Site Reliability Engineering and the Crisis/Complacency Cycle

LAURA NOLAN

Laura Nolan's background is in site reliability engineering, software engineering, distributed systems, and computer science. She wrote the "Managing Critical State" chapter in the O'Reilly Site Reliability Engineering book and was co-chair of SREcon18 Europe/Middle East/Africa. Laura is a production engineer at Slack. laura.nolan@gmail.com

This column will be published in Summer 2020, but I'm writing it in mid-March. In the past week, in a response to the spread of the new SARS-CoV-2 virus, many nations have closed down schools and imposed restrictions on travel and events. Several major technology companies are encouraging most employees to work from home. Stock markets are falling more quickly than in the first stages of the 2008 crash. Nothing is normal.

My social media feeds clearly show that SARS-CoV-2 is a source of fascination for systems engineers and SREs (site reliability engineers) because it has some characteristics of the kinds of systems problems we deal with in our work. The pandemic response is currently centered around preventing the spread of the infection, effectively an attempt to throttle admissions to intensive care in order to avoid saturating scarce medical resources. It involves gathering metrics (which are lagging and sparse due to shortage of test kits) to make analyses and projections. The mathematical analysis of the spread of the illness is very similar to the characteristics of information propagation in a dissemination gossip protocol [1], which will be familiar to anyone who has worked with Cassandra, Riak, Consul, or even BitTorrent—the major difference being that instead of modifying software parameters to adjust the propagation, we all now need to reduce our social interactions, and perhaps to partition our systems with travel restrictions.

I am not an epidemiologist, and I can't predict how this situation will unfold between now and when you read these words. Will we have endured on an international scale the kind of health crisis that northern Italy is experiencing in March, or will most nations succeed in averting the worst consequences of the pandemic, as South Korea seems to have done? If we do succeed, it's possible that many will consider the robust response to the outbreak to be an overreaction, even in light of the evidence from northern Italy and Wuhan that failure to control outbreaks leads to major public health problems.

The Job Is to Get Ahead of Problems

There is a phenomenon in operations, which I've heard called the "paradox of preparation"—an organization that is effectively managing risks and preventing problems can fail to be recognized as such. Bad outcomes aren't actually occurring, because of this preventative work, so decision-makers may come to believe that the risks are significantly lower than they actually are. Therefore, leaders may conclude that the organization that is preventing the negative events from occurring isn't an efficient use of resources anymore.

This appears to have been the fate of the White House's National Security Council Directorate for Global Health Security and Biodefense, which was set up in 2014 in response to the Ebola outbreaks in Western Africa, then shut down abruptly in 2018. It was tasked with monitoring emerging disease risks and coordinating responses and preparation. According to its former head, Beth Cameron, "The job of a White House pandemics office would have been to get ahead: to accelerate the response, empower experts, anticipate failures, and act quickly and transparently to solve problems" [2]. That is a function very much akin to what a good SRE or resilience engineering team can do within a software engineering organization.

In 2019, before the SARS-CoV-2 virus appeared, the Center for Strategic and International Studies think tank drew attention to the closure of the Directorate.

> When health crises strike—measles, MERS, Zika, dengue, Ebola, pandemic flu—and the American people grow alarmed, the U.S. government springs into action. But all too often, when the crisis fades and fear subsides, urgency morphs into complacency. Investments dry up, attention shifts, and a false sense of security takes hold. The CSIS Commission on Strengthening America's Health Security urges the U.S. government to replace the cycle of crisis and complacency that has long plagued health security preparedness with a doctrine of continuous prevention, protection, and resilience. [3]

This cycle of crisis and complacency is one we see in other kinds of organizations, including software companies—a view that reliability is only worth investing in the wake of problems, and at other times it may be deprioritized and destaffed. The last edition of this column discussed Professor Nancy Leveson's model of operations as a sociotechnical system dedicated to establishing controls over production systems in order to keep them within predefined safety constraints [4]. The crisis/complacency cycle makes it impossible to build a strong sociotechnical system that proactively manages change and emerging risks, because it means that when investment into reliability happens you have to build expertise, standards, processes, and organizations from scratch while already in crisis mode.

Against the Advice of Their Own Experts

This crisis/complacency cycle is not new, nor is it unique to either software or to pandemic prevention. The Boeing 737 Max has been in the news for most of the past year following two fatal crashes which were the consequence of design flaws in the new aircraft type. The entire 737 Max fleet was grounded in response to the accidents.

The airplane's design was certified by the US Federal Aviation Administration (FAA), a body created in 1958 to manage all aspects of safety in aviation. Air travel has become safer every decade since the FAA was set up, driven by improvements in technology and safety culture. Perhaps not coincidentally, the FAA has come under significant budgetary pressure in recent years. Partly as a result of those budgetary constraints and partly because of a shortage of relevant technical expertise, the FAA delegated much of the technical work of validating the design of the 737 Max aircraft against FAA standards to Boeing itself.

The report of the House Committee on Transportation and Infrastructure paints a clear picture of enormous pressure from Boeing's management to get the aircraft to the market as quickly as possible, at the lowest feasible cost and without any need for existing 737 pilots to take further training—regardless of any safety concerns [5]. Budgets for testing were cut, and multiple suggestions by engineers to incorporate extra alerts and indicators were rejected. Though it isn't in Boeing's commercial interest to develop an unsafe aircraft, the company's management consistently made decisions that compromised safety, contrary to the advice of their own technical experts. That they did this against the backdrop of the safest period in the history of commercial flight strongly suggests the same cycle of crisis and complacency was at work in Boeing and the FAA that led to the shutdown of the White House's pandemics office in 2018.

Disconnects between Management and Engineers

On January 28, 1986, the Space Shuttle Challenger exploded during liftoff. The accident was triggered by the failure of an O-ring seal in unusually cold weather conditions. The disaster occurred after 24 successful space shuttle launches, and these successes helped to create complacency about safety at NASA. The incident has been studied extensively, most notably by Diane Vaughan, who coined the term "normalization of deviance" to describe the process by which previously unacceptable results and practices can gradually become the norm over time [6]. Despite that phenomenon, the Rogers Commission Report on the disaster found that engineers had raised safety concerns over the design with management.

Richard Feynman, the noted physicist, was a member of the commission that investigated the Challenger accident. Feynman was particularly struck by the difference in perception of risk between the engineers who worked on the shuttle and NASA's management. The engineers mostly believed that the shuttle had a risk of catastrophic failure between 1 in 50 and 1 in 200. NASA's management claimed that the risk was 1 in 100,000. Feynman's assessment was that the engineers' estimate of the risk was far closer to the truth than management's number, which seemed based largely on wishful thinking and misunderstandings [7].

This kind of disconnect seems also to have existed at Boeing in recent years. In 2001, Boeing's executives moved from Seattle, where its engineers are located, to Chicago, and non-engineers moved into many executive roles.

> [T]he ability [was lost] to comfortably interact with an engineer who in turn feels comfortable telling you their reservations, versus calling a manager [more than] 1,500 miles away who you know has a reputation for wanting to take your pension away. It's a very different dynamic. As a recipe for disempowering engineers in particular, you couldn't come up with a better format. [8]

"Captain Hindsight Suited Up": Outcome Bias

Many of us in the software industry still remember the cautionary tale of Knight Capital, a financial firm that went bust in 2012 as a result of a bug in their trading software. As Knight Capital was an SEC (Securities and Exchange Commission) regulated company, there was an investigation and a report, which recommended that the company should have halted trading as soon as they realized there was something amiss [9].

On July 9, 2015, the New York Stock Exchange discovered a problem in their systems. They halted trading, just as the SEC said that Knight Capital ought to have done. However, as John Allspaw put it, the "clone army of Captain Hindsights suited up, ready to go" decided that the shutdown hadn't been essential and criticized the NYSE for unnecessarily halting over a "glitch" [10].

This is outcome bias, a cognitive bias that leads us to judge decisions based on their results. We can't predict the consequences of decisions perfectly at the time we make them. Many tough decisions have to be made with imperfect information—risks we can't fully quantify, information that's incomplete or missing. Sometimes, you need to make a sacrifice decision to avoid a risk of greater harm. This is likely better than simply reacting according to prevailing conditions of the crisis/complacency cycle. This closely describes the situation that the political leaders of most of the world find themselves in March 2020 with respect to SARS-CoV-2. By the time you read this, outcome bias will likely have declared their actions as overkill (if successful) or insufficient.

Risk and Rot in Sociotechnical Systems

We work in organizations made up of people, all subject to outcome bias and prone to underestimate or overestimate risks, depending on to what extent normalization of deviance has set in on our team. Executives can become far removed from the reality of life at the front line, and their appreciation of probabilities of adverse events can be strongly affected by recent outcomes.

One of the major functions of an SRE or operations team is to proactively manage risks. This kind of work covers a broad spectrum, from keeping systems patched, rotating certs and tokens, and validating backups, through to less routine things like writing runbooks and recovery tools, running disaster tests, performing production readiness reviews for new systems, and doing thorough reviews of near-miss production incidents. These are also the kinds of work that can fall by the wayside all too easily when a team is overloaded or understaffed. The eventual outcome is likely to be a crisis and the start of a new cycle of investment.

An important part of our job, therefore, is to make the value of our work visible in order to avoid the organizational rot that makes us underestimate risk and underinvest in reliability. We live in a data-driven world, but of course, we can't track the incidents that don't happen because of good preventative work. However, at times when we aren't in crisis mode, there are many other things that we can do to show how our work contributes to increasing reliability.

We can create internal SLOs for the routine jobs we do to manage risks, and set up dashboards to show whether you're meeting those SLOs or not. Write production-readiness standards that you'd like your services to meet—covering areas such as change management, monitoring and alerting, load balancing and request management, failover, and capacity planning. Track how your services meet those standards (or don't). Set up chaos engineering and game days to test how your services deal with failure, and track those results as you would postmortem action items. Load test your systems to understand how they scale, and address bottlenecks you will encounter in the next year or two. Take near-misses and surprises seriously, and track them, along with action items. All of these things help to prevent a slide into normalization of deviance as well as giving visibility into our work.

As engineers, we have a responsibility to clearly communicate about risks in our systems and the proactive work we do to reduce them. But "the fish rots from the head down": engineering leaders ultimately make critical decisions and therefore they must be acutely aware of outcome bias and the risk of disconnects in understanding of risk between front-line engineers and themselves. Most importantly, they must be mindful of the crisis/complacency cycle and maintain an appropriate continuous investment in resilience and reliability in order to avoid crisis.

References

[1] K. Birman, "The Promise, and Limitations, of Gossip Protocols": http://www.cs.cornell.edu/Projects/Quicksilver/public_pdfs/2007PromiseAndLimitations.pdf.

[2] B. Cameron, "I ran the White House pandemic office. Trump closed it," *The Washington Post,* March 13, 2020.

[3] J. S. Morrison, K. Ayotte, and J. Gerberding, "Ending the Cycle of Crisis and Complacency in U.S. Global Health Security," November 20, 2019, Center for Strategic International Studies: https://www.csis.org/analysis/ending-cycle-crisis-and-complacency-us-global-health-security.

[4] L. Nolan, "Constraints and Controls: The Sociotechnical Model of Site Reliability Engineering," *;login:*, vol. 45, no. 1 (Spring 2020), pp. 44–48.

[5] The House Committee on Transportation and Infrastructure, "Boeing 737 MAX Aircraft: Costs, Consequences, and Lessons from Its Design, Development, and Certification," March 2020: https://transportation.house.gov/imo/media/doc/TI%20Preliminary%20Investigative%20Findings%20Boeing%20737%20MAX%20March%202020.pdf.

[6] D. Vaughan, *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (University of Chicago Press, 1996).

[7] Presidential Commission on the Space Shuttle Challenger Accident, Report, 1986: https://science.ksc.nasa.gov/shuttle/missions/51-l/docs/rogers-commission/table-of-contents.html.

[8] J. Useem, "The Long-Forgotten Flight That Sent Boeing Off Course," *The Atlantic*, November 20, 2019: https://www.theatlantic.com/ideas/archive/2019/11/how-boeing-lost-its-bearings/602188/.

[9] "In the Matter of Knight Capital Americas LLC," SEC Release No. 70694, October 16, 2013: https://www.sec.gov/litigation/admin/2013/34-70694.pdf.

[10] J. Allspaw, "Hindsight and Sacrifice Decisions," March 3, 2019: https://www.adaptivecapacitylabs.com/blog/2019/03/03/hindsight-and-sacrifice-decisions/.