Editors: Patrick McDaniel, mcdaniel@cse.psu.edu; Sean W. Smith, sws@cs.dartmouth.edu

Security and Cognitive Bias: Exploring the Role of the Mind

Sean W. Smith | Dartmouth College

Computer security aims to ensure that only "good" behavior happens in computer systems, despite potential action by malicious adversaries. Consequently, practitioners have focused primarily on the technology to prohibit "bad" things—according to some set of rules—and to a lesser extent on the structure of such rules.

Unfortunately, fieldwork and anecdotes report how we continue to get the rules wrong. We keep hearing that security is hard to use and gets in the way. In the workplace, writing down passwords on Post-it notes hidden under keyboards, under tables, or in desk drawers is endemic, because humans have too many to remember—and perhaps also because the IT system forces an authentication system that doesn't meet users' needs. (Recently, a school system secretary was lambasted for misusing the superintendent's password to change grades—but no one seemed to think it odd that she knew the password in the first place.[1]) IT staffs know that keeping software updated is important to patch holes, but balancing those updates while keeping mission-critical applications running unimpaired is tricky—many users just give up. (Stuxnet was lauded for the number of 0-day holes it used, but five-year holes would suffice to penetrate much of our information infrastructure.) Savvy home users, trying to (legally) share music files with another household computer, will struggle over drop-down menu options attempting to open only the proper holes in the network perimeter. Developers might know that advanced protection technology, such as SELinux, will help keep programs in the bounds of secure behavior, but they have no easy way of formally telling the system what those bounds are.

So, it's hard to create and configure security technology and hard to use it after deployment. However, the charter of this department is to look at the broader "system" context of security—and the human mind is a component in both security creation and use. The human mind is the arena in which security engineers translate "goodness" to machine rules; it's where users experience frustration and is the medium through which that frustration is conveyed.

While we practitioners have spent the last 40 years building fancier machines, psychologists have spent those decades documenting ways in which human minds systematically (and predictably) misperceive things. Minds are part of the system, and cognitive biases tell us how minds get things wrong. (For quick introductions to this field, see *Rational Choice in an Uncertain World*, an undergraduate-level textbook;[2] *Cognitive Illusions*, a graduate-level book;[3] or *Stumbling on Happiness*, more casual reading.[4] A pioneer in this space, Daniel Kahneman—a Nobel laureate—also has a new book out, *Thinking, Fast and Slow*, for a general audience.[5])

To What Extent Might This Affect the Usable Security Problem?

Consider the creation of security policies—the formal rules stating whether subject S can perform action A on object O right now (let's call this time t_1). It's tempting to imagine that an omniscient deity hovers in the computer, looking at a request's full context and implications and making the wisest possible decision. However, in reality, this decision was probably made much earlier in time (at a time $t_0 \ll t_1$) by a security officer trying to imagine what S would be doing in the future and whether action A would be consistent with the organization's goals and values. We can pretend that the policy rules came from the deity at

t_1, but it was all in the officer's head at t_0. Cognitive bias can tell us how these rules might differ. If we don't pay attention to this difference, we risk creating incorrect policies.

Alternatively, consider the case of a subject S complaining about unusable security features (or, for that matter, other unusable aspects of IT). It's tempting to imagine that an omniscient deity is hovering in S's mind, who recorded this bad experience at time t_1. However, in reality, we have a security engineer hearing S's recollections, at time $t_2 \gg t_1$, of how S felt at t_1. We can pretend these recollections are the same as the deity's observations, but they were all filtered through S's head. Cognitive bias can tell us how the recollections and observations might differ. In this case, if we don't pay attention to this difference, we risk "fixing" the wrong thing.

The Dual-Process Model

In my lab at Dartmouth, my colleagues and I have performed some initial exploration into how two sources of cognitive bias—the dual-process model and the empathy gap—affect security policy creation.

The dual-process model partitions the mind into two parts: an intuitive, nonverbal, and almost nonconscious *system 1*, and a verbal, introspective, conscious *system 2*. Some tasks are better done by one system or the other, and the systems can interfere with each other. However, this isn't just abstract theory; what makes the last few decades of this science so interesting for people like me is that these theories are reinforced by experiments. We can use the theories to make predictions that are borne out in practice!

For example, psychologists Timothy Wilson and Jonathan Schooler carried out some experiments regarding jam (and by "jam," I mean the sweet condiment one puts on toast, not an obscure security acronym).[6] Trained taste experts ranked a set of jams. One set of test subjects ranked the jams without thinking (that is, using system 1); their rankings closely correlated with the experts' rankings. However, other sets of test subjects were asked to think carefully while ranking the jams—and their rankings were very different. Nonexperts could do the task with system 1 but not with system 2. For jam, introspection inhibits intuition.

Sticky jam made me think of sticky security policy problems. We technologists build elaborate sets of knobs—drop-down menus, check boxes, access control lists—and expect users to figure out how to map their notion of "goodness" to a setting of the knobs, perhaps moving a system 1 goal to a system 2 task. Might we see the same inhibition phenomenon here? To test this, my team created a fictional social network. Users had various categories of personal information, and the GUI told users the various levels of connection they had with each friend. We presented one group of test subjects a sequence of friends and asked them to decide which information they'd share with each friend. Another group was asked to think about various social network privacy issues, and then given the same choices. The second group made significantly different choices—but to our surprise, the difference was one-sided: the group asked to think about privacy gave more information away![7] Perhaps introspection inhibits intuition also when it comes to security policy. (In hindsight, I wonder whether the cognitive bias toward dissonance reduction might have been at play; maybe the results would have differed if we didn't call them "friends.")

The Empathy Gap

We also examined what psychologists call the empathy gap: the very different decisions people make, even about dry factual things such as an estimated selling price for a coffee mug, when they are in the situation themselves versus when they are speculating about themselves in the future or about someone else.[8–10] In our fieldwork in access control in large enterprises, we kept hearing how users needed to work around the access control system because the policy didn't allow them to do what they needed to perform their jobs. In the case of healthcare IT, some researchers have even reached the conclusion that the problem is a dearth of clinicians among the policy makers.

Could the empathy gap be playing a role here? To examine this question, we recruited nearly 200 clinicians and staff members at a large hospital and partitioned them into two groups.[11] We gave one group a series of access control scenarios we developed with a medical informatics specialist. These scenarios were all phrased in an abstract, role-based way, as is often found in security policies (for example, "Should a physician be able to see information I about patient A in this particular context?"). We gave the other group the same scenarios but instead phrased them in a way that put the test subject directly in the setting; each wildcard became specific (for example, "You are a physician treating patient Alice…").

For two-thirds of the scenarios, the direct-experience group made significantly looser judgments than the policy maker group, suggesting that even experienced medical staff will make access control policies that experienced medical staff will find overly constraining. (However, in some of the other scenarios, the direct-experience group made significantly tighter decisions, oddly.) Maybe the problem with policy creation isn't the policy makers' backgrounds but the cognitive bias built into human minds.

Bounded Rationality and the Anchoring Effect

At the University of Southern California, Milind Tambe has also been looking at the role of cognitive bias, but in the context of optimizing system defense against human adversaries. In these scenarios, defenders have a limited amount of resources to distribute across various targets. Before mounting their attack, the adversaries can make repeated observations of the defenders' actions. For instance, defenders distribute guards across a certain number of airport terminals, and the adversary can quietly scope things out and see that, perhaps, the guards go to the odd-numbered terminals on odd-numbered days and even-numbered terminals on even-numbered days.

The branch of mathematics called *game theory* analyzes these scenarios as a special type of Stackelberg game. Formalized treatments establish a set of possible adversaries, under a known distribution, and assume that adversaries choose the attack strategies that maximize their expectations of success. Under this formalized model, with "perfectly rational" adversaries, optimal strategies exist for the defenders.

In the real world, it might be nice to pretend that adversaries are perfectly rational—but in fact, they're human, with minds subject to biases and distortions. In a 2009 project, James Pita and colleagues considered the implications of two of these biases: *bounded rationality* and the *anchoring effect*.[12]

Computer scientists like to think about how problems are solved by precise, thorough algorithms. The concept of bounded rationality (attributed to Herb Simon, whom the computer science field claims as one of its own) arises from the annoying observation that, when approaching many of these problems, human minds show no evidence of actually carrying out these algorithms, and so are perhaps doing something much simpler and less correct. Tambe's group allowed for the adversaries' bounded rationality by allowing them to have only approximately optimal choices.

Looking back on the inspiration for his pioneering work in cognitive bias, Kahneman tells how he and his colleague would consistently misestimate statistical probabilities—but would misestimate the same way! (So maybe we can predict how humans get these things wrong.) The anchoring effect describes one type of distortion here: generally put, human minds like to make basic assumptions about probability distributions and only slowly change them on the basis of observation.

It might be nice to pretend that adversaries are perfectly rational—but in fact, they're human, with minds subject to biases and distortions.

As noted, formal treatments of the defender game assume adversaries make the best possible choice against the defender's strategy. An omniscient adversary sees the defender's strategy exactly; however, human adversaries can only act on their perceptions of the strategy. Tambe's group modeled this effect by initially anchoring the adversaries' perceptions on uniformity in the defender's resource distribution, regardless of what the defender was doing.

Making these changes in the adversary model leads to defender strategies that differ from what was previously considered mathematically optimal. The punch line? When evaluated in large-scale experiments against human adversaries, these new strategies outperformed the mathematically optimal ones! In subsequent work, Pita's group further improved their model by taking into account *prospect theory*, which describes how human minds tend to distort estimated probabilities of actions depending on how good or bad the perceived outcome is.[13]

Some Other Cognitive Bias Techniques

These study results stemmed from looking at a few basic ways the mind gets things wrong. However, the literature on cognitive biases provides a veritable wonderland of additional techniques. Here are just a few:

- *Peak end* (for example, see "End Effects of Rated Life Quality"[14]). Rather than considering the net amount of goodness over time, human minds measure the quality of an event with duration by considering just the maximum value and the end value. Humans judge a short, happy life to be better than the same life with a longer but not quite as happy tail. Perhaps we can make an unusable security system appear more usable just by making it end well.
- *Immune neglect* (for example, see "The Particular Longevity of Things Not So Bad"[15]). Scenarios exist in which less-bad events can have a longer negative impact (when recalled by human minds) than worse events. Perhaps we can make an unusable security system appear more usable (afterward) by making things go really wrong when they start to go wrong. Rather than simply reject a password, maybe we should crash the browser.
- *Preview-based forecasting* (for example, see "Why the Brain Talks to Itself"[16]). Humans evaluate future choices by "previewing"

their consequences in their heads. However, psychologists have identified various sources of systematic error in such previews. Perhaps this can tell us how to make a security policy tool (predicting the goodness of future actions) that creates a policy users are less likely to circumvent.

- *"Infernal" internal logic* (for example, see "Supposition and Representation in Human Reasoning"[17]). Human minds have interesting ways of drawing incorrect conclusions from a set of assertions and observations (for example, Google the "Wason selection task"). Perhaps this might shed light on how even shrewd Unix users have trouble setting file and directory permissions correctly for various scenarios. (Think of "access" as "conclusion," and "rules/settings" as "assertions and observations.")
- *Moral cognition* (for example, see "The Emotional Dog and Its Rational Tail"[18]). Human minds have interesting ways of reasoning about moral and immoral actions. Perhaps this work can shed light on why some security officers pound fists and insist that the enterprise firewall must block all recreational browsing—even though studies show that such browsing increases productivity.

One could teach a whole course on this—in fact, I've tried to.

Why should human minds behave this way? To paraphrase Tom Lehrer, that's not our department. But that's how they seem to behave, and because human minds are part of the system of usable and effective security, we'd be wise to take into account the strange ways they work. ■

References

1. E. Protalinksi, "Mom Accessed School System 110 Times to Change Kids' Grades," ZDNet, 19 July 2012; www.zdnet.com/mom-accessed-school-system-110-times-to-change-kids-grades-7000001230.
2. R.K. Hastie and R.M. Dawes, *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*, 2nd ed., Sage, 2009.
3. R.F. Pohl, *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*, Psychology Press, 2005.
4. D. Gilbert, *Stumbling on Happiness*, Vintage Books, 2007.
5. D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.
6. T.D. Wilson and J.W. Schooler, "Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions," *J. Personality and Social Psychology*, vol. 60, no. 2, 1991, pp. 181–192.
7. S. Trudeau, S. Sinclair, and S. Smith, "The Effects of Introspection on Creating Privacy Policy," *Proc. 8th ACM Workshop on Privacy in the Electronic Society* (WPES 09), ACM, 2008, pp. 1–10.
8. E.W. Dunn and S.A. Laham, "Affective Forecasting: A User's Guide to Emotional Time Travel," *Affect in Social Thinking and Behavior*, J. Forgas, ed., Psychology Press, 2006.
9. E. Pronin, C. Olivola, and K. Kennedy, "Doing unto Future Selves as You Would Do unto Others: Psychological Distance and Decision Making," *Personality and Social Psychology Bulletin*, vol. 34, no. 2, 2007, pp. 224–237.
10. L. Van Boven, D. Dunning, and G. Loewenstein, "Egocentric Empathy Gaps between Owners and Buyers: Misperceptions of the Endowment Effect" *J. Personality and Social Psychology*, vol. 79, no. 1, 2000, pp. 66–76.
11. Y. Wang, S.W. Smith, and A. Gettinger, "Access Control Hygiene and the Empathy Gap in Medical IT," *HealthSec*, Usenix Assoc., 2012; https://www.usenix.org/conference/healthsec12/access-control-hygiene-and-empathy-gap-medical-it.
12. J. Pita et al., "Effective Solutions for Real-World Stackelberg Games: When Agents Must Deal with Human Uncertainties," *Proc. 8th Int'l Conf. Autonomous Agents and Multiagent Systems*, Int'l Foundation for Autonomous Agents and Multiagent Systems, 2009, http://teamcore.usc.edu/papers/2009/COBRA.pdf.
13. R. Yang et al., "Improving Resource Allocation Strategy against Human Adversaries in Security Games," *Int'l Joint Conf. Artificial Intelligence* (IJCAI 11), AAAI, 2011; http://teamcore.usc.edu/papers/2011/ijcai11_paper148_cameraready.pdf.
14. E. Diener, D. Wirtz and S. Oishi, "End Effects of Rated Life Quality: The James Dean Effect," *Psychological Science*, vol. 12, no. 2, 2001, pp. 124–128.
15. D. Gilbert et al., "The Peculiar Longevity of Things Not So Bad," *Psychological Science*, vol. 15, no. 1, 2004, pp. 14–19.
16. D.T. Gilbert and T.D. Wilson, "Why the Brain Talks to Itself: Sources of Error in Emotional Prediction," *Philosophical Trans. Royal Soc. B*, vol. 364, no. 1521, 2009, pp. 1335–1341.
17. S.J. Handley and J. Evans, "Supposition and Representation in Human Reasoning," *Thinking and Reasoning*, vol. 6, no. 4, 2000, pp. 273–311.
18. J. Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Rev.*, vol. 108, no. 4, 2001, pp. 814–834.

Sean W. Smith is a professor of computer science at Dartmouth College. Contact him at sws@cs.dartmouth.edu.