The Human Side of Postmortems

Managing Stress and Cognitive Biases

Dave Zwieback

O'REILLY®

O'REILLY® Velocity Web Performance and Operations

The Human Side of Postmortems

Managing Stress and Cognitive Biases

Dave Zwieback

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

The Human Side of Postmortems

by Dave Zwieback

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: (800) 998-9938 or *corporate@oreilly.com*.

May 2013: First Edition

Revision History for the First Edition:

2013-05-07: First release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449365851* for release details.

ISBN: 978-1-449-36585-1

Table of Contents

Acknowledgements

The author greatfully acknowledges the contributions of the following individuals, whose corrections and ideas made this article vastly better: John Allspaw, Gene Kim, Mathias Meyer, Peter Miron, Alex Payne, James Turnbull, and John Willis.

CHAPTER 1

What's Missing from Postmortem Investigations and Write-Ups?

How would you feel if you had to write a postmortem containing statements like these?

> "We were unable to resolve the outage as quickly as we would have hoped because our decision making was impacted by extreme stress."
>
> "We spent two hours repeatedly applying the fix that worked during the previous outage, only to find out that it made no difference in this one."
>
> "We did not communicate openly about an escalating outage that was caused by our botched deployment because we thought we were about to lose our jobs."

While these scenarios are entirely realistic, I challenge the reader to find many postmortem write-ups that even hint at these "human factors." A rare and notable exception might be Heroku's "Widespread Application Outage"[1] from the April 21, 2011, "absolute disaster" of an EC2 outage, which dryly notes:

> Once it became clear that this was going to be a lengthy outage, the Ops team instituted an emergency incident commander rotation of 8 hours per shift, keeping a fresh mind in charge of the situation at all time.

The absence of such statements from postmortem write-ups might be, in part, due to the social stigma associated with publicly acknowledging the contribution of human factors to outages. And yet, people

1. *http://bit.ly/KVKqB0*

dealing with outages are subject to physical exhaustion and psychological stress and suffer from communication breakdowns, not to mention impaired reasoning due to a host of cognitive biases.

What *actually* happens during and after outages is this: from the time that an incident is detected, imperfect and incomplete information is uncovered in *nonlinear*, chaotic bursts; the full outage impact is not always apparent; the search for "root causes" often leads down multiple dead ends; and not all conditions can be immediately identified and remedied (which is often the reason for repeated outages).

The omission of human factors makes most postmortem write-ups a peculiar kind of docufiction. Often as long as novellas (see Amazon's 5,694-word take on the same outage discussed previously in "Summary of the April 21, 2011 EC2/RDS Service Disruption in the US East Region"[2]), they follow a predictable format of the Three Rs[3]:

- **Regret** — an acknowledgement of the impact of the outage and an apology.
- **Reason** — a *linear* outage timeline, from initial incident detection to resolution, including the so-called "root causes."
- **Remedy** — a list of remediation items to ensure that this particular outage won't repeat.

Worse than not being documented, human and organizational factors in outages may not be sufficiently considered during postmortems that are narrowly focused on the technology in complex systems. In this paper, I will cover two additions to outage investigations — stress and cognitive biases — that form the often-missing human side of postmortems. How do we recognize and mitigate their effects?

2. *http://amzn.to/jFdKAR*

3. McFarlan, Bill. *Drop the Pink Elephant: 15 Ways to Say What You Mean… and Mean What You Say*. Capstone, 2009.

CHAPTER 2
Stress

What Is Stress?

Outages are stressful events. But what does *stress* actually mean, and what effects does it have on the people working to resolve an outage?

The term *stress* was first used by engineers in the context of stress and strain of different materials and was borrowed starting in the 1930s by social scientists studying the effects of physical and psychological stressors on humans[1]. We can distinguish between two types of stress: absolute and relative. Seeing a hungry tiger approaching will elicit a stress reaction — the fight-or-flight response — in most or all of us. This evolutionary survival mechanism helps us react to such absolute stressors quickly and automatically. In contrast, a sudden need to speak in front of a large group of people will stress out many of us, but the effect of this relative stressor would be less universal than that of confronting a dangerous animal.

More specifically, there are four relative stressors that induce a measurable stress response by the body:

1. A situation that is *interpreted* as novel.
2. A situation that is *interpreted* as unpredictable.
3. A *feeling* of a lack of control over a situation.

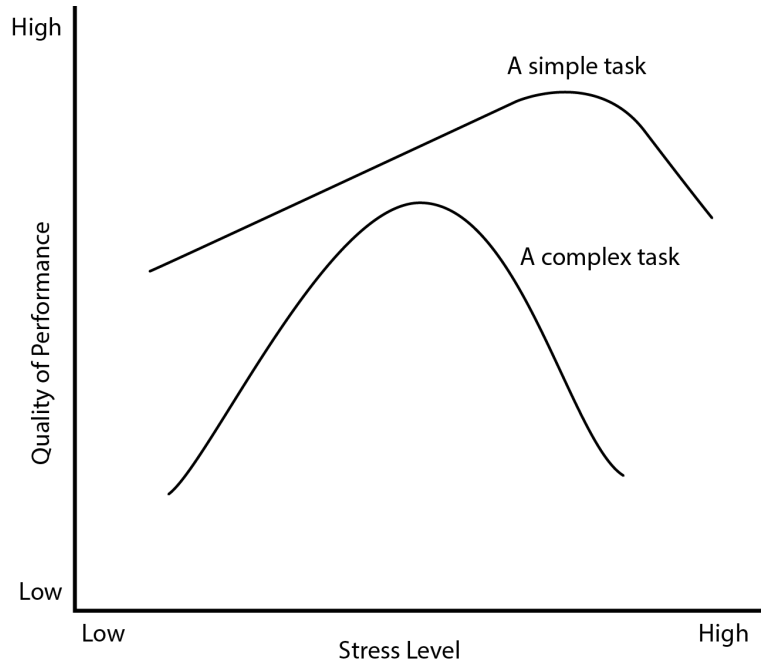
1. Lupien, Sonia J., F. Maheu, M. Tu, Al Fiocco, and T. E. Schramek. "The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition." Brain and Cognition 65, no. 3 (2007): 209-237.

4. A situation where one can be judged negatively by others (the "social evaluative threat").

While most outages are not life-or-death matters, they still contain combinations of most (or all) of the above stressors and will therefore have an impact on the people working to resolve an outage.

Performance under Stress

In 1908, the psychologists Robert Yerkes and John Dodson established a relationship between stress and performance. Although what is now known as the Yerkes-Dodson law was based on a less-than-humane experiment with a few dozen mice, subsequent research confirmed that it was "valid in an extraordinarily wide range of situations."[2]



Not all stress is bad. For instance, as you can see from the diagram above, low levels of stress are actually associated with low levels of performance. For example, it's unlikely that one will do one's best work of the day right after waking up, without taking steps to shake off the

2. Kahneman, Daniel. "Attention and effort." (1973).

grogginess (e.g., coffee, a morning run, and there's nothing like reading a heated discussion on Hacker News to get the heart rate up).

As stress increases, so does performance, at least for some time. This is the reason that a coach gives a rallying pep talk before an important sports event — a much-parodied movie cliché that can nonetheless improve team performance. Athletes are also often seen purposefully putting themselves in higher stress situations before competitions (for instance by playing loud music or warming up vigorously) in order to improve focus, motivation, and performance.

While the Yerkes-Dodson law applies universally, individuals exhibit a wide spectrum of stress tolerance. Some people are extraordinarily resilient to high levels of stress, and some of them naturally gravitate toward high-stress professions that involve firefighting (both the literal and figurative kinds). However, there is an inflection point for each individual after which additional stress will cause performance to deteriorate due to impaired attention and reduced ability to make sound decisions. The length of time that one is subject to stress also impacts the extent of its effects: playing Metallica's "Enter Sandman" at top volume might initially improve performance, but continued exposure will eventually weaken it. (Notably, this song has been used to put Guantánamo Bay detainees under extreme stress during interrogations[3].)

Simple vs. Complex Tasks

An important part of the Yerkes-Dodson law that is often overlooked is that simple tasks are much more resilient to the effects of stress than complex ones. That is, in addition to individual differences in stress resilience, the impact of stress on performance is also related to the difficulty of the task.

One way to think about "simple" tasks is that they are well-learned, practiced, and relatively effortless. For instance, one will have little difficulty recalling the capital of France, regardless of whether one is in a low- or high-stress situation. In contrast, "complex" tasks (like troubleshooting outages) are likely to be novel, unpredictable, or perceived as outside one's control. That is, complex tasks are likely to be subject to three of the four relative stressors mentioned above.

3. Stafford Smith, Clive. The Guardian, "Welcome to *the disco*." *http://bit.ly/oE3UM*.

With practice, complex tasks can become simpler. For instance, driving is initially a very complex task. Because learning to drive requires constant and effortful attention, one is unlikely to be playing "Harlem Shake" at top volume or casually chatting with friends at the same time. As we become more experienced, driving becomes more automatic and much less effortful, though we might still turn down the radio volume or pause our conversations when merging into heavy traffic. The good news is that increased experience in a particular task can make its performance more resilient to the effects of stress.

Stress Surface, Defined

The difficulty, of course, is finding precisely the point of an individual's optimal performance as it relates to stress during an outage. A precise measurement is impractical, since it would involve ascertaining the difficulty of the task, type and duration of stress, and would also have to account for individual differences in stress response.

A more pragmatic approach is to estimate the *potential* impact that stress can have on the outcome of an outage. To enable this, I'm introducing the concept of "stress surface," which measures the perception of the four relative stressors during an outage: the novelty of the situation, its unpredictability, lack of control, and social evaluative threat. These four stressors are selected because they are present during most outages, are known to cause a stress response by the body, and therefore have the potential to impact performance.

Stress surface is similar to the computer security concept of "attack surface" — a measure of the collection of ways in which an attacker can damage a system[4]. Very simply, an outage with a larger stress surface is more susceptible to the effects of stress than that with a smaller stress surface. As a result, we can use stress surface to compare the potential impact of stress on different outages as well as assess the impact of efforts to reduce stress surface over time.

To measure stress surface, we use a modified Perceived Stress Scale, the "most widely used psychological instrument for measuring the perception of stress":[5]

4. Manadhata, Pratyusa K. "Attack Surface Measurement." *http://bit.ly/10niL47*.

5. Cohen, Sheldon. "Perceived Stress Scale." *http://bit.ly/wmXLU8*.

The questions in this scale ask you about your feelings and thoughts during the outage. In each case, you will be asked to indicate how often you felt or thought a certain way.

0 = Never 1 = Almost Never 2 = Sometimes 3 = Fairly Often 4 = Very Often

During the outage, how often have you felt or thought that:

1. The situation was novel or unusual?
2. The situation was unpredictable?
3. You were unable to control the situation?
4. Others could judge your actions negatively?

We administer the above questionnaire as soon as possible after the completion of an outage. To prevent groupthink, all participants of the postmortem should complete the questionnaire independently. The overall stress surface score for each outage is obtained by summing the scores for all responses. A standard deviation should also be computed for the score to indicate the variance in responses.

Why measure stress surface? Knowing the stress surface score and asking questions like "What made this outage feel so unpredictable?" opens the door to understanding the effects of stress in real-world situations. Furthermore, one can gather data about the relationship of stress to the length of outages and determine if any particular dimension of the stress surface (for example, the threat of being negatively judged) remains stable between various outages. Most important, stress surface allows us to measure the results of steps taken to mitigate the effects of stress over time.

Reducing the Stress Surface

Two effective ways to reduce the stress surface of an outage are training and postmortem analyses. Specifically, conducting realistic game day exercises; regular disaster recovery tests; or, if operating in Amazon Web Services (AWS), surprise attacks of the Netflix Simian Army[6] — all followed by postmortem investigations — are effective in making

6. *http://nflx.it/q6fVuL*

outages less novel as well as exposing latent failure conditions. Moreover, developing so-called "muscle memory" from handling many outages (including practicing critical communication skills) can reduce the perceived complexity of tasks, making their performance more resilient to the effects of stress.

There has also been some promising research into Decision Support Systems (DSS), which have been used to improve decision making under stress in military and financial applications. In one case, researchers attached biometric monitors to bank traders, which alerted them when decision making was likely to be compromised due to high stress (measured by the stability of the frequency and shape of the heart rate waveform[7]). While DSS technology matures, organizations with awareness of the effects of stress on performance can take simple stress mitigation steps, for instance, by insisting on a "rotation of 8 hours per shift" during lengthy outages.

Why Postmortems Should Be Blameless

Unfortunately, these stress surface reduction steps do not address the effects of social evaluative threat in meaningful ways. That is especially troubling because, in my early investigations into stress surface, the component related to being negatively judged appears most stable between different outages and engineers.

Evaluative threat is social in nature — it involves both the organization's ways of dealing with failure (e.g., the extent to which blame and shame are part of the culture) and the individual's ability to cope with it. We should not dismiss the extent to which this stressor affects performance: several surveys have found that Americans are more afraid of public speaking, which is a classic example of social evaluative threat, than death[8]. Organizations where postmortems are far from blameless and where being "the root cause" of an outage could result in a demotion or getting fired will certainly have larger stress surfaces.

The most effective way of mitigating the effects of social evaluative stress is to emphasize the blameless nature of postmortems. What does

7. Martínez Fernández, Javier, Juan Carlos Augusto, Ralf Seepold, and Natividad Martinez Madrid. "Sensors in trading process: A Stress — Aware Trader." InIntelligent Solutions in Embedded Systems (WISES), 2010 8th Workshop, pp. 17-22. IEEE, 2010.

8. Garber, Richard I. "America's Number One Fear: Public Speaking - that 1993 Bruskin-Goldring Survey." Last modified May 19, 2011. *http://bit.ly/11KpT77*.

"blameless" actually mean? Very simply, *your organization must continually affirm that individuals are never the "root cause" of outages*. This can be counterintuitive for engineers, who can be quick to take responsibility for "causing" the failure or to pin it on someone else. In reality, blame is a shortcut, an intuitive jump to an incorrect conclusion, and a symptom of not going deeply enough in the postmortem investigation to identify the real conditions that enabled the failure, conditions that will likely do so again until fully remediated.

Making the effort to become more accepting of failure at an organizational level, and more specifically making postmortems "blameless," is not a new-age feel-good measure done intuitively in "evolved" organizations. It is rooted in the understanding of the real conditions of failure in complex systems and a concrete way to improve performance during outages by reducing their stress surface.

The Limits of Stress Reduction

Of course, no amount of training or experience can reduce the stress surface to zero — outages will continue to surprise (and to some extent delight) in novel, unpredictable ways. A true mark of an expert is a realistic and humble assessment of the limitations of experience and the extent to which control over complex systems is actually possible. In contrast, less mature engineers tend to develop overconfidence in their own abilities after some initial success and familiarly with systems. This is not endemic to engineers: despite overwhelming evidence that inexperience is one of the main causes of accidents in young drivers, they consistently fail to judge the extent of their own inexperience and how it affects their safety[9]. We'll cover overconfidence and other biases in more detail later in this paper.

Caveats of Stress Surface Measurements

In a poll of 2,387 U.S. residents, the mean male and female Perceived Stress Scale scores (12.1 and 13.7, respectively) had fairly high stan-

9. Ginsburg, Kenneth R., Flaura K. Winston, Teresa M. Senserrick, Felipe García-España, Sara Kinsman, D. Alex Quistberg, James G. Ross, and Michael R. Elliott. "National young-driver survey: teen perspective and experience with factors that affect driving safety." Pediatrics 121, no. 5 (2008): e1391-e1403.

dard deviations (5.9 and 6.6, respectively)[10]. We can expect a similarly high variance in stress surface measurements, in part due to the individual differences in perception of stress.

We should also remember that stress surface scores are based on a *memory* of feelings and thoughts during a stressful event. There are many conditions that could influence the ability of individuals to faithfully recall their experiences, including the duration of time that has passed since the event as well as the severity of stress they experienced during an outage. Furthermore, our recollections are likely colored by hindsight bias, which is our tendency to remember things as more obvious than they appeared at the time of the outage.

Finally, stress surface measurements in smaller teams may be subject to the Law of Small Numbers. As Daniel Kahneman warns in *Thinking, Fast and Slow*:[11]

- The exaggerated faith in small samples is only one example of a more general illusion — we pay more attention to the content of messages than to information about their reliability, and as a result, end up with a view of the world around us that is simpler and more coherent than the data justify. Jumping to conclusions is a safer sport in the world of our imagination than it is in reality.
- Statistics produce many observations that appear to beg for causal explanations but do not lend themselves to such explanations. Many facts of the world are due to chance, including accidents of sampling. Causal explanations of chance events are inevitably wrong.

Nevertheless, obtaining the stress surface score for each outage is an effective way to frame the discussion of the effects of stress, including identifying ways they can be mitigated.

10. Cohen, Sheldon. "Perceived Stress Scale." *http://bit.ly/wmXLU8*.

11. Kahneman, Daniel. Thinking, fast and slow. Farrar, Straus and Giroux, 2011.

CHAPTER 3

Cognitive Biases

The Benefits and Pitfalls of Intuitive and Analytical Thinking

To further quote Kahneman, "the law of small numbers is a manifestation of a general bias that favors certainty over doubt." What other biases affect people working with complex systems? And how can they be overcome?

To begin our discussion of cognitive biases, we should introduce the theory of two different systems (or types) of thinking. Perhaps the best way to do it is through examples:

> What is 2+2?
>
> What is the capital of France?

Did you notice that in the above examples, you arrived at the answers (4 and Paris, respectively) quickly, automatically, and without effort? In fact, there was almost nothing you could do to stop the answers from appearing in your mind. This fast, intuitive, and largely automatic thinking is known as System 1 thinking. This type of thinking is largely based on associative memory and experience. This is the thinking celebrated in Malcom Gladwell's 2005 book *Blink: The Power of Thinking Without Thinking*. A famous example from *Blink* is that of the Getty kouros, a statue bought by the Getty Museum. Despite the statue's credible documentation, expert archeologists identified it as fake, seemingly at a glance and despite their inability to specify the exact reasons.

Now, let's illustrate the other (System 2) thinking:

> What is 367 x 108?

Unless you've memorized the answer previously, or have made a career of performing feats of mental math, it did not automatically come to mind and it took some time to calculate (the answer is 39,636). To do so, you used your System 2 thinking, which, compared to System 1 thinking, is slower, non-automatic, and effortful. In fact, it uses more energy (glucose) than System 1 thinking. As a result, we spend much of our time relying on the more energy-efficient System 1 thinking. System 1 is also where we fall under stress.

The two systems of thinking are not separate, and there's no reason to look down on the lowly System 1 thinking. This is the thinking that allows us to do marvelous things like drive a car and listen to music or hold a conversation at the same time, to quickly determine if our partner is upset after the first few moments of a phone conversation, or enables an experienced firefighter to save his men by pulling them out of a dangerous fire based on a sudden "gut feeling" that something is wrong[1].

However, System 1 thinking has two major shortcomings. First, it is rooted in memory and experience. Unless we've trained for years as archeologists, and have looked at literally thousands of archeological artifacts, we won't be able to "take in" a new artifact and quickly determine its authenticity. Similarly, it is the many years of seeing complex systems function and fail that allows experienced operations people to quickly identify and deal with conditions of an outage. The second shortcoming of System 1 thinking is that it is prone to making systematic mistakes, which are called cognitive biases. Our preference for System 1 thinking, especially in stressful situations, can increase the effects of cognitive biases during outages.

Jumping to Conclusions

Consider this question:

> If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

This is one of the three questions on a Cognitive Reflection Test — questions that were selected because they reliably evoke an immediate

1. Gladwell, Malcolm. Blink: The power of thinking without thinking. Back Bay Books, 2007.

and *incorrect* answer[2]. You are in good company if your quick and intuitive answer is "100 minutes" — 90% of participants of an experiment involving Princeton students answered at least one of the three CRT questions incorrectly. Still, the unintuitive and correct answer is "5 minutes."

What's going on here? As we've seen, System 1 thinking quickly and efficiently provides what you might call a first approximation assessment of a situation, or the effortlessly intuitive answer to the question above. And in the vast majority of cases, System 1 thinking functions superbly. For instance, being able to quickly spot something that looks like a leopard is an important function of System 1. When System 1 produces a mistake — if, for instance, what looks like a leopard turns out to be an old lady in a leopard-print coat — from an evolutionary point of view, it may be vastly better to be wrong while quickly running away than to be mauled by a hungry leopard while taking the time to thoroughly analyze the potential attacker. That is, unless you, as a result, quickly run into crosstown traffic, in which case it would have been far better to slow down and actually evaluate the probability of meeting a leopard in midtown Manhattan!

System 1 is expert at quickly jumping to conclusions. It does so by employing mental shortcuts — heuristics — "which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors,"[3] otherwise known as cognitive biases.

A Small Selection of Biases Present in Complex System Outages and Postmortems

There are more than 100 cognitive biases listed in Wikipedia[4], and Daniel Kahneman's epic-yet-accessible treatment of the subject (*Thinking, Fast and Slow*) weighs in at more than 500 pages.

2. Kahneman, Daniel. Thinking, fast and slow. Farrar, Straus and Giroux, 2011.
3. Tversky, Amos, and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. Springer Netherlands, 1975.
4. *http://bit.ly/985JMi*

Both the number of biases and our understanding of them is growing, as they have been the subject of considerable research since they were first identified by Kahneman and his research partner Amos Tversky in the early 1970s.

The following discussion will give the reader familiarity with some of the more "classic" biases that are usually present during outages and postmortems.

Hindsight Bias

In early 1999, the co-founders of Google, having raised a mere $100,000 to date, attempted to sell it to more established search companies Yahoo! and Excite so they could return to their graduate studies. [5] What is your estimate of the asking price?

Clearly, the time at which the above question was posed would have an effect on your answer. If you were asked in early 1999, would you have guessed that $1 million was a reasonable estimate? Had you been asked the same question in 2012, would you have thought that Google could have been as valuable to Yahoo! as Instagram was to Facebook, and therefore worth $1 billion (or $725 million, adjusted for inflation)?

Moreover, given what we know about Google's current valuation (more than $266 billion) as well as the dwindling fortunes of Yahoo! and Excite, it appears absolutely clear that these companies missed a huge opportunity by not purchasing Google for the actual asking price of $1 million (or less).

The above statement is an example of several biases in action, most prominently the hindsight bias. This bias affects not just casual observers, but professionals as well. For example, here is how Paul Graham, an investor at the Y Combinator startup incubator, views the same situation:

> Google's founders were willing to sell early on. They just wanted more than acquirers were willing to pay … Tip for acquirers: when a startup turns you down, consider raising your offer, *because there's a good chance the outrageous price they want will later seem a bargain* [6].

5. Siegler, MG. TechCrunch. "When Google Wanted To Sell To Excite For Under $1 Million — And They Passed". *http://tcrn.ch/ctS4eM*.

6. Graham, Paul. "Why There Aren't More Googles." *http://bit.ly/z3zoX*.

In retrospect, buying Google for $1 million in 1999 certainly looks like a fantastic investment. However, Google's success was far from certain in 1999, and given that 75% of startups fail[7], was it really as good a chance as Graham seems to think?

During postmortems we evaluate what happened during an outage with the benefit of currently available information, i.e., hindsight. As we aim to identify the conditions that were necessary and sufficient for an outage to occur, we often uncover things that could have prevented or shortened the outage. We hear statements like "You shouldn't have made the change without backing up the system first" or "I don't know how I overlooked this obvious step" from solemn postmortem participants. Except that these things were as far from obvious during the outage as Google's 2013 valuation was in 1999!

Outcome Bias

When the results of an outage are especially bad, hindsight bias is often accompanied by outcome bias, which is a major contributor to the "blame game" during postmortems. Because of hindsight bias, we first make the mistake of thinking that the correct steps to prevent or shorten an outage are equally obvious before, during, and after the outage. Then, under the influence of outcome bias, we judge the quality of the actions or decisions that contributed to the outage in proportion to how "bad" the outage was. The worse the outage, the more we tend to blame the human committing the error — starting with overlooking information due to "a lack of training," and quickly escalating to the more nefarious "carelessness," "irresponsibility" and "negligence." People become "root causes" of failure, and therefore something that must be remediated.

The combined effects of hindsight and outcome bias are staggering:

> Based on an actual legal case, students in California were asked whether the city of Duluth, Minnesota, should have shouldered the considerable cost of hiring a full-time bridge monitor to protect against the risk that debris might get caught and block the free flow of water. One group was shown only the evidence available at the time of the city's decision; 24% of these people felt that Duluth should take on the expense of hiring a flood monitor. The second group was informed that debris had blocked the river, causing major flood damage;

7. Xavier, Jon. Silicon Valley Business Journal, "75% of startups fail, but it's no biggie." *http://bit.ly/QGUSdC*.

> 56% of these people said the city should have hired the monitor, *although they had been explicitly instructed not to let hindsight distort their judgment* [8].

Outcome bias is also implicated in the way we perceive risky actions that appear to have positive effects. As David Woods, Sidney Dekker and others point out, "good decision processes can lead to bad outcomes and good outcomes may still occur despite poor decisions"[9]. For example, if an engineer makes changes to a system without having a reliable backup and this leads to an outage, outcome bias will help us quickly (and incorrectly) see these behaviors as careless, irresponsible, and even negligent. However, if no outage occurred, or if the same objectively risky action resulted in a positive outcome like meeting a deadline, the action would be perceived as far less risky, and the person who took it might even be celebrated as a visionary hero. At the organizational level, there is a real danger that unnecessarily risky behaviors would be overlooked or, worse yet, rewarded.

Availability Bias

Residents of the Northeast United States experience electricity outages fairly frequently. While most power outages are brief and localized, there have been several massive ones, including the blackout of August 14-15, 2003[10]. Because of the relative frequency of such outages, and the disproportionate attention they receive in the media, many households have gasoline-powered backup generators with enough fuel to last a few hours. In late October 2012, in addition to lengthy power outages, Hurricane Sandy brought severe fuel shortages that lasted for more than a week. Very few households were prepared for an extended power outage *and* a gasoline shortage by owning backup generators *and* stockpiling fuel.

This is a demonstration of the effects of the availability bias (also known as the recency bias), which causes us to overestimate (sometimes drastically) the probability of events that are easier to recall and underestimate that of events that do not easily come to mind. For instance, tornadoes (which are, again, heavily covered by the media)

8. Kahneman, Daniel. Thinking, fast and slow. Farrar, Straus and Giroux, 2011.
9. Woods, David D., Sidney Dekker, Richard Cook, Leila Johannesen, and N. B. Sarter. "Behind human error." (2009): 235.
10. *http://en.wikipedia.org/wiki/Northeast_blackout_of_2003*

are often perceived to cause more deaths than asthma, while in reality asthma causes 20 times more deaths.[11]

In the case of Hurricane Sandy, since the median age of the U.S. population is 37, the last time fuel shortages were at the top of the news (in 1973-74 and 1979-80) was before about half of the U.S. population was born, so it's easy to see how most people did not think to prepare for this eventuality. Of course, the hindsight bias makes it obvious that such preparations were necessary.

The availability bias impacts outages and postmortems in several ways. First, in preparing for future outages or mitigating effects of past outages, we tend to consider scenarios that appear more likely, but are, in fact, only easier to remember either because of the attention they received or because they occurred recently. For instance, due to its severity, many organizations utilizing AWS vividly remember the April 21, 2011, "service disruption" mentioned previously and have taken steps to reduce their reliance on the Elastic Block Store (EBS), the network storage technology at the heart of the lengthy outage. While they would have fared better during the October 22, 2012, "service event" also involving EBS, these preparations would have done little to reduce the impact of the December 24, 2012, outage, which affected heavy users of the Elastic Load Balancing (ELB) service, like Netflix.

Furthermore, especially under stress, we often fall back to familiar responses from prior outages, which is another manifestation of the availability bias. If rebooting the server worked the last N times, we are likely to try that again, especially if the initial troubleshooting offers no competing narratives. In general, not recognizing the differences between outages could actually make the situation worse.

Although much progress has been made in standardizing system components and configurations, outages are still like snowflakes, gloriously unique. Most outages are independent events, which means that past outages have no effect on the probability of future outages. In other words, while experience with previous outages is important, it can only go so far.

11. Kahneman, Daniel. Thinking, fast and slow. Farrar, Straus and Giroux, 2011.

Other Biases and Misunderstandings of Probability and Statistics

Most of us are terrible at *intuitively* grasping probabilities of events. For instance, we often confuse independent events (e.g., the probability of getting "heads" in a coin toss remains 50% regardless of the number of tosses) from dependent ones (e.g., the probability of picking a marble of a particular color changes as marbles are removed from a bag). This sometimes manifests as *sunk cost bias*, for example, when engineers are unwilling to try a different approach to solving a problem even though a substantial investment in a particular approach hasn't yielded the desired results. In fact, they are likely to exclaim "I almost have it working!" and further escalate their commitment to the non-working approach. This can be made worse by the *confirmation bias*, which compels us to search for or interpret information in a way that confirms our preconceptions.

At other times, intuitive errors in understanding of statistics result in finding illusory correlations (or worse, causation) between uncorrelated events — e.g., "every outage that Jim participates in takes longer to resolve, therefore the length of outages must have some relation to Jim." Similarly, because large outages are relatively rare, we can become biased due to the Law of Small Numbers — e.g., "this outage is likely to look like the last outage."

Finally, we are often overly confident in our decision-making abilities. This overconfidence bias manifests most clearly and dangerously when two nations are about to go to war, and their estimates of winning often sum to greater than 100% (i.e., "both think they have more than a 50% chance of winning"). Similarly, the positive "can do" attitude on display during outages is a symptom of overconfidence in our abilities to control the situation over which, in reality, we have little or no control (think: public cloud). There's certainly nothing wrong with maintaining a positive attitude during a stressful event, but it's worth keeping in mind that confidence is nothing but a feeling that is "determined mostly by the coherence of the story and by the ease with which it comes to mind, even when the evidence for the story is sparse and unreliable"[12].

12. Kahneman, Daniel. New York Times, "Don't Blink! The Hazards of Confidence." *http://www.nytimes.com/2011/10/23/magazine/dont-blink-the-hazards-of-confidence.html*.

Reducing the Effects of Cognitive Biases, or "How Do You Know That?"

Cognitive biases are a function of System 1 thinking. This is the thinking that produces quick, efficient, effortless, and intuitive judgments, which are good enough in most cases. But this is also the thinking that is adept at maintaining cognitive ease, which can lead to mistakes due to cognitive biases. The way that we can reduce the effects of cognitive biases is by engaging System 2 thinking in an effortful way. Even so:

> biases cannot always be avoided because System 2 may have no clue to the error … The best we can do is a compromise: learn to recognize situations in which mistakes are likely and try harder to avoid significant mistakes when the stakes are high[13].

We've discussed the effects of stress on performance, and we should emphasize again that we tend to slip into System 1 thinking under stress. This certainly increases the chances of mistakes that result from cognitive biases during and after outages. So what can we do to invoke System 2 thinking, which is less prone to cognitive biases, when we need it most?

We don't typically have the luxury of knowing when our actions might become conditions for an outage or when an outage may turn out to be especially widespread. However, before working on critical or fragile systems — or, in general, before starting work on large projects — we can use a technique developed by Gary Klein called the PreMortem. In this exercise, we imagine that our work has resulted in a spectacular and total fiasco, and "generate plausible reasons for the project's failure"[14]. Discussing cognitive biases in PreMortem exercises will help improve their recognition — and reduce their effects — during stressful events.

It's often easier to recognize other people's mistakes than our own. Working in groups and openly asking the following questions can illuminate people's quick judgments and cognitive biases at work:

> How is this outage different from previous outages?
>
> What is the relationship between these two pieces of information — causation, correlation, or neither?

13. Kahneman, Daniel. Thinking, fast and slow. Farrar, Straus and Giroux, 2011.
14. Klein, Gary. Harvard Business Review. "Performing a Project Premortem." *http://hbr.org/2007/09/performing-a-project-premortem/ar/1*

What evidence do we have to support this explanation of events?

Can there be a different explanation for this event?

What is the risk of this action? (Or, what could possibly go wrong?)

Edward Tufte, who's been helping the world find meaning in ever-increasing volumes of data for more than 30 years, suggests we view evidence (e.g., during an outage) through what he calls the "thinking eye," with:

> bright-eyed observing curiosity. And then what follows after that is reasoning about what one sees and asking: what's going on here? And in that reasoning, intensely, it involves also a skepticism about one's own understanding. The thinking eye must always ask: How do I know that? *That's probably the most powerful question of all time. How do you know that?*" [15]

15. Tufte, Edward. "Edward Tufte Wants You to See Better." Talk of the Nation, by Flora Lichtman. *http://www.npr.org/2013/01/18/169708761/edward-tufte-wants-you-to-see-better*.

CHAPTER 4
Mindful Ops

Relative stressors and cognitive biases are both mental phenomena — thoughts and feelings — which nonetheless have concrete effects on our physical world, whether it is the health of operations people or the length and severity of outages. The best way to work with mental phenomena is through mindfulness. Mindfulness has two components:

> The first component involves the self-regulation of attention so that it is maintained on immediate experience, thereby allowing for increased recognition of mental events in the present moment. The second component involves adopting a particular orientation toward one's experiences in the present moment, an orientation that is characterized by curiosity, openness, and acceptance.[1]

One of the challenges with mitigating the effects of stress is the variance in individual responses to it. For instance, there is no known method to objectively determine the level of social evaluative threat that is harmful for a particular individual. Measuring stress surface, vital signs or stress hormone levels are, at best, proxies for — and approximations of — the real effects of stress. However, by practicing mindfulness, an individual can learn to recognize when they're experiencing (subjectively) harmful levels of stress and take simple corrective actions (e.g., take a break or ask for a second opinion in a high-risk situation). Mindfulness-Based Stress Reduction (MBSR) — a "meditation program created in 1979 from the effort to integrate Bud-

1. Bishop, Scott R., Mark Lau, Shauna Shapiro, Linda Carlson, Nicole D. Anderson, James Carmody, Zindel V. Segal et al. "Mindfulness: A proposed operational definition." Clinical psychology: Science and practice 11, no. 3 (2004): 230-241.

dhist mindfulness meditation with contemporary clinical and psychological practice" — is known to significantly reduce stress[2].

We can similarly mitigate the effects of cognitive biases through mindfulness — we can become aware of when we're jumping to conclusions and purposefully slow down to engage our analytical System 2 thinking.

The practice of mindfulness requires some effort, but is also simple, free, and without negative side effects. As we've seen, increased mindfulness — Mindful Ops — can reduce the effects of stress and cognitive biases, ultimately help us build more resilient systems and teams, and reduce the duration and severity of outages.

2. Chiesa, Alberto, and Alessandro Serretti. "Mindfulness-based stress reduction for stress management in healthy people: a review and meta-analysis." The journal of alternative and complementary medicine 15, no. 5 (2009): 593-600.

Author's Note

Meditation and mindfulness are huge subjects that we've barely begun to explore in this paper. I sincerely encourage the reader to investigate and experience their benefits in their work and life. The works of Thich Nhat Hanh, Jon Kabat-Zinn, Matthieu Ricard, or Sharon Salzberg (among others) are great places to get started.

About the Author

Dave Zwieback has been managing large-scale mission-critical infrastructure and teams for 17 years. He is the CTO of Lotus Outreach. He was previously the head of infrastructure at Knewton, managed UNIX Engineering at D.E. Shaw & Co., and managed enterprise monitoring tools at Morgan Stanley. He also ran an infrastructure architecture consultancy for seven years. Follow Dave *@mindweather* or on his website, *mindweather.com*.