Inc. App vault. <https://play.google.com/store/apps/details?id=com.mi.android.globalminusscreen>.
- [19] Markus Jakobsson, Elaine Shi, Philippe Golle, and Richard Chow. Implicit authentication for mobile devices. In *Proceedings of the 4th USENIX conference on Hot topics in security*, pages 9–9, 2009.
- [20] Amy K Karlson, AJ Bernheim Brush, and Stuart Schechter. Can I borrow your phone? Understanding concerns when sharing mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1647–1650, 2009.
- [21] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, 52(4):1–36, 2019.
- [22] Hassan Khan, Aaron Atwater, and Urs Hengartner. Itus: An implicit authentication framework for Android. In *Proceedings of the 20th annual international conference*

- on *Mobile computing and networking*, pages 507–518, 2014.
- [23] Hassan Khan, Urs Hengartner, and Daniel Vogel. Augmented reality-based mimicry attacks on behaviour-based smartphone authentication. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 41–53, 2018.
- [24] Masoud Mehrabi Koushki, Yue Huang, Julia Rubin, and Konstantin Beznosov. Neither access nor control: A longitudinal investigation of the efficacy of user Access-Control solutions on smartphones. In *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA, August 2022. USENIX Association.
- [25] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.
- [26] Norton Labs. Norton app lock. <https://play.google.com/store/apps/details?id=com.symantec.applock>.
- [27] Yunxin Liu, Ahmad Rahmati, Yuanhe Huang, Hyukjae Jang, Lin Zhong, Yongguang Zhang, and Shensheng Zhang. xShare: Supporting impromptu sharing of mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 15–28, 2009.
- [28] Zhiyuan Lu, Xiang Chen, Qiang Li, Xu Zhang, and Ping Zhou. A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. *IEEE transactions on human-machine systems*, 44(2):293–299, 2014.
- [29] Upal Mahbub, Vishal M Patel, Deepak Chandra, Brandon Barbelo, and Rama Chellappa. Partial face detection for continuous authentication. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2991–2995. IEEE, 2016.
- [30] Diogo Marques, Tiago Guerreiro, Luís Carriço, Ivan Beschastnikh, and Konstantin Beznosov. Vulnerability & blame: Making sense of unauthorized access to smartphones. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.
- [31] Tara Matthews, Kerwell Liao, Anna Turner, Marianne Berkovich, Robert Reeder, and Sunny Consolvo. “She’ll just grab any device that’s close”: A study of everyday device & account sharing in households. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5921–5932, 2016.
- [32] Michelle L Mazurek, JP Arsenault, Joanna Bresee, Nitin Gupta, Iulia Ion, Christina Johns, Daniel Lee, Yuan Liang, Jenny Olsen, Brandon Salmon, et al. Access control for home data sharing: Attitudes, needs and practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 645–654, 2010.
- [33] Xudong Ni, Zhimin Yang, Xiaole Bai, Adam C Champion, and Dong Xuan. DiffUser: Differentiated user access control on smartphones. In *2009 IEEE 6th International Conference on Mobile Adhoc and Sensor Systems*, pages 1012–1017. IEEE, 2009.
- [34] Saumay Pushp, Yunxin Liu, Mengwei Xu, Changyoung Koh, and Junehwa Song. PrivacyShield: A mobile system for supporting subtle just-in-time privacy provisioning through off-screen-based touch gestures. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–38, 2018.
- [35] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. “Privacy is not for me, it’s for those rich women”: Performative privacy practices on mobile phones by women in south asia. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 127–142, 2018.
- [36] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor’s new security indicators. In *IEEE Symposium on Security and Privacy*, pages 51–65. IEEE, 2007.
- [37] Julian Seifert, Alexander De Luca, Bettina Conradi, and Heinrich Hussmann. TreasurePhone: Context-sensitive user data protection on mobile phones. In *International Conference on Pervasive Computing*, pages 130–137. Springer, 2010.
- [38] Teddy Seyed, Xing-Dong Yang, and Daniel Vogel. A modular smartphone for lending. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 205–215, 2017.
- [39] Sheng Shen, He Wang, and Romit Roy Choudhury. I am a smartwatch and I can track my user’s arm. In *Proceedings of the 14th annual international conference on Mobile systems, applications, and services*, pages 85–96, 2016.
- [40] Zdeňka Sitová, Jaroslav Šeděnka, Qing Yang, Ge Peng, Gang Zhou, Paolo Gasti, and Kiran S Balagani. HMOG: New behavioral biometric features for continuous authentication of smartphone users. *IEEE Transactions on Information Forensics and Security*, 11(5):877–892, 2015.

- [41] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing*, 16(4):62–74, 2017.
- [42] Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, and Nadir Weibel. ExtraSensory app: Data collection in-the-wild with rich user interface to self-report behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.
- [43] Qing Yang, Ge Peng, David T Nguyen, Xin Qi, Gang Zhou, Zdeňka Sitová, Paolo Gasti, and Kiran S Balagani. A multimodal data set for evaluating continuous authentication performance in smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, pages 358–359, 2014.
- [44] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1080–1091, 2016.

A Appendix

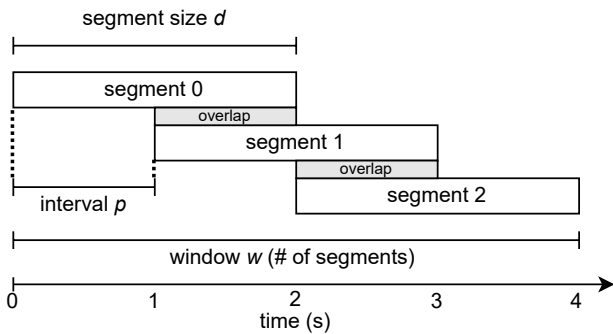
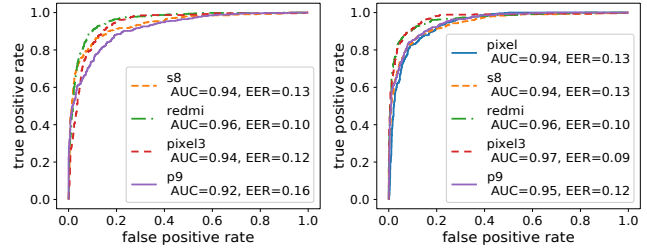


Figure 7: Data segmentation. In this example, segment size $d = 2s$, interval $p = 1s$, and windows size $w = 3$ segments.



(a) One-to-multi test (Pixel) (b) Multi-to-one test

Figure 8: Inter-device experiment.

Cross-device experiments. The gesture detection model is also expected to work across different phone models. In the cross-device experiments, we added four other Android phone models: Samsung S8, Redmi 5, Google Pixel 3, and Huawei P9 and collected motion data of two participants for each phone model. We adopt the following two protocols to test cross-device accuracy: (1) We train a model with all 12 participants’ training data on the Google Pixel and test it on the other four phones. As shown in Fig. 8(a), the model trained with one phone’s data shows a consistently good performance on the other phones, where the AUCs are always above 0.92 and the EERs are around 10 to 16%. (2) We use mixed training data from four phone models to train the model and test it on the fifth phone. As shown in Fig. 8(b), the cross-device model provides a consistently good performance across different phone models.

Balancing Power Dynamics in Smart Homes: Nannies’ Perspectives on How Cameras Reflect and Affect Relationships

Julia Bernd

International Computer Science Institute

Junghyun Choy

International Computer Science Institute

Ruba Abu-Salma

King’s College London

Alisa Frik

International Computer Science Institute

Abstract

Smart home cameras raise privacy concerns in part because they frequently collect data not only about the primary users who deployed them but also other parties—who may be targets of intentional surveillance or incidental bystanders. Domestic employees working in smart homes must navigate a complex situation that blends privacy and social norms for homes, workplaces, and caregiving. This paper presents findings from 25 semi-structured interviews with domestic childcare workers in the U.S. about smart home cameras, focusing on how privacy considerations interact with the dynamics of their employer–employee relationships. We show how participants’ views on camera data collection, and their desire and ability to set conditions on data use and sharing, were affected by power differentials and norms about who should control information flows in a given context. Participants’ attitudes about employers’ cameras often hinged on how employers used the data; whether participants viewed camera use as likely to reinforce negative tendencies in the employer–employee relationship; and how camera use and disclosure might reflect existing relationship tendencies. We also suggest technical and social interventions to mitigate the adverse effects of power imbalances on domestic employees’ privacy and individual agency.

1 Introduction

Privacy choices that individuals make regarding their own connected devices often affect the privacy of those around them. These knock-on privacy effects are becoming rapidly more urgent with the expanding use of connected smart devices—

many of which are designed to collect data in what was formerly a prototypical private place, the home. In addition to collecting data about the primary user who installed it, a smart home device may also collect data about other members of the household, incidental visitors or bystanders, and potentially targets of deliberate surveillance within the home. These secondary “users” may have more or less power to control what data the devices collect about them, depending on their social and economic position relative to the primary user(s).

In this paper, we examine how social and economic power dynamics affect the privacy consequences of smart home cameras for domestic childcare workers. Childcare workers such as nannies, au pairs, and professional babysitters may sometimes be incidental bystanders to data collection, as a result of increasing use of smart home technology in general [118, 125]. And they may sometimes be deliberate targets of monitoring by their employers—a practice that is becoming more expected, at least in some places [35, 43, 57, 127].¹ This case study of nannies contributes to a growing body of work on how socio-economic power differentials may result in differential privacy outcomes for different types of people.

In this research—the first on smart home privacy for domestic workers in the U.S.—we focus on nannies because they operate in a complex, multi-layered context that blends disparate sets of potentially conflicting norms and priorities about data collection and sharing [cf. 4, 17].² In addition to being a home, where the residents tend to have more control over decisions about technology (or whatever else) than they do elsewhere, the smart home is also the nanny’s workplace. This may imply a different set of data norms, and control over any aspect of the environment is also mediated by the employer–employee relationship and its power dynamics. Finally, it is

¹In this paper, we refer to domestic childcare workers as “nannies”, and the job as “nannying”, for the sake of brevity. But our study included au pairs and professional babysitters as well, and findings are based on all participants.

²Brief preliminary findings based on interviewers’ impressions were previously published as a work-in-progress workshop paper [17]. This paper is the first full publication based on systematic analysis of the transcripts.

a care situation, which complicates the usual professional divides—and at the same time may imply a different balance between the employer’s safety concerns and the employee’s privacy, compared to other workplaces.

Within this blended context, this paper focuses on the effects of cameras (as opposed to other devices), because they have unique implications in terms of power dynamics and employer–employee relationships. Employers may use camera data in a way that affects nannies’ day-to-day experiences of their job, as well as their job security [cf. 35, 53, 122], while devices such as smart speakers or smart thermostats rarely have such uses.³

This paper addresses the following research questions:

1. How are domestic childcare workers’ privacy attitudes, experiences, expectations, concerns, and choices with regard to working with smart home cameras shaped by their relationships with their employers?
2. How do employers’ use of and interactions with employees about cameras reflect, reinforce, or change existing power dynamics in those relationships?
3. What are potential points of intervention (social and technical) for mitigating the effects of power imbalances on how domestic childcare workers’ privacy preferences are enacted with regard to smart home cameras?

Based on a qualitative analysis of 25 interviews with nannies, au pairs, and professional babysitters in the U.S., we show how privacy attitudes and expectations in a domestic childcare context may be affected by power differentials and norms about who makes decisions about data flows in a given context, as well as childcare workers’ specific concerns about how their employers may use the data. We also show how nannies’ ability to exert control over their data is limited by both social norms and practical economic considerations. In the process, we identify potential points of intervention to mitigate the privacy effects of power differentials. We suggest corresponding solutions that focus on promoting and improving communication between employers and employees about camera use, supporting technical and social designs that are not only privacy-enhancing but also agency-enhancing.

2 Related Work

User and Bystander Privacy in Smart Homes Researchers have explored primary users’ experiences, perspectives, expectations, and privacy concerns with regard to smart home devices’ data practices [e.g. 1, 14, 23, 51, 93, 128, 138, 139, 140, 147] [overview in 72]. Many studies focus on how

³We compare nannies’ experiences, views, and adversarial models with regard to cameras vs. other smart home devices in a forthcoming paper.

particular situational factors shape people’s attitudes and concerns with regard to data collection, use, and sharing [e.g. 7, 38, 44, 67, 68, 81, 82, 94]. Most studies that compared locales have found that people are more sensitive about devices gathering data in their homes than, for example, in their workplaces or in business establishments [e.g. 4, 27, 67, 94, 112] [contra 45]. Consequently, several questions about privacy preferences and concerns arise when one person’s workplace is also another person’s home.

Researchers have used the Theory of Privacy as Contextual Integrity (CI) [15, 96, 97, 98] to examine how people reason about data collection that blurs or crosses the boundary between private and public contexts (e.g. home vs. Internet, [e.g. 7, 77, 141]). From the intersection of these contexts, new ideas about privacy and power can emerge [20]. CI asserts that it is the *context*, or particular social situation, that dictates norms about digital privacy and acceptable data sharing. Information redistribution that is considered appropriate in one situation may be too sensitive or a violation of privacy in another. If there are power imbalances, CI analysis can also uncover how the parties in a given context negotiate conflicting norms [e.g. 16, 55, 56, 63, 64].

Most research on smart home privacy preferences and expectations has focused on people’s views as primary users—including their concerns about bystanders [e.g. 52, 145, 146, 147]. However, secondary users of various kinds have recently received more research attention. Research on multi-user smart homes shows the complexity of balancing differing privacy preferences of household members [e.g. 6, 40, 42, 51, 60, 62, 144, 148]. In some cases, research on residents of smart homes has also noted potential issues for non-residents [e.g. 23, 26, 65, 81, 82, 106, 117, 128].

Other work has focused more closely on visitors and guests as bystanders, including Airbnb guests [29, 84, 123]. Situations in which people become smart home bystanders are very common, and span a variety of social and employment contexts [25, 85, 91]. Even if bystanders know devices are present, they often have incomplete or incorrect ideas about the extent of data collection and use [3, 4, 85, 86], and they may not have socially appropriate ways to express their privacy preferences even when they understand the implications [3, 4, 53, 86, 144]. Some of these studies suggest technical and/or social solutions to these issues, some of which will be described later in this section.

Privacy and Social and Economic Power Discussions about technology and social and economic power rest on a substantial body of work on the digital divide. Research on digital inequalities [e.g. 22, 39, 109, 114, 136, 150] [for IoT: 73], and on demographic differences in online privacy knowledge, behaviors, and attitudes [e.g. 30, 49, 80, 100, 132, 133, 134], has shown how vulnerabilities arise from such differences [e.g.

24, 46, 49, 100] [for IoT: 10, 47, 103]. Existing power imbalances between those who collect data and those data is collected from mean the disadvantaged tend to have less control over their privacy [e.g. 5, 19, 24, 33, 78, 83, 111, 115, 134]. Collected data can then lead to further discrimination [e.g. 79] [for IoT: 31, 103].

Power imbalances and accompanying privacy vulnerabilities can play out in employer-employee relationships within surveilled workplaces, creating complex trade-offs [11, 69, 137] [for IoT: 9, 75, 78, 99] [for care contexts: 18, 71, 124]. These trade-offs are especially prominent with in-home work [4, 53]. For example, where parents use nanny cams as a means of control [35], it can result in evasion [4, 53], and may also reduce nannies' capacity to deliver the best care [53]. We compare findings from our study with other research on domestic workers in smart homes in §6.2; however, there are no published studies from the U.S. Additionally, our study adds depth in aiming to understand the relationship between smart home privacy and employer-employee power dynamics.

Smart devices can also affect family or household dynamics in multi-user smart homes. For example, Apthorpe et al. [6] found that IoT devices benefit interpersonal relationships (e.g. easing household management) but also cause interpersonal conflicts (e.g. facilitating surveillance, causing distrust, causing disagreements over device use). On the other hand, the question of exactly who has control over devices in the home may be an indicator of existing interpersonal and/or socio-cultural dynamics [42, 53, 61, 63, 64]. In extreme scenarios, imbalances in device control can enable domestic abuse [36, 53, 70, 76, 101, 116]. If there are children involved, smart home devices [13, 66, 113, 126] and smart toys [8, 90] can turn children into targets of or bystanders to data collection.

Protections for Bystander Privacy Proposals for or attempts to implement stronger bystander protections have focused on detecting hidden cameras [e.g. 21, 74, 102, 123, 131] or clearly signaling that a device is recording or transmitting data [e.g. 3, 4, 23, 25, 32, 54, 59, 85, 86, 105, 108, 128, 129, 143, 147]. Others have proposed using objects [e.g. 2, 119] or contextual cues (such as people's locations, presence of multiple people in a room) [e.g. 12, 51, 65, 92, 95, 104] to signal preference to obfuscate or not record data about bystanders at all. However, there are limitations to such technical approaches, and not all users trust manufacturers or service providers to implement them [58, 65, 149].

There have been calls for more granular smart device settings to accommodate the privacy interests of different parties in the same household [6, 40, 42, 48, 50, 50, 51, 63, 65, 91, 105, 130, 135, 143, 144, 148]. For instance, parents have expressed interest in nuanced parental controls that would allow children to use devices without compromising safety

[6, 126]. Other potential design practices for increased bystander protections—which could especially benefit domestic employees, if extended—include simplifying the privacy control process [148] and expanding the platforms through which smart controls can be adjusted [3, 42, 60, 86, 143, 148]. Besides technical implementations, other suggestions include raising awareness of smart device function [4] and facilitating more open and transparent conversations about device usage between primary users and bystanders [25, 128, 144, 148].

Such recommendations, however, were not made with specifically domestic workers in mind. We expand upon these recommendations, and suggest our own interventions to address the privacy concerns of domestic workers and power imbalances with their employers, in §6.3.

3 Methodology

We designed and conducted semi-structured interviews with 25 domestic childcare workers (including nannies, babysitters, and au pairs) in the U.S. in late 2019.⁴ The Institutional Review Board (IRB) at University of California, Berkeley, reviewed and approved our study, and we obtained written consent from participants.

Data Collection We used a mix of offline and online recruitment. We distributed flyers in cafes, daycares, schools, colleges, and playgrounds. We also advertised in nanny-specific Facebook groups, Reddit communities, and other online venues, and used snowball sampling. We recruited both individuals who had worked with cameras and those who had not.

Our interviews included warm-up questions about participants' nannying experiences and relationships with their current and past employers. We next asked about participants' personal experiences working in houses that had cameras, and discussions with their employers about the cameras. Other questions included participants' expectations, attitudes, privacy concerns, and choices they had (or would have) made related to camera use and disclosure, as well as their knowledge of legal and technical protections. When participants had not had specific experiences, such as working with cameras or finding hidden cameras, we probed hypotheticals to explore their views. We also administered an exit questionnaire covering demographics, experiences with technology, current employment situation, and what smart devices they and their employers owned.⁵

⁴In parallel, we also conducted 15 interviews with parents who employ nannies. Information on those interviews may be found in Bernd et al. [17], and results will be published in a future paper.

⁵Our recruitment materials, screening scripts, interview scripts, and exit questionnaires can be found at <https://bit.ly/3zIEpov>, so that other researchers can use them in related work (as some already have).

We conducted one pilot to finalize and prioritize questions. Interviews took 1 to 1 1/2 hours. Two were in person and the rest were by phone or video chat. We compensated participants \$50. Interviews were professionally transcribed.

Data Analysis We used inductive coding to identify common topics and themes in our qualitative data. To develop an initial coding frame, three researchers each independently coded a separate test transcript using MAXQDA. The three researchers then discussed their coding frames and merged them after resolving disagreements. These researchers and an additional researcher then independently coded two more transcripts (the same ones) to test the merged frame. After making further changes to the coding frame, we divided up all the interviews (including the test transcripts) so that each transcript was coded by two researchers, continuing to check agreement and discuss questions about code application throughout.⁶ No further codes were identified, indicating saturation.⁷ All four researchers participated in organizing codes into themes specific to addressing our research questions. We all reviewed the excerpts on each topic to further refine the themes.

Limitations Study materials and interviews were in English. All participants were comfortable conversing in English, but some did not speak it as their first language; 20% spoke another language primarily or equally. This may have increased the possibility of misunderstanding. We may also have missed insights about the effects of limited English fluency on communication and power imbalances for U.S. nannies that we could have captured by offering interviews in other languages.

Around 28% of domestic childcare workers in the U.S. are immigrants [142], and we believe that understanding the experiences of immigrant workers is key to understanding power imbalances and privacy in the workplace [cf. 46]. We do not know whether our sample was representative in this regard, as we did not ask about immigration status. (We did not believe the limited scientific benefit would outweigh distress undocumented participants could have experienced at the question.) Immigrant workers may be less likely to take part in studies, due to language barriers or—especially for those with precarious immigration status—enhanced privacy concerns [e.g. 28, 107]. As we describe in §6.4, this qualitative work should be expanded and quantified, increasing generalizability with multiple languages and focused recruitment of immigrants.

⁶We checked agreement rates to ensure coders were using codes similarly. We do not report formal agreement measures as we do not make quantitative claims [cf. 89]. Rather, we aim for transparency and thick description [41].

⁷A text copy of the codebook can be found at <https://bit.ly/3aEDRoU>, or as a MAXQDA file upon request, again so it can be used in future comparative research.

4 Participants

Demographics and Job Experience With the potential exception of immigration status, our sample is representative of the demographics of nannies and childcare workers in the U.S. Participants' ages ranged from 19 to 55, with a median of 30. All participants self-identified as female/women. Asked to self-describe ethnicity, 72% said white or Caucasian, 16% Hispanic, Latina/x, or Mexican, and 8% Asian or Indian-from-India. Most participants were either nannies (60%) or nannies/household managers (16%), while 12% were professional babysitters and 8% were au pairs. 4% had other similar nanny jobs. Additional details about participant demographics (and comparisons with the target population) may be found in Bernd et al. [17], Appendix A.

Experience With Cameras 22 out of 25 participants had worked with indoor cameras on while they were present. All who had worked with cameras had encountered livestreaming ones, usually Internet-enabled. Most had worked with cameras that recorded (often simultaneously streaming), though some recorded only when triggered. (The other participants were not sure about recording.) Cameras were most commonly located in children's bedrooms/playrooms and entryways, and some in common areas.

5 Findings

In this section, we describe how smart home cameras intersect with the employer–employee relationship and its power dynamics in an in-home childcare context. In §5.1, we discuss participants' privacy attitudes and concerns about how cameras *reflect* the employer–employee relationship—i.e. what camera use or disclosure may indicate about employers' attitudes toward the nannies or the operation of power dynamics in the relationship (RQ1). At the same time, many participants were concerned about how cameras could *affect* the employer–employee relationship, e.g. by reinforcing power dynamics or encouraging parents to be critical of nannies (RQ2); we discuss these perceptions in §5.2. However, as we discuss in §5.3, potentially-conflicting contextual norms about control of information flows constrained participants' ability and desire to make choices about their collected camera data (RQ1, RQ2). In §5.4, we show how participants' choices with regard to accepting or restricting data collection (including whether to accept a job at all) could be either motivated or constrained by the power differentials in those relationships (RQ1, RQ2). Throughout our analysis, we saw that purposes of data collection and how data was used were central themes—in particular, whether the purpose of the camera was related to the nanny's employment and job prospects. Some of the views discussed in this section implicate potential *points of intervention*, where intervening could have substantive effects on nannies' experiences with cameras (RQ3).

5.1 Cameras Reflecting Employer–Employee Relationships

With a well-matched family, participants said they could build a good working relationship with the parents, and strongly bond with children: “Both for me and for the family, we both have to trust each other. And that’s not as important in a lot of other positions.” (N20) Participants considered mutual trust, respect, and open and honest communication as essential components in building a common ground and effectively resolving disagreements. Our participants often expressed opinions about how cameras, or interactions around cameras, might illuminate or *reflect* those important relationship values.

5.1.1 Cameras and Trust

What cameras signaled about trust was a frequent theme. Many participants viewed cameras as at least a *potential* sign that parents might not trust them:⁸ “I think [the cameras] made me a little bit more cautious about if they trust me or not.” (N36)⁹ (Participants occasionally mentioned specific ways parents might not trust them, such as not trusting them to be attentive, but usually phrased it more generally.) However, participants had more nuanced perspectives on the likelihood that a camera signaled distrust and how uncomfortable they were with it, depending on the specific context.

Some participants viewed cameras fairly broadly as a sign of outright mistrust, to a degree that always made them uncomfortable: “It’s just a feeling, like, that you’re not trusted and you’re being watched.” (N33) Some were concerned about lack of trust when the camera’s purpose was specifically to monitor or even micromanage the participant: “[*Interviewer: If the main reason was actually to check in on you, how would you feel about that?*] [...] I would be uncomfortable. Cause again, I think the whole circumstances of the nanny and the family is the trust and the communication.” (N37)

Different participants also noted that their judgment about whether cameras indicated a lack of trust might depend on how often employers checked the cameras or how many cameras were in the house: “I wouldn’t be comfortable working in a home where there are cameras in all of [...] the communal spaces, like the living room and the kitchen [...]. And if those cameras are being monitored constantly. I just wouldn’t be able to relax ever and I wouldn’t feel trusted at all.” (N6) In particular, some participants did not view it as a serious trust problem even if the employers said the cameras were there to monitor the nanny, as long as it really seemed to be a just-in-case protection rather than constant monitoring: “I

⁸As this research is qualitative, we did not try to count and verify the number of participants who expressed a given view. We use words like *most*, *many*, *some*, or *a few* only to give a rough insight into prevalence.

⁹Numbers run higher than 25 because we assigned them when contacting potential participants to set up interviews; some did not follow through, or canceled.

don’t have a problem with there being observational devices. [...] But if they felt the need to monitor me 24/7, [...] I would be uncomfortable with that, because that shows us a level of distrust that would make me probably leave and go find another position.” (N14)

A few opined that a trust gap at the beginning was not unexpected, but that trust should build over time and be reflected in decreasing camera use: “[An employer should] maybe consider using it only as the trust is being built. And then, once [...] she realize[s] that she can trust this person, then to stop using the cameras.” (N19) However, some worried that employers who frequently used cameras might end up relying too much on them: “I would see something parents need to avoid would be to use [cameras] to build trust as opposed to actually building trust with the person.” (N14)

5.1.2 Disclosure and Respect

Mentions of trust *per se* did not necessarily explicitly relate to power dynamics (though such dynamics might be implicit, in who had power to entrust whom with what). However, trust was often discussed together with other relationship aspects with more clear power implications, such as respect: “[If] a camera just shows up all of a sudden one day without any discussion, that’s not gonna make me feel very trusted and like they respect my profession.” (N27)

With both trust and respect, participants had a range of views on whether having cameras in itself was a bad sign—and some noted the meaning was changing as smart homes became more popular: “Earlier in my career, it was very odd to see a camera in a house, and it meant that a family didn’t trust you. But now it’s become so much more commonplace.” (N27) However, opinions of *undisclosed* cameras tended to be more negative—and in particular, not disclosing cameras to a nanny was seen by many participants as a bad sign in terms of relationship values like trust, respect, and good communication: “That would feel better [if employers had disclosed their cameras] because [...] you can see a good communication between you and them.” (N15) An undisclosed camera could make someone feel untrusted even when they believed the camera was not there to watch them: “They never told me that they have [cameras]. [...] I understand that they don’t have them for me, they have them for the children, they’re like, you know, cameras to watch the kids, so... But I had that feeling of feeling, like, not trusted.” (N36)

In particular, several participants drew a strong connection between disclosure and respecting or valuing the nanny: “I have no problem being recorded as long as you’re telling me you’re doing it. You know, as long as there’s some respect for privacy. [...] Respecting you enough to let you know.” (N10)

A few participants also viewed parents’ sharing of data with others through the lens of respect: “[I: Do you know if either of

[your employers] have ever shared any recordings or showed anybody what was happening?] I don't know but I would be shocked if they had. [I: Why would you be shocked?] Because they respect me." (N20) Others were less concerned about sharing generally, though they might draw inferences about employers' attitudes if it was done without consent: "If they've already shared [camera footage] without my consent, I would kind of assume that they are finding an issue with my work in some way." (N7)

5.2 Cameras Affecting Employer–Employee Relationships

In addition to being an indicator of relationship tendencies, cameras can also actively *affect* relationships. Nannies' expectations or concerns about how cameras might change their relationships with their employers—positively, negatively, or not at all—could, in turn, affect participants' privacy attitudes about cameras: whether they were comfortable, uncomfortable, or simply resigned.

5.2.1 Uses, Power Implications, and Discomfort With Cameras

Participants' attitudes about cameras (including attitudes about audio vs. video) often depended on the employer's camera usage. Such attitudes were often entangled with how camera usage and purpose impacted power dynamics and relationship quality. Even nannies who might generally be comfortable with cameras in an employer's home might be less comfortable if they thought the devices could facilitate or exacerbate poor treatment or intrusive supervision: "If the purpose is to babysit me while I'm nanning the children, then I really feel uncomfortable with that." (N37) However, there was notable variation amongst participants in what they considered intrusive supervision.

Catching the Nanny Out Cameras could reinforce existing power imbalances by giving employers new evidence to excuse firing a nanny over small infractions: "I [got] fired over the cameras last summer, or that was their official excuse. Because they denied unemployment cause they said I got fired 'for cause' instead of, 'she's not Christian' or whatever. [...] Other ladies have also had [...] personality conflicts [...], and then all of a sudden there's something on the camera that they do, because the parents are watching [...] for the first wrong thing that's a little bit out of line." (N27) Even participants who had not been fired could be concerned they might be: "If I curse in front of an audio system, even if I'm not with their child, I could get fired." (N3) Concerns about employers trying to catch nannies out were amplified when cameras were not disclosed: "'We have hidden this camera because we believe that you will be lying to us,' is the message that I get when I see or suspect a hidden camera." (N29)

Cameras could also make nannies nervous about their perceived job performance, or make them feel as if they had to perform for an audience [cf. 128]: "I could see [the camera] turn, it would make me feel extremely uncomfortable. [...] It would almost feel like I'm putting on a show." (N16)

Micromanaging Many participants mentioned that nanny-ing provided them with more autonomy and flexibility than other jobs; however, this benefit could be undermined by excessive supervision. How cameras enabled or even encouraged micromanaging was a major concern for many participants. Some participants explicitly discussed the power implications of micromanaging: "I'm a grown adult. They don't have their boss sitting at their desk watching them do the minutia of their day. I deserve to be treated with the same amount of respect." (N3)

Even when power was not mentioned, the term *micromanaging* evoked the employer's ability to exert additional and unwelcome control over the employee: "So long as they're not using the cameras to micromanage. I've had friends who get [...] little messages throughout the day to show that the parents are watching and criticizing their work." (N29)

Participants were especially uncomfortable with micromanaging via camera when it was used to enforce completion of tasks unrelated to childcare: "If I had just done anything [with their child] that they didn't like, that would be okay [for employers to talk about something they saw on-camera]. But if it was something really nitpicky or if it was something like, 'Oh, I saw that when our daughter was napping, you were on your phone. Can you clean the kitchen next time?' that would be something [...] I would take more offense to." (N26)

"Spying" or Illegitimate Use The connection between camera purposes and power also played out in discussions about employers using cameras for "spying", a term that had different meanings for different participants. In addition to using that term to refer to undisclosed cameras generally, a few referred to it as "spying" when employers observed them when they were not directly with the kids. Such nannies were comfortable with being monitored only while they were with the kids (because then they did not view it as spying on them): "When they're doing it to spy, then I'm less comfortable about it. [But], if it's centered around the kids, I'll accept most explanations." (N7)

Some participants even viewed it as "spying" if they were monitored via camera while they were with the kids: "When you feel like you're being observed by camera, that's different. That's an invasion. [...] If you're watching your nanny do something and then you text her [...] [about something you saw], that's different than if you notice something in the house. That feels like you're being spied on." (N12)

Abuse In addition to concerns about how cameras could negatively impact supervision practices, some nannies were concerned that cameras provided the means for intimidating or creepy behavior: “If they were watching me [when] I wasn’t even with the child, I probably would leave the job. [...] When you’re in someone’s house, it’s their territory. And when they make that unsafe [...] I just wouldn’t feel comfortable in their house again.” (N26)

Risks could be higher for employees who belonged to more vulnerable populations or who knew less about the technology (discussed further in §6.2): “A lot of nannies are older and they might not even understand what some of this technology is, and how it’s used. [...] A lot of domestic workers don’t speak English very well. A lot of domestic workers are from different countries. So, there’s a lot of potential for vulnerable populations to be taken advantage of using this technology.” (N27)

5.2.2 Disclosure, (Dis)Trust, and Mitigating Discomfort

One participant pointed out that hidden cameras in particular aroused suspicions of harassment, whatever the actual reason for nondisclosure might be, and connected it to the equation of disclosure with respect: “I don’t want to be giving anybody a private show by accident and not know. [...] I don’t think it’s respectful to have a camera and hide it. [...] Like it just feels creepy.” (N16)

Several participants highlighted how undisclosed cameras—or undisclosed uses of disclosed cameras—could erode the participant’s trust in parents: “I would feel like [a camera is] a violation of my trust and my privacy if I don’t know about it.” (N20) On the flip side, some participants noted that disclosing cameras and their purposes could facilitate nannies’ trust in their employers’ intentions and attitudes, and thus reduce their discomfort with cameras: “If they give me a good explanation [...] I am generally okay with that. It’s the hiding of it, and then the spying, and the saying that it’s all just so they can look at the kids, when [...] they would not have those cameras if there was not a nanny present. [...] There needs to be two-way communication, so that I feel trust, so that I can provide good care while still feeling watched.” (N7)

However, a few participants preferred their employer *not* disclose, to avoid the discomfort of feeling watched: “I don’t want to know [whether it captures audio], because I don’t want to be self-conscious. I want to do my job without thought of the camera.” (N12)

A couple of participants also mentioned that discomfort or concerns about potential problematic use of cameras could be averted if data were not retained indefinitely: “It just seems like there’s less potential for abuse or misusing a camera if you can’t [...] save tons and tons of video.” (N18)

5.2.3 Power and Comfort With Cameras

A few participants expressed positive views of cameras based on their benefits to relationships, such as how cameras might support good communication and employers’ respect for them as professionals: “What I try to do when there are cameras around is to model for parents how I would handle situations. So, if [the cameras are recording audio] the parents can hear what I’m saying to their child, [...] that’s all the better. [...] And it’s also a way of making sure that we’re on the same page.” (N32)

Relationship benefits like increasing trust might be traded off against potential sources of discomfort: “[Having cameras] gives them the sense [...] that I am who I am with their kids, and who I said I was at the interview. And that’s why I kind of don’t mind the cameras, in a big sense? [...] [Even though] you’re conscious of how you look and [...] all these little things, [...] they don’t really bother me.” (N12)

Another mentioned benefit was that if something went wrong where the nanny was not at fault, cameras could provide exculpatory evidence: “If [the child] runs and falls and smacks her head and gets a bruise, there’s now proof on camera that I’m not the one who caused that to happen. [...] So I definitely prefer working with cameras.” (N20) Further, cameras could mitigate some of the negative consequences of a difficult dynamic, by providing evidence when the nanny would have no other recourse against an employer looking for an excuse to reprimand or fire them: “That way, they can’t say, he said, she said. It’s on the footage.” (N40)

A few nannies expressed comfort with cameras not because of work relationship benefits, but because they had not experienced negative relationship effects with their current employer: “The cameras I feel are perfectly comfortable, within the context of how they’re being utilized and the specific family that I work with.” (N4)

5.2.4 Social Factors in Privacy Resignation

Often participants were resigned to camera data collection because it accorded with the norms for employee–employer or caregiver–child relationships; we describe these norms in §5.3.

A couple of nannies pointed out that cameras could reinforce a general dynamic they were already resigned to, where being in someone else’s home compromised their privacy: “For the most part, there’s no breaks. So, there’s no privacy. [...] Last month, my aunt passed away [...] and at an office job, I might have taken the day off to maintain some privacy. [...] And I feel like I’m overstimulated by kids clinging onto me, no privacy. And then when you add cameras in the home, there’s no privacy.” (N7)

5.3 Prerogatives and Privacy Expectations

Our participants often phrased their expectations about cameras in terms of *prerogatives*. In these examples, participants' expectations were based on privacy norms about who had the prerogative to make decisions about data collection and sharing in a given situation—where those privacy norms were part of a broader set of social norms about who made decisions in that situation [cf. 3, 88, 110].¹⁰

As we noted in §1, domestic childcare work combines three contexts with different norms about information sharing—home, work, and caregiving [cf. 4, 17]). We observed three common ways of framing prerogatives to control data flows, loosely related to those three contexts. First, participants could see it as *the homeowners' prerogative* to install what technology they chose and use it how they liked, and to protect the safety of their homes. Second, it could be *employers' prerogative* to dictate working conditions and rules. Finally, some participants viewed it as *parents' prerogative* to make choices about how to protect their children's safety, or to keep track of what is happening with their children.

The examples below are roughly grouped according to home, work, and caregiving contexts, but many explicitly highlight the tensions that arise from the overlap—and it is worth noting that participants did not always endorse prerogatives, even when they referred to them as norms.

5.3.1 Homeowner Prerogatives and Home as Baseline

Some participants explained their acceptance of cameras by invoking homeowner prerogatives as an *a priori* assertion, without further explanation: “We have to respect too that we are not in our house, you know, so...” (N15) For some participants, homeowner prerogatives (or device-owner prerogatives) also precluded negotiation about specifics (see §5.4.2): “[I: Do you feel like parents should ask nannies if there's any preferences that they have about the privacy settings? [...]] No. It's your camera. It's your life.” (N16)

A couple of participants invoked the home context in explaining that they understood the use of cameras for monitoring because it could be difficult or strange having someone in your home: “It doesn't get any more personal and private than your home. That's where you go to retreat from the world. If you need a camera there because there's a stranger...” (N16) Other participants were resigned to having less control in someone else's home—especially if it was also a workplace: “I feel that there's a level of relinquishment of my privacy here in the house when I'm working in somebody else's home; I recognize that that is part of the job.” (N4)

¹⁰In general, we assume that someone's *privacy expectations* in a given context are a result of their individual *past experiences* with information flows in that context, their *accumulated knowledge* of how information usually flows in that context, and the *social norms* they are aware of about how information should flow in that context.

As a counterpoint, some viewed cameras as potentially concerning because they felt out of sync with general expectations of privacy when in someone's home: “Like just blowing noses, or just like random stuff like that, that you think, ‘Oh, I'm in someone's house. Like, it's private. I'm fine.’ But it's not private, you know, cause you're on camera. So it's stuff like that that I've had to just think twice about.” (N26)

In particular, the difference between a home and other workplaces made some nannies feel personally targeted: “The daycare center is [...] less sort of, targeted because there's lots of different kids and I assume [...] different employees instead of just sort of the more one-on-one kind of thing.” (N19) A few thought that cameras were *less* expected in a home-based care situation because of the close relationship: “I feel like in a daycare, camera, it's more normal cause [...] you don't have a personal relationship. [...] It's just like a standard. Whereas, if people are trusting you to be in their house with their kids, that's different to me.” (N18)

Several nannies also pointed out how their presence disrupted the privacy the *parents* might expect in their home [cf. 86, 91, 127]: “You're in somebody's home, it's their privacy. [...] I am learning a lot of very intimate details about their lives that they might not show to the outside world.” (N29) A couple of participants viewed cameras as a sort of trade-off for this, even if they would have preferred to work without cameras: “We're kind of like outsiders here, in their private home. So we need to maybe give in a little bit of our privacy.” (N10)

5.3.2 Employer Prerogatives and Workplace as Baseline

Some participants mentioned they would expect to be monitored by their employers in any workplace: “It's not isolated to domestic work; I think that it is not terribly uncommon to work with a camera. [...] I don't expect privacy necessarily while I'm working.” (N4) However, the same participant referenced workplace-based privacy norms she felt should be respected even in a private home: “Because it also your job, for you as a nanny it feels more like a public area, because [...] you're in somebody else's private space, that for you is a workspace. So, I do think that it's important to know when and where there are cameras, for basic privacy reasons.” (N4)

Several nannies pointed out that differences between privacy norms in a home and a workplace caused them to evaluate home cameras differently from other workplace cameras: “[I: Do you feel like [having cameras] is different in a [preschool] versus if it was in somebody's house?] Yeah. [Laughs] Yeah, because it's their house. I mean, like I said, it's something more personal. But, the job is your professional job. [...] This is the difference.” (N24) For this participant, N24, being a live-in au pair—where her employers' home was also hers—introduced additional needs: “If I don't live there, I don't care [whether the camera collects audio], but, I'm going to live, like an au pair, [...] I prefer just video.” (N24)

The tension between expectations based on home norms versus workplace norms manifested in several ways, such as what counted as public versus private space: “I think making sure that the current laws are more clear on what is a private area versus a public area when a home becomes a workplace would be great, [...] especially for live-in nannies, and where they can be recorded and stuff like that.” (N3)

5.3.3 Parental Prerogatives and Shared Caregiving as Baseline

Many participants said that they expected and accepted monitoring at least in part because it was a care situation, and parents were expected to prioritize their children’s safety: “It’s their home and their children, and they have every right to do whatever is in their power to keep their children safe, and if they think that includes video recording, then that is their right.” (N29) Some framed it more generally as parents’ prerogative to make decisions about their children’s care: “That is pretty much my feelings on cameras. It’s your house. It’s your child. You can raise it and do whatever you choose to. [...] I’m not gonna judge a parent on that.” (N16)

In some cases, participants discussed power dynamics and prerogatives negatively, in terms of potential harms or constraints on their choices (see §5.4). However, especially when talking about the care relationship and parental prerogatives, some participants said their understanding of prerogatives made them less uncomfortable with monitoring: “I understand the big brother overtones, but I also understand that parents want to be able to see if their children are doing okay while they’re in the care of somebody else. So it doesn’t bother me, because I recognize my role in the house.” (N4)

While most of the discussion of prerogatives and power dynamics situated the employers as having the most power in the relationship (with nannies having, at most, the power to choose to leave), a few participants highlighted that—whatever the economics of the situation—parents might feel like they were losing power by relinquishing control of their children and their homes: “I understand that in someone’s home, I’m by myself. If I choose to [...] abuse a kid, there’s no one to stop me. So I understand the need for cameras in that sense, where I have all the power with their child.” (N26)

5.4 Power Dynamics Motivating and Constraining Privacy-Related Choices

Participants sometimes made job choices based on how their employers used cameras. But at the same time, their ability to make choices or express preferences about camera data collection was constrained by the power dynamics of employer–employee relationships.

5.4.1 Power Implications Motivating Job Choices

As we described in §5.1 and §5.2, participants were concerned about both what cameras could indicate about their relationships with their employers, and how they might affect those relationships. Those factors might affect whether a nanny accepted a job, or kept a job they had, in a house with cameras. For example, participants might quit or consider quitting a job when video surveillance exacerbated the problem of micro-managing: “That’s actually a reason why I left my previous nanny family. They would constantly check the cameras and text me on certain things that they would do differently or things I was doing wrong in their eyes.” (N35) Participants might also quit if cameras were used in ways that indicated disrespect, distrust, or other negative dynamics (as described in §5.1), e.g. watching at inappropriate times or failing to disclose the cameras at all: “I might actually consider leaving [if I found a hidden camera], because [...] if they didn’t trust me enough to, one, not have them; two, tell me they were putting them up, then there’s the underlying issue there that needs to be addressed. And if they don’t feel comfortable talking to me about it, then maybe we’re not the right fit.” (N33)

5.4.2 Power Constraining Choices About Taking or Keeping Jobs With Cameras

Concerns about the presence and use of cameras might be weighed against other factors—including socioeconomic factors that determined how selective a participant could be. As we noted earlier, participants might not feel they had the power to refuse or leave a job with cameras, given that jobs in camera-free houses were increasingly hard to find: “They’re your boss. You can’t really say no to them. And it’s their house, not only are they your boss, their house, they’re allowed to do what they want. So, saying no, whether or not my feelings [about cameras] are valid for whatever reason is... Yeah, there’re probably gonna be consequences for that.” (N33) On the flip side, a few participants said they might be more willing to put up with cameras—and even micromanaging via cameras—as a trade-off for a higher salary: “I felt like they really expected a lot of me. And that’s why they had cameras, which made it okay for me because they were paying for high expectations. [...] If someone wasn’t paying me well and they wanted to put me on camera, I think I would not.” (N26)

5.4.3 Power Constraining Discussions and Condition-Setting About Data Flows

In discussing the downsides of nannying as a career, many participants noted that having no intermediary (or having only agencies) put them at a disadvantage in negotiating working conditions: “The different characteristics of fair employment are really on you, and it’s a very vulnerable position to be in, especially because you’re in somebody else’s house. The power dynamics are really different, and that way it can be

really tricky for a lot of nannies.” (N4) Against that background, participants had a range of views on what they could reasonably expect an employer to discuss about a camera.

Some participants did not feel that employers were—or even should be—obligated to disclose the existence of cameras at all. However, most believed they had a right to know. But even participants who thought employers *should* disclose did not necessarily assume they *would* disclose, if not forced to do so: “I would expect [employers to tell me if there were cameras], but I know they don’t have to. They’re not obligated to. I think they should [be obligated].” (N32)

The right to know might be framed in terms of *consent* (including implicit consent by accepting a job or continuing to work after camera disclosure): “I probably would not return to that family [to babysit]. [Because of] trust and respect. If they don’t tell me that there are cameras recording, then I do not consent to being recorded.” (N29) However, when we asked whether employers should seek *permission* from their employees, most participants said they did not view that as appropriate. While the words *consent* and *permission* can describe the same interaction, the two words profile a different power balance between the parties involved. While a couple of participants seemed to use the two words interchangeably, others drew an explicit distinction:¹¹ “[I: *Do you think employers should ask nannies for their permission to install cameras inside the house?*] I don’t know if I would probably use the word ‘permission’. I think it is up to the parent. It’s their home, it’s their kids. But I do think asking for the nanny’s consent is [...] necessary.” (N6)

Opinions also differed as to whether it was reasonable to expect employers to discuss details such as purpose and planned use, or specifics about data flows and privacy settings. Some considered it reasonable to ask about these details, especially about the purpose of the camera: “I would still [...] advise [a first-time nanny] to ask about it and ask where they are, and [...] what their plan is with those. How often they’re going to be checking those, and things like that.” (N34) Some viewed it as inappropriate or risky to ask too many questions, even if there were things they would have liked to discuss: “I don’t want to make it seem like I don’t want to be videotaped, [...] like, ‘Oh gosh, what have you seen?’ But then, I would like to ask because I’m curious.” (N26) Others viewed feasibility of asking for details as depending on the current relationship and how good communication was with that employer: “I feel, like, with the boy’s family, I would be comfortable discussing it. And I’d be kind of afraid to discuss it with the girls’ parents because [...] I feel they would get on the defensive.” (N19)

Very few participants thought they were in a position to set conditions on camera use, such as requesting changes to privacy settings. Most viewed cameras as take-it-or-leave-it—

¹¹We were not deliberately varying the wording of our questions to compare participants’ reactions; this was an accidental experiment.

even in the rare cases where they were offered a say in whether cameras were used: “My current nanny families, they both asked me if I’m okay with [cameras], and if not, they would take them down, but prefer to leave them up to monitor the kids in case anything happens. [...] [I: *If you had access to the privacy settings, would you change anything, or would you ask the parents to change anything in these settings?*] [...] No, I think that that is pretty much up to the parents and [...] I’ve already known about the cameras, so the privacy settings are really up to them.” (N35) However, some believed they would have no problem requesting changes if they had an issue: “[I: *Have you ever asked parents about your preferences about the privacy settings of the cameras?*] [...] No, I have never felt like that boundary was crossed. If I did, I would feel very comfortable saying something.” (N4)

6 Discussion

6.1 Summary of Findings

The major takeaways from our findings above are:

- Participants’ views on camera data collection depended in large part on the purpose of data collection, how they thought the data would be used, and how those data uses might affect their relationships with their employers; they were most concerned about whether and how cameras would be used to supervise their work. (RQ1)
- Participants believed that the way employers used cameras reflected relationship qualities, such as trust, respect, and open communication. (RQ2)
- Participants’ views on whether and how employers should use cameras—and how that use interacted with nannies’ privacy rights—often made reference to prerogatives or social norms about who should control data flows, based on general social norms about who made decisions in a given context. (RQ1, RQ2)
- Even where participants believed they had a *right* to control data collection about them, most saw themselves as having a limited *ability* to make choices or express preferences about it, due to power dynamics in employer–employee relationships. (RQ1, RQ2)

In §6.3, we recap specific problems where interventions could have the most potential to mitigate the effects of power dynamics and promote clear, open communication about cameras (RQ3), and suggest corresponding interventions.

6.2 Comparison With Similar Studies

Johnson et al. [53] collected ethnographic data about Filipino migrant domestic workers in Hong Kong, including nannies, and their perceptions of home cameras used for surveillance.

The research showed that pervasive digital surveillance resulted in workers finding ways to evade control, not delivering the best care, and showing signs of negligence. These practices undermined trust between domestic workers and employers. In addition, the study found that control aligned with social hierarchies of gender, race, and class.

The participants in Johnson et al.'s study were in a more precarious position than most of ours reported being, in part because they were subject to rigid immigration rules that required them to live in their employers' house. (Only the two au pairs in our study lived with their employers; no other participants mentioned being dependent on employers for visas.) Also, our study was not limited to cameras used for surveillance. We therefore found a greater range of perspectives on cameras and how they helped or hindered employer–employee relationships, job performance, and job satisfaction. But at the same time, some of the same patterns were reflected in our participants' concerns about excessive surveillance, loss of trust, and control or micromanaging of work.

Albayaydh and Flechais [4] conducted qualitative interviews with domestic workers and employers of domestic workers, exploring privacy attitudes about smart home devices in the home workplace (not specific to cameras). The study was conducted in Jordan and focused on how religion and customs influenced perceptions of smart home devices. The study did not explore device purposes and uses in depth, but noted that some employees expected employers not to hide or use monitoring devices maliciously, due to norms based on Islamic religious beliefs that forbid breaching the privacy rights of others. However, in the end, many employers did not disclose smart devices, either purposely or because they assumed employees already knew—and similarly to our study, employees viewed nondisclosure as an indicator of distrust.

6.3 Points of Intervention and Recommended Mitigation Strategies

We found that power imbalances had adverse effects on nannies' privacy and individual agency with regard to camera data collection—mainly due to the fact that employers own the cameras, and, hence, employers have the power to choose who can access cameras and their settings. Efforts to create advanced controls (see §2) and education about device functions and configuration are insufficient. Social interventions such as those we suggest here are needed to guide parents and nannies in negotiating privacy matters and increase nannies' agency in the smart home context.

Privacy and Security Discussion Guides for Parents and Nannies Participants valued transparency about the existence and uses of cameras, to the point where they might quit if they found a hidden camera, or even hidden uses of a disclosed camera. They identified *communication at time of hire*

as an especially effective point of intervention at which to mitigate concerns. Even with very obviously visible cameras, they said they would like an opportunity to ask questions. To help both parents and nannies navigate such conversations and make it easier to introduce potentially sensitive questions, we propose designing digital privacy and security discussion guides. Such guides might include advice about the mutual benefits of transparency around cameras; a list of possible discussion points to structure the conversation; and guidelines on how different smart home stakeholders can be involved in deciding on the configuration of a camera.

Further research is needed to expand on, verify, and quantify the considerations to prioritize for such a guide. However, our findings so far suggest these frequently-mentioned questions:

- Whether there are cameras present, how many there are, and where they are located.
- What type of data cameras collect (audio/video), whether it is recorded (as opposed to livestreamed), and, if so, how long recordings will be kept.
- How often cameras will be checked and how camera data will be used, especially whether they will be used to supervise nannies' work.
- Whether the nanny will be able to use the camera as a baby monitor, or otherwise have access to the data.
- Under what conditions the nanny is comfortable with the employer sharing video of her with third parties or on social media—and when nannies may share pictures or videos of children.

Encouraging such open conversations about cameras may function as a trust-building intervention to help address power imbalances.

Discussion guides should be co-designed with participation from domestic childcare workers and employers thereof [cf. 121, 122], and iteratively tested, refined, and validated. Guides should be jargon-free and accessible and translated into non-English languages commonly spoken in the U.S.

These discussion guides can supplement existing materials for nannies and other domestic workers about privacy issues and rights with respect to cameras and other smart home devices, provided by organizations like the National Domestic Workers Alliance in the U.S. or Voice of Domestic Workers in the UK [121]. Agencies are also well-positioned to mitigate power imbalances by encouraging discussions, and when asked, some participants thought *agencies could facilitate conversations about cameras*—or at least inform parents that the conversation should be had [cf. 4].

Promoting Domestic Worker Agency at Point of Configuration Few of our participants had discussed camera con-

figuration and privacy settings with employers, and none had been actively involved in choosing settings. When asked, many did not have strong opinions about what the settings should be, or thought it was not their place—but more did at least want to know what the current configurations were. Some participants opined that employers did not think to bring it up because it was not a normal *expectation to discuss device settings with non-household members*. Meanwhile, some employers could configure or use cameras in problematic ways unintentionally; in such cases, an alert might be effective in averting privacy infringement.

Prior work on bystander and secondary-user privacy has suggested nudging a device owner about a visitor’s known preferences [144], or incorporating social interventions such as alerts and nudges into the interfaces of smart home devices [42, 91, 148] that would encourage the owner of the device to involve other occupants of the home in setup processes and alert them to potential violations of information-sharing norms. This idea could be expanded to encourage owners to consider the needs of non-occupants as well [25, 105, 128], including domestic workers, and could incorporate a discussion guide such as that suggested above.

Relatedly, some participants noted that nannies’ *control over or access to camera data* was a point of intervention where employers could feasibly share power, reducing uncertainty about data handling and allowing nannies to use the same device to monitor children. Nudges could encourage configuring options to allow this.

Design Guidelines for Smart Home Camera Product Teams To bridge the gap between academic research findings and industry practice, we suggest creating design guidelines that explicitly foreground the needs of domestic workers, and provide practical recommendations for balancing conflicting needs and privacy concerns of employers and employees. Different stakeholders should take part in developing and refining the design guidelines [cf. 34]: primary users/device owners, domestic childcare workers, and camera product teams, e.g. in participatory design workshops [87, 120]. Guidelines could also incorporate findings from other user studies with diverse types of bystanders (see §2). In addition, guidelines could promote value-sensitive tech product design (VSD) [survey in 37] [for privacy: 10] (e.g. accounting for potential use of cameras in covert surveillance) and educate developers about relevant privacy regulations.

In many cases, there are existing solutions that could be adapted to the use case; however, guidelines should emphasize that enhanced features may need to accommodate nuance. For example, as we noted in §5.4.3, some participants found it overly distracting to know when a camera was being watched live. It should therefore be possible for the nanny to choose to turn this feature off—and yet the design ought not to make

it easy for employers to hide that they are watching.

6.4 Future Research

Different smart home devices have different purposes of use and data practices (collection, use, storage, and sharing), leading nannies to think differently about them. Besides cameras, our interview script included questions asking participants about their views on smart speakers, smart TVs, and location trackers. In a forthcoming paper, we compare nannies’ perspectives on these different devices.

In this paper, we focused on the employees’ (nannies’) perspectives. In future work, we will explore the other side of the equation: employers (parents), based on the interviews we conducted (see §3). In that work, we will compare the privacy threats perceived by nannies with those perceived by parents in smart homes, and examine how those threat models may influence the choices of each.

This paper explored the perspectives of nannies; future work should explore and compare the needs and concerns of other groups of bystanders with regard to cameras as well as other smart home devices, as specific needs and threats may differ, and different vulnerabilities may need to be addressed. Even amongst domestic workers, different job types may lead to differences in experiences and views, due to differences in social prestige, central management, and opportunities for building relationships and trust with employers. We are currently designing large-scale surveys to quantify our findings with domestic childcare workers and compare that situation with other bystander contexts in smart homes.

7 Conclusion

We conducted 25 semi-structured interviews with domestic childcare workers in the U.S. about smart home cameras, investigating how domestic employees navigate a multi-layered context that blends privacy and social norms for homes, workplaces, and caregiving. We examined how privacy considerations interact with the dynamics of employer–employee relationships in an in-home childcare context. Power differentials and norms about who should decide how information flows in a given situation affected participants’ perspectives on camera data practices, as well as their ability to make choices and requests about camera data collection. Purposes and manner of use especially influenced participants’ attitudes about cameras, because those factors both reflected and affected their relationships with their employers. (E.G., employers using cameras to micromanage and excessively monitor participants signaled disrespect and a lack of trust on the employers’ part.) Drawing on the findings of this study, we suggest a set of technical and social interventions that balance power dynamics in smart homes with a focus on cameras, to improve domestic employees’ privacy and support their individual agency.

Acknowledgments

We are grateful to Maritza Johnson for proposing this project, and to others who have helped with suggestions and resources along the way, including Franziska Roesner, Serge Egelman, Yasemin Acar, Sascha Fahl, Julia Słupska, and Wael Albayaydh. We also thank participants at the 2019 Symposium on the Applications of Contextual Integrity and the 2020 Workshop on Free and Open Communications on the Internet (FOCI), as well as anonymous reviewers for FOCI and SOUPS, for their helpful comments.

This research was supported in part by grants from the Center for Long-Term Cybersecurity at the University of California, Berkeley, the U.S. National Security Agency (contract H98230-18-D-0006), and the U.S. National Science Foundation (award CNS-2114229).

References

- [1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, pages 451–466, Santa Clara, CA, USA, 2019.
- [2] Paarijaat Aditya, Rijurekha Sen, Peter Druschel, Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele, Bobby Bhattacharjee, and Tong Tong Wu. I-pic: A platform for privacy-compliant image capture. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 235–248, New York, NY, USA, 2016.
- [3] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J Lee. Tangible privacy: Towards user-centric sensor designs for bystander privacy. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW):1–28, 2020.
- [4] Wael Albayaydh and Ivan Flechais. Exploring bystanders’ privacy concerns with smart homes in Jordan. In *ACM Conference on Human Factors in Computing Systems (CHI)*, New York, NY, USA, 2022.
- [5] Mark Andrejevic. Big data, big questions: The big data divide. *International Journal of Communication*, 8(0), 2014.
- [6] Noah Apthorpe, Pardis Emami-Naeini, Arunesh Mathur, Marshini Chetty, and Nick Feamster. You, me, and IoT: How Internet-connected consumer devices affect interpersonal relationships. *arXiv:2001.10608 [cs]*, July 2020.
- [7] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering smart home Internet of Things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies (IMWUT)*, 2(2), June 2018.
- [8] Noah Apthorpe, Sarah Varghese, and Nick Feamster. Evaluating the contextual integrity of privacy regulation: Parents’ IoT toy privacy norms versus COPPA. In *USENIX Security Symposium (USENIX Security)*, pages 123–140, Santa Clara, CA, USA, August 2019.
- [9] Andrew Baerg. Big data, sport, and the digital divide: Theorizing how athletes might respond to big data monitoring. *Journal of Sport and Social Issues*, 41(1):3–20, 2017.
- [10] Gianmarco Baldini, Maarten Botterman, Ricardo Neisse, and Mariachiara Tallacchini. Ethical design in the Internet of Things. *Science and Engineering Ethics*, 24(3):905–925, June 2018.
- [11] Kirstie Ball. Workplace surveillance: An overview. *Labor History*, 51(1):87–106, 2010.
- [12] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. Proxemic interaction: Designing for a proximity and orientation-aware environment. In *ACM International Conference on Interactive Tabletops and Surfaces (ITS)*, pages 121–130, 2010.
- [13] William Balmford, Larissa Hjorth, and Ingrid Richardson. Supervised play: Intimate surveillance and children’s mobile media usage. *The Routledge Companion to Digital Media and Children*, page 185, 2020.
- [14] Natā M. Barbosa, Zhuohao Zhang, and Yang Wang. Do privacy and security matter to everyone? Quantifying and clustering user-centric considerations about smart home device adoption. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, pages 417–435, August 2020.
- [15] Adam Barth, Anupam Datta, John C. Mitchell, and Helen Nissenbaum. Privacy and contextual integrity: Framework and applications. In *IEEE Symposium on Security and Privacy (SP)*, pages 184–198, Washington, DC, USA, 2006.
- [16] Sebastian Benthall and Bruce D. Haynes. Contexts are political: Field theory and privacy. Presentation at the Symposium on Applications of Contextual Integrity, Berkeley, CA, USA, August 19–20, 2019.
- [17] Julia Bernd, Ruba Abu-Salma, and Alisa Friik. Bystanders’ privacy: The perspectives of nannies on smart home surveillance. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, August 2020.

- [18] Clara Berridge, Jodi Halpern, and Karen Levy. Cameras on beds: The ethics of surveillance in nursing home rooms. *AJOB Empirical Bioethics*, 10(1):55–62, 2019.
- [19] danah boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society*, 15(5):662–679, 2012.
- [20] Alison Burrows, David Coyle, and Rachael Gooberman-Hill. Privacy, boundaries and smart homes for health: An ethnographic study. *Health & Place*, 50:112–118, 2018.
- [21] Anadi Chaman, Jiaming Wang, Jiachen Sun, Haitham Hassanieh, and Romit Roy Choudhury. Ghostbuster: Detecting the presence of hidden eavesdroppers. In *ACM International Conference on Mobile Computing and Networking (MobiCom)*, pages 337–351, New Delhi, India, 2018.
- [22] Menzie D. Chinn and Robert W. Fairlie. The determinants of the global digital divide: A cross-country analysis of computer and Internet penetration. *Oxford Economic Papers*, 59(1):16–44, 2007.
- [23] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, Shwetak N. Patel, and Julie A. Kientz. Investigating receptiveness to sensing and inference in the home using sensor proxies. In *ACM Conference on Ubiquitous Computing (UbiComp)*, pages 61–70, 2012.
- [24] Ian Clark. The digital divide in the post-Snowden era. *Journal of Radical Librarianship*, 2, March 2016.
- [25] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. “I would have to evaluate their objections”: Privacy tensions between smart home device owners and incidental users. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2021(4):54–75, 2021.
- [26] Camille Cobb, Milijana Surbatovich, Anna Kawakami, Mahmood Sharif, Lujo Bauer, Anupam Das, and Limin Jia. How risky are real users’ IFTTT applets? In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, pages 505–529, August 2020.
- [27] Anupam Das, Martin Degeling, Xiaoyou Wang, Junjue Wang, Norman Sadeh, and Mahadev Satyanarayanan. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1387–1396, July 2017.
- [28] Mario De La Rosa, Rosa Babino, Adelaida Rosario, Natalia Valiente Martinez, and Lubna Aijaz. Challenges and strategies in recruiting, interviewing, and retaining recent Latino immigrants in substance abuse and HIV epidemiologic studies. *The American Journal on Addictions*, 21(1):11–22, 2012.
- [29] Rajib Dey, Sayma Sultana, Afsaneh Razi, and Pamela J Wisniewski. Exploring smart home device use by Airbnb hosts. In *ACM Conference on Human Factors in Computing Systems (CHI): Extended Abstracts*, pages 1–8, New York, New York, United States, 2020.
- [30] Leen d’Haenens and Christine Ogan. Internet-using children and digital inequality: A comparison between majority and minority Europeans. *Communications: The European Journal of Communication Research*, 38(1):41–60, 2013.
- [31] David Eckhoff and Isabel Wagner. Privacy in the smart city: Applications, technologies, challenges, and solutions. *IEEE Communications Surveys & Tutorials*, 20(1):489–516, Firstquarter 2018.
- [32] Serge Egelman, Raghudeep Kannavara, and Richard Chow. Is this thing on? Crowdsourcing privacy indicators for ubiquitous sensing platforms. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1669–1678, New York, NY, USA, 2015.
- [33] Ame Elliott and Sara Brody. Straight talk: New yorkers on mobile messaging and implications for privacy. Technical report, 2015. Accessed: 10 December, 2018.
- [34] P. Emami-Naeini, Y. Agarwal, L. Cranor, and H. Hibshi. Ask the experts: What should be on an IoT privacy and security label? In *IEEE Symposium on Security and Privacy (SP)*, pages 771–788, Los Alamitos, CA, USA, may 2020.
- [35] Angella Foster. When parents eavesdrop on nannies. *New York Times*, August 2019. Accessed: 8 June, 2020.
- [36] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. “A stalker’s paradise”: How intimate partner abusers exploit technology. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–13, New York, NY, USA, 2018.
- [37] Batya Friedman, David G Hendry, and Alan Borning. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(2):63–125, 2017.
- [38] Alisa Frik, Julia Bernd, and Serge Egelman. A model of contextual factors affecting older adults’ information-sharing decisions in the US. *ACM Trans-*

actions on *Computer-Human Interaction*, 2022. To appear.

- [39] Jon P. Gant, Nicole E. Turner-Lee, Ying Li, and Joseph S. Miller. Minority broadband adoption: Comparative trends in adoption, acceptance and use. Technical report, Joint Center for Political and Economic Studies, Washington, DC, USA, February 2010.
- [40] Radhika Garg and Hua Cui. Social contexts, agency, and conflicts: Exploring critical aspects of design for future smart home technologies. *ACM Transactions on Computer-Human Interaction*, 29(2):11:1–11:30, January 2022.
- [41] George Gaskell and Martin W. Bauer. Towards public accountability: Beyond sampling, reliability and validity. In *Qualitative Researching with Text, Image and Sound*, pages 337–350. SAGE Publications Ltd., London, UK, 2000.
- [42] Christine Geeng and Franziska Roesner. Who’s in control? Interactions in multi-user smart homes. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–13, New York, NY, USA, 2019.
- [43] Emily Starbuck Gerson. Nanny cams: What parents need to know before installing a home security camera, January 2019. Blog post; accessed: 8 June, 2020.
- [44] Marco Ghiglieri, Melanie Volkamer, and Karen Renaud. Exploring consumers’ attitudes of smart TV related privacy risks. In *International Conference on Human Aspects of Information Security, Privacy, and Trust (HAS)*, Lecture Notes in Computer Science, pages 656–674, Cham, 2017. Springer.
- [45] Jessica Groopman and Susan Etlinger. Consumer perceptions of privacy in the Internet of Things: What brands can learn from a concerned citizenry. Technical report, June 2015. Accessed: 17 February, 2018.
- [46] Tamy Guberek, Allison McDonald, Sylvia Simioni, Abraham H Mhaidli, Kentaro Toyama, and Florian Schaub. Keeping a low profile?: Technology, risk and privacy among undocumented immigrants. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–15, New York, NY, USA, 2018.
- [47] Loni Hagen. Overcoming the privacy challenges of wearable devices: A study on the role of digital literacy. In *ACM International Conference on Digital Government Research*, dg.o ’17, pages 598–599, New York, NY, USA, 2017.
- [48] Julie M. Haney, Susanne M. Furman, and Yasemin Acar. Smart home security and privacy mitigations: Consumer perceptions, practices, and challenges. In *International Conference on Human-Computer Interaction*, pages 393–411. Springer International Publishing, 2020.
- [49] Eszter Hargittai and Eden Litt. New strategies for employment? Internet skills and online privacy practices during people’s job search. *IEEE Security & Privacy*, 11(3):38–45, May 2013.
- [50] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlene Fernandes, and Blase Ur. Rethinking access control and authentication for the home Internet of Things (IoT). In *USENIX Security Symposium (USENIX Security)*, pages 255–272, Baltimore, MD, USA, August 2018.
- [51] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. Amazon vs. my brother: How users of shared smart speakers perceive and cope with privacy risks. In *ACM Conference on Human Factors in Computing Systems (CHI)*, page 1–13, New York, NY, USA, 2020.
- [52] Haojian Jin, Boyuan Guo, Rituparna Roychoudhury, Yaxing Yao, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. Exploring the needs of users for supporting privacy-protective behaviors in smart homes. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–19, New York, NY, USA, April 2022.
- [53] Mark Johnson, Maggy Lee, Michael McCahill, and Ma Rosalyn Mesina. Beyond the ‘all seeing eye’: Filipino migrant domestic workers’ contestation of care and control in Hong Kong. *Ethnos*, 85(2):276–292, 2020.
- [54] Saba Kazi, Omead Kohanteb, Thidanun Saensuksopa, Owen Tong, and Heidi Yang. Decoding sensors: Creating guidelines for designing connected devices. Technical report, Carnegie Mellon University, Summer 2015. Accessed: 7 March, 2018.
- [55] Jennifer King. *Privacy, Disclosure, and Social Exchange Theory*. PhD thesis, University of California, Berkeley, CA, 2018.
- [56] Jennifer King and Andreas Katsanevas. Blending contextual integrity and social exchange theory: Assessing norm building under conditions of “informational inequality”. Presentation at the 2nd Symposium on Applications of Contextual Integrity, Berkeley, CA, USA, August 19–20, 2019.
- [57] Thorin Klosowski. Your visitors deserve to know they’re on camera. *New York Times*, October 2019. Accessed: 8 June, 2020.
- [58] Marion Koelle, Katrin Wolf, and Susanne Boll. Beyond LED status lights: Design requirements of privacy no-

- tices for body-worn cameras. In *ACM International Conference on Tangible, Embedded, and Embodied Interaction (TEI)*, pages 177–187, New York, NY, USA, 2018.
- [59] Omead Kohanteb, Owen Tong, Heidi Yang, Thidanun Saensuksopa, and Saba Kazi. signifiers.io guidelines for designing connected devices, 2015. Accessed: 26 February, 2018.
- [60] Vinay Koshy, Joon Sung Sung Park, Ti-Chung Cheng, and Karrie Karahalios. “We just use what they give us”: Understanding passenger user perspectives in smart homes. In *ACM Conference on Human Factors in Computing Systems (CHI)*, New York, NY, USA, 2021.
- [61] Martin Kraemer and Ivan Flechais. Disentangling privacy in smart homes, 2018. Presentation at the 1st Symposium on Applications of Contextual Integrity, Princeton, NJ, USA, September 13–14, 2018. Accessed: 20 November, 2018.
- [62] Martin J. Kraemer, Ivan Flechais, and Helena Webb. Exploring communal technology use in the home. In *ACM Halfway to the Future Symposium (HTTF)*, New York, NY, USA, 2019.
- [63] Martin J. Kraemer, Ulrik Lyngs, Helena Webb, and Ivan Flechais. Further exploring communal technology use in smart homes: Social expectations. In *ACM Conference on Human Factors in Computing Systems (CHI): Extended Abstracts*, pages 1–7, New York, New York, United States, 2020.
- [64] Martin J. Kraemer, William Seymour, and Ivan Flechais. Responsibility and privacy: Caring for a dependent in a digital age. In *CHI Workshop on Privacy and Power (Networked Privacy)*, 2020.
- [65] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–31, 2018.
- [66] Tama Leaver. Intimate surveillance: Normalizing parental monitoring and mediation of infants online. *Social Media+ Society*, 3(2):2056305117707192, 2017.
- [67] Hosub Lee and Alfred Kobsa. Understanding user privacy in Internet of Things environments. In *IEEE World Forum on Internet of Things (WF-IoT)*, pages 407–412, December 2016.
- [68] Linda Lee, Joong Hwa Lee, Serge Egelman, and David Wagner. Information disclosure concerns in the age of wearable computing. In *NDSS Workshop on Usable Security (USEC)*. Internet Society, 2016.
- [69] Samantha Lee and Brian H Kleiner. Electronic surveillance in the workplace. *Management Research News*, 26:72–81, 2003.
- [70] Roxanne Leitão. Anticipating smart home security and privacy threats with survivors of intimate partner abuse. In *ACM Conference on Designing Interactive Systems (DIS)*, page 527–539, New York, NY, USA, 2019.
- [71] Karen Levy, Lauren Kilgour, and Clara Berridge. Regulating privacy in public/private space: The case of nursing home monitoring laws. *The Elder Law Journal*, February 2019.
- [72] Heather Richter Lipford, Madiha Tabassum, Paritosh Bahirat, Yaxing Yao, and Bart P Knijnenburg. Privacy and the Internet of Things. In *Modern Socio-Technical Perspectives on Privacy*, page 233. Springer, 2022.
- [73] Eden Litt and Eszter Hargittai. Smile, snap, and share? A nuanced approach to privacy and online photo-sharing. *Poetics*, 42:1–21, 2014.
- [74] Tian Liu, Ziyu Liu, Jun Huang, Rui Tan, and Zhen Tan. Detecting wireless spy cameras via stimulating and probing. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 243–255, 2018.
- [75] Steve Lohr. Unblinking eyes track employees. *New York Times*, June 2014. Accessed: 23 July, 2018.
- [76] Isabel Lopez-Neira, Trupti Patel, Simon Parkin, George Danezis, and Leonie Tanczer. ‘Internet of Things’: How abuse is getting smarter. *Safe – The Domestic Abuse Quarterly*, 63:22–26, 2019. Available at SSRN: <https://ssrn.com/abstract=3350615>.
- [77] Lesa Lorenzen-Huber, Mary Boutain, L. Jean Camp, Kalpana Shankar, and Kay H. Connelly. Privacy, technology, and aging: A proposed framework. *Ageing International*, 36(2):232–252, June 2011.
- [78] Deborah Lupton. Self-tracking cultures: Towards a sociology of personal informatics. In *ACM Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design (OzCHI)*, pages 77–86, New York, NY, USA, 2014.
- [79] Mary Madden, Michele E. Gilman, Karen Levy, and Alice E. Marwick. Privacy, poverty, and big data: A matrix of vulnerabilities for poor Americans. *Washington University Law Review*, 95(1):53–125, 2017.
- [80] Mary Madden and Lee Rainie. Americans’ attitudes about privacy, security, and surveillance. Technical report, Pew Research Center, May 2015. Accessed: 9 February, 2018.

- [81] Nathan Malkin, Julia Bernd, Maritza Johnson, and Serge Egelman. “What *can’t* data be used for?” Privacy expectations about smart TVs in the U.S. In *European Workshop on Usable Security (EuroUSEC)*, London, UK, 2018.
- [82] Nathan Malkin, Joe Deatruck, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2019(4):250–271, 2019.
- [83] Lev Manovich. Trending: The promises and the challenges of big social data. In *Debates in the Digital Humanities*, pages 460–475. The University of Minnesota Press, Minneapolis, MN, 2011.
- [84] Shrirang Mare, Franziska Roesner, and Tadayoshi Kohno. Smart devices in Airbnbs: Considering privacy and security for both guests and hosts. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2020(2):436 – 458, 2020.
- [85] Karola Marky, Sarah Prange, Florian Krell, Max Mühlhäuser, and Florian Alt. “You just can’t know about everything”: Privacy perceptions of smart home visitors. In *International Conference on Mobile and Ubiquitous Multimedia (MUM)*, pages 83–95, 2020.
- [86] Karola Marky, Alexandra Voit, Alina Stöver, Kai Kunze, Svenja Schröder, and Max Mühlhäuser. “I don’t know how to protect myself”: Understanding privacy perceptions resulting from the presence of bystanders in smart environments. In *ACM Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI)*, New York, NY, USA, 2020.
- [87] Michael Massimi and Ronald Baecker. Participatory design process with older users. In *UbiComp Workshop on Future Media*, 2006.
- [88] Michelle L. Mazurek, J. P. Arsenault, Joanna Bresee, Nitin Gupta, Iulia Ion, Christina Johns, Daniel Lee, Yuan Liang, Jenny Olsen, Brandon Salmon, Richard Shay, Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor, Gregory R. Ganger, and Michael K. Reiter. Access control for home data sharing: Attitudes, needs, and practices. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 645–654, New York, NY, USA, 2010.
- [89] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.
- [90] Emily McReynolds, Sarah Hubbard, Timothy Lau, Aditya Saraf, Maya Cakmak, and Franziska Roesner. Toys that listen: A study of parents, children, and Internet-connected toys. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 5197–5207, New York, NY, USA, 2017.
- [91] Nicole Meng, Dilara Keküllüoğlu, and Kami Vaniea. Owning and sharing: Privacy perceptions of smart speaker users. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW), April 2021.
- [92] Simon Moncrieff, Svetha Venkatesh, and Geoff West. Dynamic privacy in a smart house environment. In *IEEE International Conference on Multimedia and Expo*, pages 2034–2037, 2007.
- [93] Alessandro Montanari, Afra Mashhadi, Akhil Mathur, and Fahim Kawsar. Understanding the privacy design space for personal connected objects. In *International BCS Human Computer Interaction Conference: Fusion! (HCI)*, pages 18:1–18:13, Swindon, UK, 2016. BCS Learning & Development Ltd.
- [94] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy expectations and preferences in an IoT world. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, pages 399–412, Santa Clara, CA, 2017.
- [95] Carman Neustaedter and Saul Greenberg. The design of a context-aware home media space for balancing privacy and awareness. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 297–314. Springer, 2003.
- [96] Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(119):101–139, 2004.
- [97] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.
- [98] Helen Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, Fall 2011.
- [99] Parmy Olson. Wearable tech is plugging into health insurance. *Forbes*, June 2014. Accessed: 23 July, 2018.
- [100] Yong Jin Park. Digital literacy and privacy behavior online. *Communication Research*, 40(2):215–236, 2013.
- [101] Simon Parkin, Trupti Patel, Isabel Lopez-Neira, and Leonie Tanczer. Usability analysis of shared device ecosystem security: Informing support for survivors of IoT-facilitated tech-abuse. In *New Security Paradigms Workshop (NSPW)*, pages 1–15, 2019.

- [102] Shwetak N Patel, Jay W Summet, and Khai N Truong. Blindspot: Creating capture-resistant spaces. In *Protecting Privacy in Video Surveillance*, pages 185–201. Springer, 2009.
- [103] Scott R. Peppet. Regulating the Internet of Things: First steps toward managing discrimination, privacy, security & consent. *Texas Law Review*, 93:85–178, 2014.
- [104] Alfredo J Perez, Sherali Zeadally, and Scott Griffith. Bystanders’ privacy. *IT Professional*, 19(3):61–65, 2017.
- [105] James Pierce, Claire Weizenegger, Parag Nandi, Isha Agarwal, Gwenna Gram, Jade Hurlle, Betty Lo, Aaron Park, Aivy Phan, Mark Shumskiy, and Grace Sturlaugson. Addressing adjacent actor privacy: Designing for bystanders, co-users, and surveilled subjects of smart home cameras. In *ACM Conference on Designing Interactive Systems (DIS)*, 2022. To appear.
- [106] James Pierce, Richmond Y. Wong, and Nick Merrill. Sensor illumination: Exploring design qualities and ethical implications of smart cameras and image/video analytics. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–19, New York, NY, USA, 2020.
- [107] Lucinda Platt, Renee Luthra, and Tom Frere-Smith. Adapting chain referral methods to sample new migrants: Possibilities and limitations. *Demographic Research*, 33:665–700, 2015.
- [108] Rebecca S. Portnoff, Linda N. Lee, Serge Egelman, Pratyush Mishra, Derek Leung, and David Wagner. Somebody’s watching me?: Assessing the effectiveness of webcam indicator lights. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1649–1658, New York, NY, USA, 2015.
- [109] Alison Powell, Amelia Bryne, and Dharma Dailey. The essential Internet: Digital exclusion in low-income American communities. *Policy & Internet*, 2(2):161–192, 2010.
- [110] Nicholas Proferes. The development of privacy norms. In *Modern Socio-Technical Perspectives on Privacy*, pages 79–90. Springer, Cham, 2022.
- [111] Lee Rainie and Janna Anderson. The future of privacy. Technical report, Pew Research Center, December 2014. Accessed: 17 July, 2018.
- [112] Lee Rainie and Maeve Duggan. Privacy and information sharing. Technical report, Pew Research Center, January 2016. Accessed: 16 February, 2022.
- [113] Olivia Richards and Gabriela Marcu. Children’s agency in the age of smart things. In *CHI Workshop on Privacy and Power (Networked Privacy)*, 2020.
- [114] Laura Robinson, Shelia R. Cotten, Hiroshi Ono, Anabel Quan-Haase, Gustavo Mesch, Wenhong Chen, Jeremy Schulz, Timothy M. Hale, and Michael J. Stern. Digital inequalities and why they matter. *Information, Communication & Society*, 18(5):569–582, 2015.
- [115] Laura Robinson and Brian K. Gran. No kid is an island: Privacy scarcities and digital inequalities. *American Behavioral Scientist*, 2018.
- [116] Ignacio Rodríguez-Rodríguez, José-Víctor Rodríguez, Aránzazu Elizondo-Moreno, Purificación Heras-González, and Michele Gentili. Towards a holistic ICT platform for protecting intimate partner violence survivors based on the IoT paradigm. *Symmetry*, 12(1):37, 2020.
- [117] Franziska Roesner, Tamara Denning, Bryce Clayton Newell, Tadayoshi Kohno, and Ryan Calo. Augmented reality: Hard problems of law and policy. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp): Adjunct Publication*, pages 1283–1288, New York, NY, USA, 2014.
- [118] Safe Smart Living. 16 smart home statistics and predictions, October 2019. Web page; accessed: 16 July, 2020.
- [119] Jeremy Schiff, Marci Meingast, Deirdre K Mulligan, Shankar Sastry, and Ken Goldberg. Respectful cameras: Detecting visual markers in real-time to address privacy concerns. In *Protecting Privacy in Video Surveillance*, pages 65–89. Springer, 2009.
- [120] Douglas Schuler and Aki Namioka. *Participatory design: Principles and practices*. CRC Press, 1993.
- [121] Julia Słupska, Marissa Begonia, Nayana Prakash, Selina Cho, Ruba Abu-Salma, Mallika Balakrishnan, and Natalie Sedacca. Digital privacy & security guide for migrant domestic workers. Technical report, University of Oxford, King’s College London, Voice of Domestic Workers, and Migrants Organise, September 2021. Web page; accessed: 14 February, 2022.
- [122] Julia Słupska, Selina Cho, Marissa Begonia, Ruba Abu-Salma, Nayanatara Prakash, and Mallika Balakrishnan. “They look at vulnerability and use that to abuse you”: Participatory threat modelling with migrant domestic workers. In *USENIX Security Symposium (USENIX Security)*, Boston, MA, USA, August 2022. To appear.
- [123] Yunpeng Song, Yun Huang, Zhongmin Cai, and Jason I. Hong. I’m all eyes and ears: Exploring effective locators for privacy awareness in IoT scenarios. In *ACM*

Conference on Human Factors in Computing Systems (CHI), pages 1–13, New York, NY, USA, 2020.

- [124] Luke Stark and Karen Levy. The surveillant consumer. *Media, Culture, and Society*, 40(8):1202–20, November 2018.
- [125] Statista. Smart home penetration rate forecast worldwide from 2017 to 2024, June 2020. Web page; accessed: 16 July, 2020.
- [126] Kaiwen Sun, Yixin Zhou, Jenny Radesky, Christopher Brooks, and Florian Schaub. Child safety in the smart home: Parents’ perceptions, needs, and mitigation strategies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW), 2021.
- [127] Janos Mark Szokolczai. “What have you caught?”: Nannycams and hidden cameras as normalised surveillance of the intimate. In *The Technologisation of the Social*. Routledge, 2021.
- [128] Neilly H. Tan, Richmond Y. Wong, Audrey Desjardins, Sean A. Munson, and James Pierce. Monitoring pets, deterring intruders, and casually spying on neighbors: Everyday uses of smart home cameras. In *ACM Conference on Human Factors in Computing Systems (CHI)*, New York, NY, USA, 2022.
- [129] Parth Kirankumar Thakkar, Shijing He, Shiyu Xu, Danny Yuxing Huang, and Yaxing Yao. “It would probably turn into a social faux-pas”: Users’ and bystanders’ preferences of privacy awareness mechanisms in smart homes. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–13, New York, NY, USA, April 2022.
- [130] Peter Tolmie, Andy Crabtree, Tom Rodden, James Colley, and Ewa Luger. “This has to be the cats”: Personal data legibility in networked sensing systems. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW):491–502, 2016.
- [131] Khai N Truong, Shwetak N Patel, Jay W Summet, and Gregory D Abowd. Preventing camera recording by designing a capture-resistant environment. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 73–86. Springer, 2005.
- [132] Joseph Turow, Lauren Feldman, and Kimberly Meltzer. Open to exploitation: America’s shoppers online and offline. Technical report, Annenberg Public Policy Center of the University of Pennsylvania, June 2005. Accessed: 3 June, 2015.
- [133] Joseph Turow, Michael Hennessy, and Nora Draper. The tradeoff fallacy: How marketers are misrepresenting American consumers and opening them up to exploitation. Technical report, Annenberg Public Policy Center of the University of Pennsylvania, June 2015. Accessed: 24 February, 2018.
- [134] Joseph Turow, Michael Hennessy, Nora Draper, Ope Akanbi, and Diami Virgilio. Divided we feel: Partisan politics drive Americans’ emotions regarding surveillance of low-income population. Technical report, Annenberg School for Communication at the University of Pennsylvania, 2018. Accessed: 24 December, 2018.
- [135] Blase Ur, Jaeyeon Jung, and Stuart Schechter. Intruders versus intrusiveness: Teens’ and parents’ perspectives on home-entryway surveillance. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, page 129–139, New York, NY, USA, 2014.
- [136] Alexander J.A.M. van Deursen and Jan A.G.M. van Dijk. Internet skills and the digital divide. *New Media & Society*, 13(6):893–911, 2010.
- [137] Myria Watkins Allen, Stephanie J Coopman, Joy L Hart, and Kasey L Walker. Workplace surveillance and managing privacy boundaries. *Management Communication Quarterly*, 21(2):172–200, 2007.
- [138] Meredydd Williams, Jason R. C. Nurse, and Sadie Creese. “Privacy is the boring bit”: User perceptions and behaviour in the Internet-of-Things. In *IEEE International Conference on Privacy, Security, and Trust (PST)*, August 2017.
- [139] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. Smart homes and their users: A systematic analysis and key challenges. *Personal and Ubiquitous Computing*, 19(2):463–476, February 2015.
- [140] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. Benefits and risks of smart home technologies. *Energy Policy*, 103:72–83, April 2017.
- [141] Jenifer Sunrise Winter. Citizen perspectives on the customization/privacy paradox related to smart meter implementation. *International Journal of Technoethics*, 6(1), 2015.
- [142] Julia Wolfe, Jori Kandra, Lora Engdahl, and Heidi Shierholz. Domestic workers chartbook: A comprehensive look at the demographics, wages, benefits, and poverty rates of the professionals who care for our family members and clean our homes. Technical report, Economic Policy Institute, May 2020. Accessed: 21 July, 2020.
- [143] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending my castle: A co-design study of

- privacy mechanisms for smart homes. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–12, New York, NY, USA, May 2019.
- [144] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata McDonough, and Yang Wang. Privacy perceptions and designs of bystanders in smart homes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, November 2019.
- [145] Yaxing Yao, Huichuan Xia, Yun Huang, and Yang Wang. Free to fly in public spaces: Drone controllers’ privacy perceptions and practices. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 6789–6793, New York, NY, USA, 2017.
- [146] Yaxing Yao, Huichuan Xia, Yun Huang, and Yang Wang. Privacy mechanisms for drones: Perceptions of drone controllers and bystanders. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 6777–6788, New York, NY, USA, 2017.
- [147] Eric Zeng, Shirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, pages 65–80, Santa Clara, CA, 2017.
- [148] Eric Zeng and Franziska Roesner. Understanding and improving security and privacy in multi-user smart homes: A design exploration and in-home user study. In *USENIX Security Symposium (USENIX Security)*, pages 159–176, Santa Clara, CA, August 2019.
- [149] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home IoT privacy. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–20, 2018.
- [150] Nicole Zillien and Eszter Hargittai. Digital distinction: Status-specific types of Internet usage. *Social Science Quarterly*, 90(2):274–291, 2009.