Leveraging Behavioral Science to Mitigate Cyber Security Risk

Shari Lawrence Pfleeger
Institute for Information Infrastructure Protection
Dartmouth College
4519 Davenport St. NW
Washington, DC 20016
Phone: +1 603 729-6023
Email: shari.l.pfleeger@dartmouth.edu

Deanna D. Caputo (corresponding author)
MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7539
Phone: +1 703 983-3846
Email: dcaputo@mitre.org

Leveraging Behavioral Science to Mitigate Cyber Security Risk

Shari Lawrence Pfleeger (Dartmouth College)

Deanna D. Caputo (MITRE Corporation)

Abstract: Most efforts to improve cyber security focus primarily on incorporating new technological approaches in products and processes. However, a key element of improvement involves acknowledging the importance of human behavior when designing, building and using cyber security technology. In this survey paper, we describe why incorporating an understanding of human behavior into cyber security products and processes can lead to more effective technology. We present two examples: the first demonstrates how leveraging behavioral science leads to clear improvements, and the other illustrates how behavioral science offers the potential for significant increases in the effectiveness of cyber security. Based on feedback collected from practitioners in preliminary interviews, we narrow our focus to two important behavioral aspects: cognitive load and bias. Next, we identify proven and potential behavioral science findings that have cyber security relevance, not only related to cognitive load and bias but also to heuristics and behavioral science models. We conclude by suggesting several next steps for incorporating behavioral science findings in our technological design, development and use.

Keywords: cyber security, cognitive load, bias, heuristics, risk communication, health models

2 Introduction

"Only amateurs attack machines; professionals target people" (Schneier, 2000).

What is the best way to deal with cyber attacks? Cyber security promises protection and prevention, using both innovative technology and an understanding of the human user. Which aspects of human behavior offer the most promise in making cyber security processes and products more effective? What role should education and training play? How can we encourage good security practices without unnecessarily

interrupting or annoying users? How can we create a cyber environment that provides users with all of the functionality they need without compromising enterprise or national security? We investigate the answers to these questions by examining the behavioral science literature to identify behavioral science theories and research findings that have the potential to improve cyber security and reduce risk. In this paper, we report on our initial findings, describe several behavioral science areas that offer particularly useful applications to security, and describe how to use them in a general risk-reduction process.

The remainder of this paper is organized in five sections. Section 2 describes some of the problems that a technology-alone solution cannot address. Section 3 explains how we used a set of scenarios to elicit suggestions about the behaviors of most concern to technology designers and users. Sections 4 and 5 highlight several areas of behavioral science with demonstrated and potential relevance to security technology. Finally, Section 6 suggests possible next steps toward inclusion of behavioral science in security technology's design, construction and use.

3 Why Technology Alone is Not Enough

The media frequently express the private sector's concern about liability for cyber attacks and its eagerness to minimize risk. The public sector has similar concerns, because aspects of everyday life (such as operation and defense of critical infrastructure, protection of national security information, and operation of financial markets) involve both government regulation and private sector administration.[1] The government's concern is warranted: the Consumer's Union found that government was the source of one-fifth of the publicly-reported data breaches between 2005 and mid-2008 (Consumer's Union, 2008). The changing nature of both technology and the threat environment makes the risks to information and infrastructure difficult to anticipate and quantify.

[1] See, for example, the video at http://www.cbsnews.com/video/watch/?id=5578986n&tag=related;photovideo

Problems of appropriate response to cyber incidents are exacerbated when security technology is perceived as an obstacle to the user. The user may be overwhelmed by difficulties in security implementation, or may mistrust, misinterpret or override the security. A recent study of users at Virginia Tech illustrates the problem (Virginia Tech, 2011). Bellanger et al. examined user attitudes and the "resistance behavior" of individuals faced with a mandatory password change. The researchers found that, even when passwords were changed as required, the changes were intentionally delayed and the request perceived as being an unnecessary interruption. "People are conscious that a password breach can have severe consequences, but it does not affect their attitude toward the security policy implementation." Moreover, "the more technical competence respondents have, the less they favor the policy enhancement. … In a voluntary implementation, that competence may be a vector of pride and accomplishment. In a mandatory context, the individual may feel her competence challenged, triggering a negative attitude toward the process."

In the past, solutions to these problems have ranged from strict, technology-based control of computer-based human behavior (often with inconsistent or sometimes rigid enforcement) to comprehensive education and training of system developers and users. Neither extreme has been particularly successful, but recent studies suggest that a blending of the two can lead to effective results. For example, the U.K. Office for Standards in Education, Children's Services and Skills (Ofsted) evaluated the safety of online behavior at 35 representative schools across the U.K. "Where the provision for e-safety was outstanding, the schools had managed rather than locked down systems. In the best practice seen, pupils were helped, from a very early age, to assess the risk of accessing sites and therefore gradually to acquire skills which would help them adopt safe practices even when they were not supervised" (Ofsted, 2010). In other words, the most successful security behaviors were exhibited in schools where students were taught appropriate behaviors and then trusted to behave responsibly. The Ofsted report likens the approach to teaching children how to cross the road safely, rather than relying on adults to accompany the children across the road each time.

This approach is at the core of our research. Our overarching hypothesis is that if humans using computer systems are given the tools and information they need, taught the meaning of responsible use, and then trusted to behave appropriately with respect to cyber security, desired outcomes may be obtained without security being perceived as onerous or burdensome. By both understanding the role of human behavior and leveraging behavioral science findings, the designers, developers and maintainers of information infrastructure can address real and perceived obstacles to productivity and provide more effective security. These behavioral changes take time, so plans for initiating change should include sufficient time to propose the change, implement it, and have it become part of the culture or common practice.

Other evidence (Predd et al., 2008; Pfleeger et al., 2010) is beginning to emerge that points to the importance of understanding human behaviors when developing and providing cyber security.[2] There is particular interest in using trust to mitigate risk, especially online. For example, the European Union funded a several-year, multi-disciplinary project on online trust (iTrust),[3] documenting the many ways that trust can be created and broken. Now, frameworks are being developed for analyzing the degree to which trust is built and maintained in computer applications (Riegelsberger, Sasse and McCarthy, 2005). More broadly, a rich and relevant behavioral science literature addresses critical security problems, such as employee deviance, employee compliance, effective decision-making, and the degree to which emotions (Lerner and Tiedens, 2006) or stressful conditions (Klein and Salas, 2001) can lead to riskier choices by decision-makers.[4] At the same time, there is much evidence that technological advances can have unintended consequences that reduce trust or increase risk (Tenner, 1991). For these reasons, we conclude that it is important to include the human element when designing, building and using critical systems.

---

[2] See the First Interdisciplinary Workshop on Security and Human Behavior, described at http://www.schneier.com/blog/archives/2008/06/security_and_http://www.cl.cam.ac.uk/~rja14/shb08.html

[3] See workshop papers at http://www.informatik.uni-trier.de/~ley/db/conf/itrust/itrust2006.html

[4] The National Science Foundation program is interested in the connections between social science and cyber security. It has announced a new program that encourages computer scientists and social scientists to work together (Secure and Trustworthy Cyberspace, described at http://www.nsf.gov/pubs/2012/nsf12503/nsf12503.htm?WT.mc_id=USNSF_25&WT.mc_ev=click).

To understand how to design and build systems that encourage users to act responsibly when using them, we identified two types of behavioral science findings: those that have already been shown to demonstrate a welcome effect on cyber security implementation and use, and those with potential to have such an effect. In the first case, we documented the relevant findings, so that practitioners and researchers can determine which approaches are most applicable to their environment. In the second case, we are designing a series of studies to test promising behavioral science results in a cyber security setting with the goal of determining which results (with associated strategies for reducing or mitigating the behavioral problems they reflect) are the most effective.

However, applying behavioral science findings to cyber security problems is an enormous undertaking. To maximize the likely effectiveness of outcomes, we used a set of interviews to elicit practitioners' opinions about behaviors of concern so that we could focus on those perceived as most significant. We describe the interviews and results in Section 3. These findings suggest hypotheses about the role of behavior in addressing cyber security issues.

4 Identifying Behavioral Aspects of Security

Designers and developers of security technology can leverage what is known about people and their perceptions to provide more effective security. A former Israeli airport security chief said,

> "I say technology should support people. And it should be skilled people at the center of our security concept rather than the other way around" (Amos, 2010).

To implement this kind of human-centered security, technologists must understand the behavioral sciences as they design, develop and use technology. However, translating behavioral results to a technological environment can be a difficult process. For example, system designers must address the human elements obscured by computer mediation. Consumers making a purchase online trust that the

merchant represented by the website is not simply taking their money, but is also fulfilling its obligation to provide goods in return. The consumer infers the human involvement of the online merchant behind the scenes. Thus, at some level, the buyer and seller are humans enacting a transaction enabled by a system designed, developed and maintained by humans. There may be neither actual human contact nor direct knowledge of the other human actors involved, but the transaction process reflects its human counterpart.

Preventing or mitigating adverse cyber security incidents requires action at many stages: designing the technology being incorporated in the infrastructure; implementing, testing and maintaining the technology; and using the technology to provide essential products and services. Behavioral science has addressed notions of cyber security in these activities for many years. Indeed, Sasse and Flechais (2005) note that secure systems are socio-technical systems in which we should use an understanding of behavioral science to "prevent users from being the 'weakest link.'" For example, some behavioral scientists have investigated how trust mechanisms affect cyber security. Others have reported findings related to the design and use of cyber systems, but the relevance and degree of effect have not yet been tested.

Some of the linkage between behavioral science and security is specific to certain kinds of systems. For example, Castelfranchi and Falcone (1998 and 2002) analyze trust in multi-agent systems from a behavioral perspective. They view trust as having several components, including beliefs that must be held to develop trust (the social context, as described by Riegelsberger, Sasse and McCarthy (2003)) and relationships to previous interactions (the temporal context of the Riegelsberger-Sasse-McCarthy framework).They use psychological factors to model trust in multi-agent systems. In addition to social and temporal concerns, we add expectations of fulfillment, where someone trusting someone or something else expects something in return (Baier, 1986). This behavioral research sheds light on the nature of a user's expectation and on perceived trustworthiness of technology-mediated interactions and has important implications related to the design of protective systems and processes.

Sasse and Flechais (2005) view security from three distinct perspectives: product, process and panorama.

- **Product.** This perspective includes the effect of the security controls, such as the policies and mechanisms on stakeholders (e.g., designers, developers, users). The controls involve requirements affecting physical and mental workload, behavior, and cost (human and financial). Users trust the product to maintain security while getting the primary task done.

- **Process.** This aspect addresses how security decisions are made, especially in early stages of requirements-gathering and design. The process should allow the security mechanisms to be "an integral part of the design and development of the system, rather than being 'added on'" (Sasse and Flechais, 2005). Because "mechanisms that are not employed in practice, or that are used incorrectly, provide little or no protection," designers must consider the implications of each mechanism on workload, behavior and workflow (Sasse and Flechais, 2005). From this perspective, the stakeholders must trust the process to enable them to make appropriate and effective decisions, particularly about their primary tasks

- **Panorama.** This aspect describes the context in which the security operates. Because security is usually not the primary task, users are likely to "look for shortcuts and workarounds, especially when users do not understand why their behavior compromises security… A positive security culture, based on a shared understanding of the importance of security… is the key to achieving desired behavior" (Sasse and Flechais, 2005). From this perspective, the user views security mechanisms as essential even when they seem intrusive, limiting, or counterproductive.

4.1 Scenario Creation

Because the infrastructure types and threats are vast, we used interview results to narrow our investigation to those behavioral science areas with demonstrated or likely potential to enhance an actor's confidence in using any information infrastructure. To guide our interviews, we worked with two dozen U.S. government and industry employees familiar with information infrastructure protection issues to define three threat scenarios relevant to protecting the information infrastructure. The methodology and resulting

analyses were conducted by the paper's first author and involved five steps:

- *Choosing topics.* We chose three security topics to discuss, based on recent events. The combination of the three was intended to represent an (admittedly incomplete but) significant number of typical concerns, the discussion of which would reveal underlying areas ripe for improvement.
- *Creating a representative, realistic scenario for each topic.* Using our knowledge of recent cyber incidents and attacks, we created an attack scenario for each plausible topic, portraying a cyber security problem for which a solution would be welcomed by industry and government.
- *Identifying people with decision making authority about cyber security products and usage to interview about the scenarios.* We identified people from industry and government who were willing to participate in interviews.
- *Conducting interviews.* Our discussions focused on two questions: Are these scenarios realistic, and how could the cyber security in each situation be improved?
- *Analyzing the results and their implications.* We analyzed the results of these interviews and their implications for our research.

Scenario 1: Improving Security Awareness Among Builders of Information Infrastructure

Security is rarely the primary task of those who use the information infrastructure. Typically, users seek information, analyze relationships, produce documents, and perform tasks that help them understand situations and take action. Similarly, system developers often focus on these primary tasks before incorporating security into an architecture or design. Moreover, system developers often implement security requirements by choosing security mechanisms that are easy to build and test or that meet some other technical system objective (e.g., reliability). Developers rarely take into account the usability of the mechanism or the additional cognitive load it places on the user. Scenario 1 describes ways to improve

security awareness among system builders so that security is more likely to be useful and effective.

Suppose software engineers are designing and building a system to support the creation and transmission of sensitive documents among members of an organization. Many aspects of document creation and transmission are well known, but security mechanisms for evaluating sensitivity, labeling documents appropriately and transmitting documents securely have presented difficulties for many years. In our scenario, software engineers are tasked to design a system that solicits information from document creators, modifiers and readers, so that a trust designation can be assigned to each document. Security issues include understanding they types of trust-related information needed, determining the role of a changing threat environment, and defining the frequency at which the trust information should be refreshed and re-evaluated (particularly in light of cyber security incidents that may occur during the life of the document). In addition, the software engineers must implement some type of summary trust designation that will have meaning to document creators, modifiers and readers alike.

This trust designation, different from the classification of document sensitivity, represents the degree to which both the content and provider (or modifier) can be trusted and for how long. For example, a document about a nation's emerging military capability may be highly classified (that is, highly sensitive), regardless of whether the information provider is highly trusted (because, for example, he has repeatedly provided highly useful information in the past) or not (because, for example, he frequently provides incorrect or misleading information).

There are two important aspects of the software engineers' security awareness. First, they must be able to select security mechanisms for implementing the trust designation that allow them to balance security with performance and usability requirements. This balancing entails appreciating and accommodating the role of security in the larger context of the system's intended purpose and multiple uses. Second, the users must be able to trust that the appropriate security mechanism is chosen. Trust means that the mechanism itself must be appropriate to the task. For example, the Biba Integrity Model (Biba, 1977), a system of computer security policies expressed as access control rules, is designed to ensure data integrity. The

model defines a hierarchy of integrity levels, and then prevents participants from corrupting data of an integrity level higher than the subject, or from being corrupted by data from a level lower than the subject. The Biba model was developed to extend the Bell-La Padula (1973) model, which addresses only data confidentiality. Thus, understanding and choice of policies and mechanisms are important aspects in which we trust software engineers to exercise discretion. In addition, software engineers must be able to trust the provenance, correctness and conformance to expectations of the security mechanisms. Here, "provenance" means not only the applicability of the mechanisms and algorithms but also the source of architectural or implementation modules. With the availability of open source modules and product line architectures (see, for example, Clements and Northrup, 2001), it is likely that some parts of some security mechanisms will have been built for a different purpose, often by a different team of engineers. Builders and modifiers of the current system must know to what degree to trust someone else's modules.

Scenario 2: Enhancing Situational Awareness During a "Cyber Event"

Situational awareness is the degree to which a person or system knows about a threat in the environment. When an emergency is unfolding, the people and systems involved in watching it unfold must determine what has already happened, what is currently happening, and what is likely to happen in the future; then, they make recommendations for reaction based on their situational awareness. The people or systems perceiving the situation have varying degrees of trust in the information they gather and in the providers of that information. When a cyber event is unfolding, information can come from primary sources (such as sensors in process control systems or measurements of network activity) and secondary sources (such as human or automated interpreters of trends).

Consider analysts using a computer system that monitors the network of power systems around the United States. The system itself interacts with a network of systems, each of which collects and analyzes data about power generation and distribution stations and their access points. The analysts notice a series of network failures around the country: first, a power station in California fails, then one in Missouri, and so

on during the first few hours of the event.[5] The analysts must determine not only what is really unfolding but also how to respond appropriately. Security and human behavior are involved in many ways. First, the analyst must know whether to trust the information being reported to her monitoring system. For example, is the analyst viewing a failure in the access point or in the monitoring system? Next, the analyst must be able to know when and whether she has enough information to make a decision about which reactions are appropriate. This decision must be made in the context of an evolving situation, where some evidence at first considered trustworthy is eventually determined not to be (and vice versa). Finally, the analyst must analyze the data being reported, form hypotheses about possible causes, and then determine which interpretation of the data to use. For instance, is the sequence of failures the result of incorrect data transmission, a cyber attack, random system failures, or simply the various power companies' having purchased some of their software from the same vendor (whose system is now failing)? Choosing the wrong interpretation can have serious consequences.

Scenario 3: Supporting Decisions About Trustworthiness of Network Transactions

On Christmas Day, 2009, a Nigerian student flying from Amsterdam to Detroit attempted to detonate a bomb to destroy the plane. Fortunately, the bomb did little damage, and passengers prevented the student from completing his intended task. However, in analyzing why the student was not detected by a variety of airport security screens, it was determined that important information was never presented to the appropriate decision-makers (Baker and Hulse, 2009). This situation forms the core of Scenario 3, where a system queries an interconnected set of databases to find information about a person or situation.

In this scenario, an analyst uses an interface to a collection of data repositories, each of which contains information about crime and terrorism. When the analyst receives a warning about a particular person of interest, she must query the repositories to determine what is known about that person. There are many security issues related to this scenario. First, the analyst must determine the degree to which she can trust

[5] Indeed, at this stage it may not be clear that the event is actually a cyber event. A similar event with similar characteristics occurred on August 14, 2003, in the United States. See http://www.cnn.com/2003/US/08/14/power.outage/index.html

that all of the relevant information resides in at least one of the connected repositories. After the Christmas bombing attempt, it was revealed that the U.K. had denied a visa request by the student, but information about the denial was not available to the Transportation Security Administration when decisions were made about whether to subject the student to extra security screening. Spira (2010) points out that the problem is not the number of databases; it is the lack of ability to search the entire "federation" of databases.

Next, even if the relevant items are found, the most important ones must be visible at the appropriate time. Libicki and Pfleeger (2004) have documented the difficulties in "collecting the dots" before an analyst can take the next step to connect them. If a "dot" is not as visible as it should be, it can be overlooked or given insufficient attention during subsequent analysis. Moreover, Spira (2010) highlights the need for viewing the information in its appropriate context.

Third, the analyst must also determine the degree to which each piece of relevant information can be trusted. That is, not only must she know the accuracy and timeliness of each data item, but she also must determine whether the data source itself can be trusted. There are several aspects to this latter degree of trust, such as knowing how frequently the data source provides the information (that is, whether it is old news), knowing whether the data source is trustworthy enough, and whether circumstances may change the source's trustworthiness. For example, Predd et al. (2008) and Pfleeger et al. (2010) point out the varying types of people with legitimate access to systems taking unwelcome action. A trustworthy insider may become a threat because of a pending layoff or personal problem, inattention or confusion, or her attempt to overcome a system weakness. So the trustworthiness of information and sources must be re-evaluated repeatedly and perhaps even forecast based on predictions about a changing environment.

Finally, the analyst must also determine the degree to which the analysis is correct. Any analysis involves assumptions about variables and their importance, as well as the relationships among dependent and independent variables. Many times, it is a faulty assumption that leads to failure, rather than faulty data.

4.2 Analysis of Results

The three scenarios were intriguing to our interviewees, and all agreed that they were realistic, relevant and important. However, having the interviewees scrutinize the scenarios revealed fewer behavioral insights than we had hoped. In each case, the interviewee viewed each scenario from his or her particular perspective, highlighting only a small portion of the scenario to confirm an opinion he or she held. For example, one of the interviewees used Scenario 3 to emphasize the need for information sharing; another interviewee said that privacy is a key concern, especially in situations like Scenario 2 where significant monitoring must be balanced with protecting privacy.

Nevertheless, many of the interviewees had good suggestions for shaping the way forward. For instance, one said that there is much to be learned from command and control algorithms, where military actors have learned to deal with risk perception, uncertainty, incomplete information, and the need to make an important decision under extreme pressures. There is rich literature addressing decision-making under pressure, from Ellsberg (1964) through Klein (Klein, 1998; Klein, 2009). In particular, Klein's models of adaptive decision-making may be applicable (Klein and Calderwood, 1991; Klein and Salas, 2001). While the scenario methodology was not a structured idea generation approach, to the extent possible, we endeavored to be unbiased in our interpretation of interviewee responses. We were not trying to gather support for preconceived ideas and were genuinely trying to explore new ideas where behavioral science could be leveraged to address security issues.

There were several themes that emerged from the interviews:

- **Security is intertwined with the way humans behave when trying to meet a goal or perform a task.** The separation of primary task from secondary, as well as its impact on user behavior, was first clearly expressed in Smith et al. (1997) and elaborated in the security realm by Sasse et al. (2002). Our interviews reconfirmed that, in most instances, security is secondary to a user's primary task (e.g., finding a piece of information, processing a transaction, making a decision).

When security interferes, the person may ignore or even subvert the security, since the person is rewarded for the primary task. In some sense, the person trusts the system to take care of security concerns. That perspective can lead to at least two unwelcome events. First, when confronted with uncertainty about the security of a course of action, the person *trusts* that the system has assured the safety of the action (for example, when a user opens an attachment assuming that the system has checked it for viruses, or, as in Scenario 3, the users assumed the bomber was not a security risk because his name was not revealed by the security system). Second, when, in the past, security features have prevented or slowed task completion, a user subverts the security because he or she may no longer *trust* the system to enable effective task completion in the future. Thus, understanding the behavioral science (rather than the security itself) can offer new ways to design, build and use systems whose security is understood and respected by the user.

- Interviewees noted in all scenarios **how limitations on memory or analysis capability interfered with an analyst's ability to perform**. One interviewee noted the abundance of information being generated by automated systems, and the increasing likelihood that important events would go unnoticed (Burke 2010). In the behavioral sciences, the term *cognitive load* refers to the amount of stress placed on working memory. First addressed by Miller (1956), who claimed that a person's "working memory" could deal with at most five to nine pieces of information at once, the notion was extended by Chase and Simon (1973) to address memory overload during problem-solving. Several empirical results (see, for example, Scandura, 1971) suggest that individuals vary in their ability to process a given amount of information.

- ***Inattentional blindness* is a particular aspect of cognitive load that played a role in each scenario.** First acknowledged by Mack and Rock (1998) and studied extensively by Simons and his colleagues (see, for example, Simons and Chabris, 1999 and Simons and Jensen, 2009), inattentional blindness refers to a person's inability to notice unexpected events when concentrating on a primary task. For example, inattentional blindness may cause an analyst in

Scenario 2 to miss seeing a pattern in the failure of power plants (e.g., that all failing power plants were in areas experiencing severe drought), or to lead an analyst in Scenario 3 to overlook a warning from the bomber's father because attention was restricted to the bomber himself.

- **There is significant bias in the way each interviewee thinks about security.** This bias reflects the interviewee's experience, goals and expertise, evidencing itself in the way that two people view the same situation in very different ways. For example, interviewees with jobs that focus primarily on privacy thought of the scenarios as protecting data from outsiders but did not consider inadvertent corruption. By understanding biases, security designers and developers can anticipate likely perceptions and account for them when designing approaches to encourage good security behavior.

- **There is a significant element of risk in each scenario, and decision-makers have a difficult time both understanding the nature of the risk (expressed as a combination of likelihood and impact) and balancing multiple perceptions of the risk to make the best decision in the time available.** There is a considerable literature on risk perception and risk communication, with important papers included in the compilations by Mayo and Hollander (1991) and Slovic (2000). By applying behavioral science findings to system design, development and use, users can be made more aware of the likely impact of their security-related decisions.

The interviews revealed how practitioners (i.e., users and developers) do and do not involve security-related concerns in their decision-making process. Several points became clear to us as a result of these discussions:

- Practitioners do not have a common understanding of security.

- Practitioners do not have a heightened awareness of how security can affect all of their job functions and roles. For example, people feel comfortable revealing small amounts of information in each situation but do not realize how easily the information can aggregate into a full picture

that becomes a security concern.

- Practitioners have limited experience in dissecting a situation to identify necessary security relationships.
- The combination of narrow focus with a large (and often growing) quantity of information continues to cause failure to "connect the dots." Finding a pattern or connection among only a few dots within a large set of data is akin to the problem of identifying a constellation in a star-filled nighttime sky. Some people can find the Big Dipper easily, when others see only too many stars. Our interviews made clear that practitioners need training and assistance in identifying important aspects of a situation and in knowing how and when to focus.

Based on the outcomes from our scenario discussions, we narrowed our focus to cognitive load and bias as organizing principles for an investigation of relevant behavioral science theory and research findings that offer promise of more secure systems. We also sought information about people's heuristics and models that might be useful in helping us convey cyber security information and implement relevant results. In the next two sections, we examine both those behavioral science findings that have already been demonstrated to have bearing on cyber security and those with the potential to do so.

5 Areas of Behavioral Science with Demonstrated Relevance

We begin this section by examining several key behavioral science findings that have been demonstrated as relevant to cyber security in general and information infrastructure protection in particular. Then, in the next section we look at behavioral science research that has potential to improve cyber security. In addition, we include descriptions of heuristics and health-related models that may assist designers in building good security into products and processes. In each case, we document the possible implications of each.

5.1 Findings with Demonstrable Relevance to Cyber Security

Behavioral science findings improve product, process and panorama in these examples.

Recognition Easier Than Recollection

The behavioral science literature demonstrates that recognition is significantly easier than recall. After Rock and Engelstein (1959) showed people a single meaningless shape, the participants' ability to recall it declined rapidly, but they could recognize it almost perfectly a month later. In other words, asking participants to recall a shape without being shown examples was far less successful than displaying a collection of shapes and asking them to identify which one had been shown to them initially. Over the next two decades, many large scale empirical studies reinforced this finding. For example, Standing (1973) showed participants a set of complex pictures; the number of pictures in each set ranged from 10 to 10,000. The participants could recognize subsets of them with 95 percent accuracy.

Dhamija and Perrig (2000) studied how well people remember images compared with passwords, and found that people can more reliably recognize their chosen image than remember a selected password. This result is being applied to user-to-computer authentication; either the user selects an image as an authentication picture, or selects a one-time password based on a shape or configuration. Similarly, Zviran and Haga (1990) showed that even text-based challenge-response mechanisms and associative passwords are an improvement over unaided password recall.

Commercial products are using these results. Lamandé (2010) reports that the GrIDSure authentication system (http://www.gridsure.com) has been integrated into Microsoft's Unified Access Gateway (UAG} platform. This system allows a user to authenticate herself with a one-time passcode based on a pattern of squares chosen from a grid. When the user wishes access, she is presented with a grid containing randomly-assigned numbers; she then enters as her passcode the numbers that correspond to her chosen pattern. Because the displayed grid numbers change each time the grid is presented, the pattern enables

the entered passcode to be a one-time code. Many researchers (see, for example, Sasse, 2007; Bond, 2008; Biddle, Chiasson and van Oorschot, 2009) have examined aspects of GrIDSure's security and usability.

Other commercial products use images called Passfaces. Introduced over ten years ago (Brostoff and Sasse, 2000) and evaluated repeatedly (Everitt et al., 2009), Passfaces offer an option that addresses the drawbacks of products like GrIDSure. However, the Consumer's Union study (2008) and others document the degree to which the average user manages multiple passwords—sometimes dozens! This security-in-the-large leads to problems that are also shared with image recognition: interference.

Interference

Frequent changes to a memorized item interfere with remembering the new version of the item. That is, the newest version of the item competes with the previous ones. The frequency of change is important; for example, Underwood (1957) discovered that, in studies in which participants were required to memorize only a few prior lists, their level of forgetting was much less than in studies where the participants were required to memorize many prior lists. Wixted (2004) points out that even dissimilar things can interfere with something a subject is trying to memorize: "…recently formed memories that have not yet had a chance to consolidate are vulnerable to the interfering force of mental activity and memory formation (even if the interfering activity is not similar to the previously learned material)."

In empirical studies applying these findings to password memorability, Sasse, Brostoff and Weirich (2002) showed that login failures increased sharply as required password changes became more frequent. In addition, Brostoff and Sasse (2003) showed that allowing more login attempts led to more successful login sessions; they suggest that forgiving systems result in better compliance than very restrictive ones.

Everitt et al. (2009) and Chiasson et al. (2009) have examined the use of multiple graphical passwords. They found that users with multiple graphical passwords made fewer errors when recalling them, did not create passwords that were directly related to account names, and did not use similar passwords across

multiple accounts. Moreover, even after two weeks, recall success rates remained good with graphical passwords and were better than those with text passwords. Thus, there seemed to be less interference with graphical objects than with textual ones.

Recent studies have addressed additional concerns about recall and interference. For example, Jhawar et al. (2011) suggest that good design can overcome these issues, and that graphical recall can form the basis for effective security practices.

Other Studies at the Intersection

In addition to the findings cited above, most of which are drawn from basic cognitive psychology literature, there are many examples of applied studies from other disciplines where behavioral scientists studied cyber-related problems directly. For example,

- **Sociology**. Cheshire and Cook (2004) applied experimental sociological research results to four different categories of computer-mediated interaction. They offer guidance to computer scientists about how to build trust in online networks. For example, they suggest treating computer-mediated interaction as an architectural problem, using the nature of the mediation to shape desired behavior. They distinguish between random and fixed partners in a transaction, and suggest appropriate mechanisms for interaction based on this characterization (see Figure 1).

	Continuity	
	Random Partner	*Fixed Partner*
Frequency *Iterated Interaction*	• Solicitation by email • Email attachments from unknown individuals	(none)
One-shot Interaction	• Peer-to-peer digital goods exchange • Online "pickup" games	• Online communities • Online auctions • Chat groups • Massively multiplayer online games

Figure 1: Example Architectural Recommendations (Cheshire and Cook, 2004)

- **Economics**. Economists study the role of reputation in establishing trust, and this literature is frequently referenced in work at the intersection of economics and cyber security. For example, many of the papers at the Workshops on the Economics of Information Security leveraged economic results from reputation research. Yamagishi and Matsuda (2003) propose the use of experience-based information about reputation to address the problem of lemons: disappointment in expectation. They show that disappointment is substantially reduced when online traders can freely change their identities and cancel their reputations.
- **Psychology and economics**. There is an interaction between actual costs and perceived costs when people interact, particularly online. Research in this area spans both psychology (the perception) and economics (the real costs). Datta and Chatterjee (2008) have applied some of this research to the transference of trust in electronic markets. They show that the transference is complete only if agency costs from intermediation lie within consumer thresholds.

These examples convince us that mining the behavioral science literature more thoroughly will lead to an empirical basis for improvements in the quality and effectiveness of cyber security defense. This section has provided examples of the direct application of behavioral science research to problems in cyber security. In the next section, we consider other areas where leveraging behavioral science may reap significant benefits in protecting the information infrastructure.

6 Areas of Behavioral Science with Potential Relevance

The is a significant amount of behavioral science research on methods or concepts that influence a person's or group's perceptions, attitudes, and behaviors. Many findings may have bearing on the design, construction and use of information infrastructure protection, but the relevance and degree of effect have

not yet been tested empirically.

In this section, we identify a variety of well-studied behavioral science findings from psychology, behavioral medicine, and other disciplines where techniques have been demonstrated to affect behavior related to cognition and bias. We also describe several heuristics and health-related models that have potential for improving cyber security. However, unlike the findings in Section 4, these findings have not been evaluated specifically in terms of changing cyber security-related behavior. In this section, we introduce each behavioral science finding, discuss a sampling of research results, and describe the possible implications for cyber security.

6.1 Cognition

Cognition refers to the way people process and learn information. There are several findings from research on human cognition that may be relevant to cyber security.

Identifiable Victim Effect

The identifiable victim effect refers to the tendency of individuals to offer greater aid when a specific, identifiable person (the victim) is observed under hardship, when compared to a large, vaguely-defined group with the same need. For example, many people are more willing to help a homeless person living near the office than the several hundred homeless living in their city. **(**Example: K. Jenni and G. Loewenstein, "Explaining the 'Identifiable Victim Effect'," *Journal of Risk and Uncertainty,* 14, 1997, pp. 235-257.) **Implications:** Users may choose stronger security when possible negative outcomes are tangible and personal, rather than abstract.

Elaboration Likelihood Model

The Elaboration Likelihood Model describes how attitudes are formed and persist. It is based on the notion that there are two main routes to attitude change: the central route and the peripheral route. Central processes are logical, conscious, and require a great deal of thought. Therefore, central route processes to

decision-making are only used when people are motivated and able to pay attention. The result of central route processing is often a permanent change in attitude, as people adopt and elaborate on the arguments being made by others. By contrast, when people take the peripheral route, they do not pay attention to persuasive arguments; rather, they are swayed by surface characteristics such as the popularity of the speaker. In this case, attitude change is more like to be only temporary. Research has focused on how to get people to use the central route instead of the peripheral route. (Example: R.E. Petty and J.T. Cacioppo, *Attitudes and Persuasion: Classic and Contemporary Approaches*. Dubuque, IA: W. C. Brown, 1981. R.E. Petty and J.T. Cacioppo, *Communication and Persuasion: Central and Peripheral Routes to Attitude Change,* New York: Springer-Verlag, 1986.) **Implications**: One of the best ways to motivate users to take the central route when receiving a cyber security message is to make the message personally relevant. Fear can also be effective in making users pay attention, but only if levels of fear are moderate and a solution to the fear-inducing situation is also offered; strong fear leads to fight-or-flight (physical) reactions. The central route leads to consideration of arguments for and against, and the final choice is carefully considered. This distinction can be particularly important in security awareness training.

Cognitive Dissonance

Cognitive dissonance is the feeling of discomfort that comes from holding two conflicting thoughts in the mind at the same time. A person often feels strong dissonance when she believes something about herself (e.g., "I am a good person") and then does something counter to it (e.g., "I did something bad"). The discomfort often feels like tension between the two opposing thoughts. Cognitive dissonance is a very powerful motivator that can lead people to change in one of three ways: change behavior, justify behavior by changing the conflicting attitude, or justify behavior by adding new attitudes. Dissonance is most powerful when it is about self-image (e.g., feelings of foolishness, immorality, etc.). (Examples: L. Festinger, *A Theory of Cognitive Dissonance*, Stanford, CA: Stanford University Press, 1957; L. Festinger and J.M. Carlsmith, "Cognitive Consequences of Forced Compliance," *Journal of Abnormal and Social*

Psychology, 58, 1959, pp. 203-211.) **Implications:** Cognitive dissonance is central to many forms of persuasion to change beliefs, values, attitudes and behaviors. To get users to change their cyber behavior, we can first change their attitudes about cyber security. For example, a system could emphasize a user's sense of foolishness concerning the cyber risks he is taking, enabling dissonant tension to be injected suddenly or allowed to build up over time. Then, the system can offer the user ways to relieve the tension by changing his behavior.

Social Cognitive Theory

Social Cognitive Theory is a theory about learning based on two key notions: (1) people learn by watching what others do, and (2) human thought processes are central to understanding personality. This theory asserts that some of an individual's knowledge acquisition can be directly related to observing others within the context of social interactions, experiences, and outside media influences. (Examples: A. Bandura, "Organizational Application of Social Cognitive Theory," *Australian Journal of Management*, 13(2), 1988, pp. 275-302; A. Bandura, "Human Agency in Social Cognitive Theory," *American Psychologist*, 44, 1989, pp. 1175-1184.) **Implications:** By taking into account gender, age, and ethnicity, a cyber awareness campaign could reduce cyber risk by using social cognitive theory to enable users to identify with a recognizable peer and have a greater sense of self-efficacy. The users would then be likely to imitate the peer's actions in order to learn appropriate, secure behavior.

Bystander Effect

The bystander effect is a psychological phenomenon in which someone is less likely to intervene in an emergency situation when other people are present and able to help than when he or she is alone. (Example**:** J.M. Darley and B. Latané, "Bystander Intervention in Emergencies: Diffusion of Responsibility," *Journal of Personality and Social Psychology,* 8, 1968, pp. 377-383.) **Implications:** During a cyber event, users may not feel compelled to increase situational awareness or take necessary security measures because they will expect others around them to do so. Thus, systems can be designed

with mechanisms to counter this effect, encouraging users to take action when necessary.

6.2 Bias

Bias describes a person's tendency to view something from a particular perspective. This perspective prevents the person from being objective and impartial. The following findings about bias may be useful in designing, building and using information infrastructure.

Status Quo Bias

Status quo bias describes the tendency of people to not change an established behavior without a compelling incentive to do so. (Example: W. Samuelson and R. Zeckhauser, "Status Quo Bias in Decision Making," *Journal of Risk and Uncertainty,* 1*,* 1988, pp. 7-59.) **Implications:** Users will need compelling incentives to change their established cyber security behavior. For example, information infrastructure can be designed to provide incentives for people to suspect documents sent from unknown sources. Similarly, the infrastructure can provide designers, developers and users with feedback about their reputations (e.g., "Sixty-three percent of your attachments are never opened by the recipient.") or the repercussions of their actions (e.g., "It was your design defect that enabled this breach") to reduce status quo bias.

Framing Effects

Scientists usually expect people to make rational choices based on the information available to them. Expected utility theory is based on the notion that people choose options that provide the most benefit (i.e., the most utility to them) based on the information available to them. However, there is a growing literature providing evidence that when people must choose among alternatives involving risk, where the probabilities of outcomes are known, they behave contrary to the predictions of expected utility theory. This area of study, called prospect theory, is descriptive rather than predictive; prospect theorists report on how people actually make choices when confronted with information about each alternative.

One of the earliest findings in prospect theory (Tversky and Kahneman, 1981) demonstrated that the framing of a message can affect decision making. Framing refers to the context in which someone interprets information, reacts to events, and makes decisions. For example, the efficacy of a drug can be framed in terms of number of lives saved or number of lives lost; studies have shown that equivalent data framed in opposite ways (gain vs. loss) lead to dramatically different decisions about whether and how to use the same drug. The context or framing of a problem can be accomplished by manipulating the decision options or by referring to qualities of the decision-makers, such as their norms, habits and temperament. (Examples: D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decisions Under Risk," *Econometrica*, 47, 1979, pp. 313-327; A. Tversky and D. Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science,* 211*,* 1981, pp. 453-458.) **Implications:** User choices about cyber security may be influenced by framing them as gains rather than losses, or by appealing to particular user characteristics. Possible applications include classifying anomalous data from an intrusion detection system log, presenting the interface to a firewall as admitting (good) traffic versus blocking (bad) traffic, or describing a data mining activity as exposing malicious behavior.

Optimism Bias

Given the minuscule chances of winning the lottery, it is amazing that people buy lottery tickets. Many people believe they will do better than most others engaged in the same activity, so they buy tickets despite evidence to the contrary. This optimism bias shows itself in many ways, such as overestimating the likelihood of positive events and underestimating the likelihood of negative events. (Examples: N. D. Weinstein, "Unrealistic Optimism About Future Life Events," *Journal of Personality and Social Psychology* 39(5), November 1980, pp. 806–820; D. Dunning, C. Heath and J. M. Suls, "Flawed Self-Assessment: Implications for Health, Education, and the Workplace," *Psychological Science in the Public Interest* 5(3), 2004, pp. 69–106.) **Implications:** Because they underestimate the risk, users may think they are immune to cyber attacks, even when others have been shown to be susceptible. For example, optimism bias may enable spear phishing (messages seeming to come from a trusted source, trying to gain

unauthorized access to data at a particular organization). Optimism bias may also induce people to ignore preventive care measures, such as patching, because they think they are unlikely to be affected. To counter optimism bias, systems can be designed to convey risk impact and likelihood in ways that relate to people's real experiences.

Control Bias

Control bias refers to the tendency of people to believe they can control or influence outcomes that they clearly cannot; this phenomenon is sometimes called the illusion of control. (Example: E. J. Langer, "The Illusion of Control," *Journal of Personality and Social Psychology* 32(2), 1975, pp. 311-328.)
Implications: Users may be less likely to use protective measures (such as virus scanning, clearing cache, checking for secure sites before entering credit card information, or paying attention to spear phishing) when they feel they have control over the security risks.

Confirmation Bias

Once someone takes a position on an issue, she is more likely to notice or give credence to evidence that supports that position than to evidence that discredits it. This confirmation bias (i.e., looking for evidence to confirm a position) results in situations where people are not as open to new ideas as they think they are. They often reinforce their existing attitudes by selectively collecting new evidence, interpreting evidence in a biased way, or selectively recalling information from memory. For example, an analyst finding a perceived pattern in a series of failures will tend to cease looking for other explanations and instead seek confirming evidence for his hypothesis. (Example: M. Lewicka, "Confirmation Bias: Cognitive Error or Adaptive Strategy of Action Control?" in M. Kofta, G. Weary and G. Sedek, *Personal Control in Action: Cognitive and Motivational Mechanisms*. New York: Springer. 1998, pp. 233–255.)
Implications: Users may have initial impressions about how protected (or not) the information infrastructure is that they are using. To overcome their confirmation bias, the system must provide users with an arsenal of evidence to encourage them to change their current beliefs or to mitigate their over-

confidence.

Endowment Effect

The endowment effect describes the fact that people usually place a higher value on objects they own than objects they do not own. A related effect is that people react more strongly to loss than to gain; that is, they will take stronger action to keep from losing something than to gain something. (Example: R. Thaler, "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organization*, 1, 1980, pp. 39-60.) **Implications:** Users may pay more (both figuratively and literally) for security when it lets them keep something they already have, rather than gain something new. This effect, coupled with a framing effect, may have particular impact on privacy. When an action is expressed as a loss of privacy (rather than a gain in capability), people may react to it negatively.

6.3 Heuristics

In psychology, a heuristic is a simple rule inherent in human nature or learned in order to reduce cognitive load. Thus, we find them appealing for addressing the cognitive load issues described earlier. The heuristics' rules are used to explain how people make judgments, decide issues, and solve problems; heuristics are particularly helpful in explaining how people deal with complex problems or incomplete information. When heuristics fail, they can lead to systematic errors or cognitive biases.

Affect Heuristic

The affect heuristic enables someone to make a decision based on an affect (i.e., a feeling) rather than on rational deliberation. If someone has a good feeling about a situation, he may perceive that it has low risk; likewise, a bad feeling can lead to a higher risk perception. (Example: M. Finucane, E. Peters and D. G. MacGregor, "The Affect Heuristic," in T. Gilovich, D. Griffin and D. Kahneman, *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, 2002, pp. 397–420.) **Implications:** If users perceive little risk, the system may need a design that creates a more critical affect toward computer

security that will encourage them to take protective measures. The system should also reward the system administrator who looks closely at a system audit log because something just doesn't "feel" right.

Availability Heuristic

The availability heuristic refers to the relationship between ease of recall and probability. In other words, because of the availability heuristic, someone will predict an event's probability or frequency in a population based on the ease with which instances of an event come to mind. The more recent, emotional, or vivid an event is, the more likely it will come to mind. (Example: A. Tversky and D. Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology* 5, 1973, pp.207-232.) **Implications:** Users will be more persuaded to act responsibly if the system is designed to use vivid, personal events as examples, rather than statistics and facts. Moreover, if the system reports recent cyber events, it may be more effective in encouraging users to take measures to prevent future adverse events. Users' choices may also be heavily biased by the first thing that comes to mind. Therefore, frequent security exercises may encourage more desirable security behavior. On the other hand, a system that has gone for some time without a major cyber incident may lull the administrators into a false sense of security because of the low frequency of events. The administrators may then become lax in applying security updates because of the long run of incident-free operation.

6.4 Health-Related Behavioral Models

In cyber security, we frame many issues using health-related metaphors because they are, in many ways, analogous. For example, we speak of viruses and infections when describing attacks. Similarly, we discuss increasing immunity to intrusions, or to increasing resilience after a successful attack. For this reason, we believe that security design strategies can leverage the significant research into health-related behavioral models. We discuss several candidate models here.

Health Belief Model

The Health Belief Model, developed in the 1950s after the failure of a free tuberculosis screening program, helped the U.S. Public Health Service by attempting to explain and predict health behaviors. It focused on attitudes and beliefs. Six constructs describe an individual's core beliefs based on their perceptions of: susceptibility, severity, benefits, barriers, cues to action, and self-efficacy of performing a given health behavior. The perceived benefits must outweigh the barriers or costs. (Example: I. Rosenstock, "Historical Origins of the Health Belief Model," *Health Education Monographs,* 2(4), 1974.)

Implications: The health and security education models are similar. If the Health Belief Model translates to cyber security awareness, a user will take protective security actions if he feels that a negative condition can be avoided (e.g. computer viruses can be avoided), has a positive expectation that by taking a recommended action he will avoid a negative condition (e.g., doing a virus scan will prevent a viral infection), and believes that he can successfully perform the recommended action (e.g., is confident that he knows how to install virus protection files). The model suggests success only if the benefits (e.g., keeping himself, his organization, and the nation safe) outweigh the costs (e.g., download time, loss of work).

Extended Parallel Process Model

The Extended Parallel Process Model (EPPM) is an extension of the Health Belief Model that attempts to improve message efficacy by using threats. Based on Leventhal's danger control/fear control framework, EPPM, which has multiple components, explains why many fear appeals fail, incorporates fear as a key variable, and describes the relationship between fear and efficacy. Leventhal defines the danger control process as an individual seeking to reduce the risk presented by taking direct action and making adaptive changes but the fear control process focuses on maladaptive changes to the perception, susceptibility and severity of the risk. The EPPM provides guidance about how to construct effective fear-appeal messages: As long as efficacy perceptions are stronger than threat perceptions, the user will go into danger control

mode (accepting the message and taking recommended action to prevent danger from happening). (Examples: K. Witte, "Putting the Fear Back into Fear Appeals: The Extended Parallel Process Model," *Communication Monographs*, 59, 1992, pp. 329-349; H. Leventhal, "Findings and Theory in the Study of Fear Communications," in L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 5, New York: Academic Press, 1970, pp. 119-186.) **Implications**: When used appropriately, threats and fear can be useful in encouraging users to comply with security. However, the messages cannot be too strong, and users must believe that they are able to comply successfully with the security advice. This model may explain how to encourage users to apply security and performance patches, use and maintain anti-virus tools, and avoid risky online behavior.

Illness Representations

The health care community has a great deal of experience with representing the nature and severity of illness to patients, so that patients can make informed decisions about treatment choices and health. In particular, there are lessons to be learned from the way fear messages are used in relatively acute situations to encourage people to take health-promoting actions such as wearing seat belts or giving up smoking. Health researchers (Leventhal, Meyer, and Nerenz, 1980) have found that different types of information are needed to influence both attitudes and reactions to a perceived threat to health and well-being, and that the behavior changes last only for short periods of time. In extending their initial model, the researchers sought adaptations and coping efforts for those patients experiencing chronic illness. The resulting illness representations integrate the coping mechanisms with existing schemata (i.e., the normative guidelines that people hold), enabling patients to make sense of their symptoms and guiding any coping actions. The illness representations have five components: identity, timeline, consequences, control/cure, and illness coherence. (Examples: H. Leventhal, D. Meyer and D.R. Nerenz, "The Common Sense Representation of Illness Danger," in S. Rachman, ed., *Contributions to Medical Psychology*, New York: Pergamon Press, 1980, pp. 17–30; H. Leventhal, I. Brissette and E.A. Leventhal, "The Common-sense Model of Self-Regulation of Health and Illness," in L.D. Cameron and H. Leventhal, eds., *The Self-*

Regulation of Health and Illness Behaviour, London: Routledge, 2003, pp. 42–65.) **Implications**: In a well-designed system, users concerned about whether to trust a site, person, or document can obtain new information about their security posture and evaluate their attempts to deal (e.g., moderate, cure or cope) with its effects. Then, the users form new representations based upon their experiences. These representations are likely to be cumulative, with security information being adopted, discarded or adapted as necessary. Thus, the representations are likely to be linked to the selection of coping procedures, action plans and outcomes. These results could be of significance for developing incident response strategies.

Theory of Reasoned Action/Theory of Planned Behavior

The Theory of Reasoned Action and the Theory of Planned Behavior are based on two notions: (1) people are reasonable and make good use of information when deciding among behaviors, and (2) people consider the implications of their behavior. Behavior is directed toward goals or outcomes, and people freely choose those behaviors that will move them toward those goals. They can also choose not to act if they think acting will move them away from their goals. The theories take into account four concepts: behavioral intention, attitude, social norms, and perceived behavioral control. Intention to behave has a direct influence on actual behavior as a function of attitude and subjective norms. Attitude is a function both of the personal consequences expected from behaving and the affective value placed on those consequences. (Example: I. Ajzen, "From Intentions to Actions: A Theory of Planned Behavior," in J. Kuhl and J. Beckmann, eds., *Action Control: From Cognition to Behavior*. Berlin, Heidelberg, New York: Springer-Verlag, 1985.) **Implications**: To encourage users to change their security behavior, the system must create messages that affect users' intentions; in turn, the intentions are changed by influencing users' attitudes through identification of social norms and behavioral control. The users must perceive that they can control the successful completion of their tasks securely and safely.

Stages of Change Model

The Stages of Change Model assesses a person's readiness to initiate a new behavior, providing strategies

or processes of change to guide her through the stages of change to action and maintenance. Change is a process involving progression through six stages: precontemplation, contemplation (thoughts), preparation (thoughts and action), action (actual behavior change), maintenance, and termination. Therefore, interventions to change behaviors must match and affect the appropriate stage. To progress through the early stages, people apply cognitive, affective, and evaluative processes. As people move toward maintenance or termination, they rely more on commitments and conditioning, (Examples: J.O. Prochaska, J.C. Norcross and C.C. DiClemente, *Changing for Good: The Revolutionary Program That Explains the Six Stages of Change and Teaches You How to Free Yourself From Bad Habits*. New York: W. Morrow; 1994; J.O. Prochaska and C.C. DiClemente, "The Transtheoretical Approach," in J.C. Norcross and M.R. Goldfried, eds. *Handbook of Psychotherapy Integration*, 2nd ed., New York: Oxford University Press, 2005. pp. 147-171.) **Implications**: To change security-related behaviors, it is necessary first to assess the users' stage before developing processes to elicit behavior change. For example, getting software developers to implement security in the code development life cycle, and especially throughout the life cycle, is notoriously difficult. Currently, much effort is directed at moving developers directly to stage four (action), without appropriate attention to the importance of the earlier stages.

Precaution-Adoption Process Model

Theories that try to explain behavior by examining the perceived costs and benefits of behavior change work only if the person has enough knowledge or experience to have formed a belief. The Precaution-Adoption Process Model seeks to understand and explain behavior by looking at seven consecutive stages: unaware; unengaged; deciding about acting; decided not to act; decided to act; acting; and maintenance. People should respond better to interventions that are matched to the stage they are in. (Examples: N.D. Weinstein, "The Precaution Adoption Process," *Health Psychology*, 7(4), 1988, pp. 355-386; N.D. Weinstein and P.M. Sandman, "A Model of the Precaution Adoption Process: Evidence From Home Radon Testing," *Health Psychology*, 11(3), 1992, pp. 170-180.) **Implications**: Security actions may be related to the seven stages. It may be necessary to assess a user's stage before developing a

process to elicit the desired behavior change.

7 Applying Behavioral Science Findings: The Way Forward

We have presented some early results that show why this multi-disciplinary approach is likely to yield useful insights. In this final section, we describe next steps for determining the best ways to blend behavioral science with computer science to yield improved cyber security. The recommended steps involve encouraging multi-disciplinary workshops, performing empirical studies across disciplines, and building an accessible repository of multi-disciplinary findings.

7.1 Workshops Bridging Communities

Multi-disciplinary work can be challenging for many reasons. First, as noted by participants in a National Academy of Science workshop (2010), there are inconsistent terminologies and definitions across disciplines. Particularly for words like "trust" or "risk," two different disciplines can use the same word but with very different meanings and assumptions. Second, there are few incentives to publish findings across disciplines, so many researchers work in distinct and separate areas that do not customarily share information. For this reason, we recommend the establishment of workshops that bridge communities so that each community's knowledge can benefit the others'.

In July 2010, the Institute for Information Infrastructure Protection (I3P) held a two-day workshop to bring together members of the behavioral science community and the cyber security community, examine how to move successfully-evaluated findings into practice, and establish groups of researchers willing to empirically evaluate promising findings and assess their applicability to cyber security. The workshop created an opportunity for the formation of groups of researchers and practitioners eager to evaluate and adopt more effective ways of integrating behavioral science with cyber security. That is, the workshop is the first step in what we hope will be a continuing partnership between computer science and behavioral

science that will improve the effectiveness of cyber security.

The output of the workshop included:

- Identification of existing findings that can enhance cyber security in the near term.
- Identification of potential behavioral science findings that could be applied but necessitate empirical evaluations of their effects on cyber security.
- Identification of cyber security areas and problems where application of concepts from behavioral science could have a positive impact.
- Establishment of an initial repository of information about behavioral science and cyber security.

As a result of this workshop, several spear phishing studies were conducted in university and industrial settings, and an incentives study, to empirically demonstrate what kinds of incentives (i.e., money, convenient parking spots, public recognition, etc.) would most motivate users to have good cyber hygiene, was designed for future administration. A second workshop was held in October 2011 to report on the studies' findings and to organize further studies.

Workshops of this kind can not only act as catalysts for the initiation of new research but can also encourage continued interaction and cooperation across disciplines. Similar efforts are being encouraged in several areas of cyber security, particularly in usable security (Pfleeger, 2011).

7.2 Empirical Evaluation Across Disciplines

We hope to expand the body of knowledge on the interactions between human behavior and cyber security via investigations that will produce both innovative experimental designs and data that can form the basis of experimental replication and tailoring of applications to particular situations. However, there are challenges to performing this type of research, especially when resources are constrained. For example, it is not usually possible to build the same system twice (one as control, one as treatment) and

compare the results, so good experimental design is crucial in producing strong, credible results with sufficient levels of external validity.

Empirical evaluation of the effects of change on cyber security involves many things, including identifying variables, controlling for bias and interaction effects, and determining the degree to which results can be generalized. These are fundamental principles of the empirical method but are often not understood or not applied appropriately. We hope to produce more comprehensive guidelines for experimental design, aimed at assisting cyber security practitioners and behavioral scientists in designing evaluations that will produce the most meaningful results. These guidelines will highlight several issues:

- The need to design a study so that confounding variables and bias are reduced as much as possible.
- The need to state the experimental hypothesis and identify dependent and independent variables.
- The need to identify the research participants and determine which population is under scrutiny.
- The need for clear and complete sampling procedures, so that the sample represents the identified population.
- The need to describe experimental conditions in enough detail so that the reader can understand the study and also replicate it.
- The need to do an effective post-experiment debriefing, especially for studies where the actual intent of the study is not revealed until the study is completed.

There are several examples of good experimental design for studies at the intersection of behavioral science and cyber security. For instance, many lessons were learned in an experiment focused on insider threat (Caputo, Maloof and Stephens, 2009). In this study, the researchers encountered several challenges in selecting the best sample and following strict empirical procedures. They documented the importance of pilot testing their experimental design before engaging their targeted participants. In particular, it was

difficult to get corporate participants to perform the experimental tasks with the same motivation that the average users have when doing their regular jobs. Therefore, the researchers used pilot testing to determine what would motivate participants. Then, the motivation was built into the study design. Although this study used corporate employees, real networks, and plausible tasks to make the research environment as realistic as possible, generating data sets in any controlled situation reduced the researchers' ability to generalize the findings to complex situations.

There are many studies that can benefit from better data collection and better study design. Pfleeger et al. (2006) suggest a roadmap for improved data collection and analysis of cyber security information. In addition, Cook and Pfleeger (2010) describe how to build improvements on existing data sets and findings.

7.3 Repository of Findings

We are building a repository of relevant findings, including data sets where available, to serve at least two purposes. First, it will provide the basis for decision-making about when and how to include behavioral considerations in the specification, design, construction and use of cyber security products and processes. Second, it will enable researchers and practitioners to replicate studies in their own settings, to confirm or refute earlier findings and to tailor methods to particular needs and constraints. Such information will lay the groundwork for evidence-based cyber security.

This paper reports on the findings of our initial foray into the blending of behavioral science and cyber security. In recent years, there has been much talk about inviting both disciplines to collaborate, but little work has been done to open discussion broadly to both communities. Our workshops took bold and broad steps, and it is hoped that the activities reported here, built on the shoulders of work performed in both communities over the past two decades, will encourage others to join us in thinking more expansively about cyber security problems and possible solutions. In particular, we encourage others engaged in research across disciplines to contact us, so that we can establish virtual and actual links that move us

toward understanding and implementation of improved cyber security.

8 References

Amos, Deborah, "Challenge: Airport Screening Without Discrimination," Morning Edition, National Public Radio, January 14, 2010, available at http://www.npr.org/templates/story/story.php?storyId=122556071

Baier, Annette, "Trust and Antitrust," *Ethics,* Vol. 96, No. 2, 1986, pp. 231-260.

Baker, Peter and Carl Hulse, "U.S. Had Early Signals of Terror Plot, Obama Says," *New York Times*, 30 December 2009, page 1.

Bell, David E. and Leonard J. La Padula, "Secure Computing Systems: Mathematical Foundations," MITRE Technical Report MTR-2547, The MITRE Corporation, Bedford, MA, 1973.

Biba, Kenneth J., "Integrity Considerations for Secure Computer Systems," MITRE Technical Report MTR-3153, The MITRE Corporation, Bedford, MA, April 1977.

Biddle, Robert, Sonia Chiasson, P.C. van Oorschot, "Graphical Passwords: Learning from the First Generation," Technical Report 09-09, School of Computer Science, Carleton University, Ottawa, Canada, 2009.

Bond, Michael, "Comments on GrIDSure Authentication," 28 March 2008, available at http://www.cl.cam.ac.uk/~mkb23/research/GridsureComments.pdf

Brostoff, Sacha and M. Angela Sasse, "Are Passfaces more usable than passwords? A field trial investigation," in S. McDonald et al. (Eds) "People and Computers XIV - Usability or Else," *Proceedings of HCI 2000,* Sunderland, UK, Springer, 2000, pp. 405-424.

Brostoff, Sacha and M. Angela Sasse, "Ten Strikes and You're Out: Increasing the Number of Login

Attempts Can Improve Password Usability, *Proceedings of CHI 2003 Workshop on Human-Computer Interaction and Security Systems*, Ft. Lauderdale, FL, 2003.

Burke, Cody, "Intelligence Gathering Meets Information Overload," *Basex TechWatch*, 14 January 2010, available at http://www.basexblog.com/2010/01/14/intelligence-gathering-meets-io/

Caputo, Deanna, Marcus Maloof and Gregory Stephens, "Detecting Insider Theft of Trade Secrets," *IEEE Security and Privacy* 7(6), November/December 2009, pp. 14-21.

Castelfranchi, Cristiano and Rino Falcone, "Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification, *Proceedings of the Third International Conference on Multi Agent Systems,* 1998.

Castelfranchi, Cristiano and Rino Falcone, "Social Trust: A Cognitive Approach," in Cristano Castelfranchi and Yao-Hua Tan, eds., *Trust and Deception in Virtual Societies*, Kluwer Academic Publishers, Amsterdam, 2002.

Chase, W.G. and H. A. Simon, "Perception in Chess," *Cognitive Psychology* 4(1), 1973, pp. 55–81.

Cheshire, Coye and Karen Cook, "The Emergence of Trust Networks Under Uncertainty: Implications for Internet Interactions," *Analyse & Kritik* 26, 2004, pp. 220-240.

Chiasson, Sonia, Alain Forget, Elizabeth Stobert, Paul C. van Oorschot and Robert Biddle, "Multiple password interference in text passwords and click-based graphical passwords," *ACM Computer and Communications Security* (CCS), November 2009, pp. 500-511.

Clements, Paul and Linda Northrup, *Software Product Lines: Practices and Patterns*, Addison-Wesley, Reading, MA, 2001.

Consumer's Union, "ID Leaks: A Surprising Source is Your Government at Work," *Consumer Reports*, September 2008, available at http://www.consumerreports.org/cro/money/credit-loan/identity-theft/government-id-leaks/overview/government-id-leaks-ov.htm

Cook, Ian P. and Shari Lawrence Pfleeger, "Security Decision Support Challenges in Data Collection and Use," *IEEE Security and Privacy*, 8(3), May-June 2010, pp. 28-35.

Datta, Pratim and Sutirtha Chatterjee, "The Economics and Psychology of Consumer Trust in Intermediaries in Electronic Markets: the EM-Trust Framework, *European Journal of Information Systems*, 17(1), February 2008, pp. 12-28.

Dhamija, Rachna and Adrian Perrig, "Déjà Vu: A User Study Using Images for Authentication," *Proceedings of the 9th USENIX Security Symposium*, Denver, CO, August 2000.

Ellsberg, Daniel J., Risk, *Ambiguity and Decision*, RAND Report D-12995, RAND Corporation, Santa Monica, CA, 1964.

Everitt, Katherine, Tanya Bragin, James Fogarty and Tadayoshi Kohno, "A comprehensive study of frequency, interference, and training of multiple graphical passwords," *ACM Conference on Human Factors in Computing Systems* (CHI), April 2009.Jhawar, Ravi, Philip Inglesant, Martina Angela Sasse and Nicolas Courtois, "Make Mine a Quadruple: Strengthening the Security of Graphical One-Time PIN Authentication," *Proceedings of the Fifth International Conference on Network and Systems Security*, September 6-8, 2011, Milan, Italy.

Klein, G. A. and R. Calderwood, "Decision Models: Some Lessons From the Field," *IEEE Transactions on Systems, Man and Cybernetics* 21(5), September/October 1991, pp. 1018-1026.

Klein, Gary A., *Sources of Power: How People Make Decisions*, MIT Press, Cambridge, MA, 1998.

Klein, Gary A. and Eduardo Salas, eds., *Linking Expertise and Naturalistic Decision Making,* Erlbaum, 2001.

Klein, Gary A., *Streetlights and Shadows: Searching for the Keys to Adaptive Decision Making*, MIT Press, Cambridge, MA, 2009.

Lamandé, Emmanuelle, "GrIDSure Authenticates Microsoft's Latest Remote Application Platform,"

Global Security Mag, 27 April 2010, available at http://www.globalsecuritymag.com/GrIDsure-authenticates-Microsoft-s,20100427,17307.html

Lerner, J.S. and L.Z. Tiedens, "Portrait of the Angry Decision Maker: How Appraisal Tendencies Shape Anger's Influence on Cognition," *Journal of Behavioral Decision Making* (Special Issue on Emotion and Decision Making), 19, 2006, pp. 115-137.

Libicki, Martin C. and Shari Lawrence Pfleeger, "Collecting the Dots: Problem Formulation and Solution Elements," RAND Occasional Paper OP-103-RC, RAND Corporation, Santa Monica, CA, 2004.

Mack, A. and I. Rock, *Inattentional Blindness.* MIT Press, Cambridge, MA, 1998.

Mayo, Deborah and Rachelle Hollander, eds., *Acceptable Evidence: Science and Values in Risk Management*, Oxford University Press, 1991.

Miller, George A., "The Magic Number Seven Plus or Minus Two: Some Limits on Our Capacity to Process Information," *Psychological Review* 63, 1956, pp. 81–97.

National Academy of Science, *Toward Better Usability, Security and Privacy of Information Technology: Report of a Workshop*, National Academies Press, Washington, DC, 2010.

Ofsted (U.K. Office for Standards in Education, Children's Services and Skills), "The Safe Use of New Technologies," OFSTED Report 090231, Manchester, UK, February 2010.

Pfleeger, Shari Lawrence, "Draft Report on the NIST Workshop," March 2011, available at http://www.thei3p.org/docs/publications/436.pdf

Pfleeger, Shari Lawrence, Joel Predd, Jeffrey Hunker and Carla Bulford, "Insiders Behaving Badly: Addressing Bad Actors and Their Actions," *IEEE Transactions on Information Forensics and Security*, 5(2), March 2010.

Pfleeger, Shari Lawrence, Rachel Rue, Jay Horwitz and Aruna Balakrishnan, Investing in Cyber Security: The Path to Good Practice, *Cutter IT Journal*, 19(1), January 2006, pp. 11-18.

Predd, Joel, Shari Lawrence Pfleeger, Jeffrey Hunker and Carla Bulford, "Insiders Behaving Badly," *IEEE Security and Privacy* 6(4), July/August 2008, pp. 66-70.

Riegelsberger, Jens, M. Angela Sasse, and John D. McCarthy, "The Researcher's Dilemma: Evaluating Trust in Computer-Mediated Communication," *International Journal of Human-Computer Studies,* Vol. 58, No. 6, 2003, pp. 759-781.

Riegelsberger, Jens, M. Angela Sasse, and John D. McCarthy, "The Mechanics of Trust: A Framework for Research and Design," *International Journal of Human-Computer Studies,* Vol. 62, No. 3, 2005, pp. 381-422.

Rock, I. And P. Engelstein, "A Study of Memory for Visual Form." *American Journal of Psychology*, 72, 1959, pp. 221-229.

Sasse, M. Angela, "GrIDsure Usability Trials," 2007, available at http://www.gridsure.com/uploads/UCL%20Report%20Summary%20.pdf

Sasse, M. Angela, Sacha Brostoff and Dirk Weirich, "Transforming the 'weakest link: A Human-computer Interaction Approach to Usable and Effective Security," in R. Temple and J. Regnault. eds., *Internet and Wireless Security*, IEE Press, London, 2002, pp. 243-258.

Sasse, M. Angela and Ivan Flechais, "Usable Security: Why Do We Need It? How Do We Get It?," in Lorrie Faith Cranor and Simson Garfinkel, eds., *Security and Usability*, O'Reilly Publishing, Sebastopol, CA, 2005, pp. 13-30.

Scandura, J.M. "Deterministic Theorizing in Structural Learning: Three Levels of Empiricism," *Journal of Structural Learning* 3, 1971, pp. 21–53.

Schneier, Bruce, "Semantic Attacks: The Third Wave of Network Attacks," in *Crypto-Gram Newsletter*, October 15, 2000, available at http://www.schneier.com/crypto-gram-0010.html

Simons, Daniel J. and C. F. Chabris, "Gorillas in Our Midst: Sustained Inattentional Blindness for

Dynamic Events." *Perception*, 28, 1999, pp. 1059-1074.

Simons, Daniel J. and Melinda S. Jensen, "The Effects of Individual Differences and Task Difficulty on Inattentional Blindness," *Psychonomic Bulletin & Review* 16(2), 2009, pp. 398-403.

Slovic, Paul, ed. *The Perception of Risk*, Earthscan Ltd., London, 2000.

Smith, Walter, Becky Hill, John Long and Andy Whitefield, Andy, "A Design-Oriented Framework for Modeling the Planning and Control of Multiple Task Work in Secretarial Office Administration," *Behaviour and Information Technology*, 16(3), 1997, pp. 161-183.

Spira, Jonathan B., "The Christmas Day Terrorism Plot: How Information Overload Prevailed and Counterterrorism Knowledge Sharing Failed," *Basex TechWatch*, 4 January 2010, available at http://www.basexblog.com/category/analysts/jonathan-b-spira/

Standing, L. "Learning 10,000 Pictures," *Quarterly Journal of Experimental Psychology*, 27, 1973, pp. 207-222.

Tenner, Edward, *Why Things Bite Back: Technology and the Revenge of Unintended Consequences*, Vintage Press, 1991.

Underwood, B.J., "Interference and Forgetting," *Psychological Review* 64, 1957, pp. 49-60.

Virginia Tech, "When Users Resist: How to change management and user resistance to password security," *Pamplin*, Fall 2011, available at http://www.magazine.pamplin.vt.edu/fall11/passwordsecurity.html

Wixted, John T., "The Psychology and Neuroscience of Forgetting," *Annual Review of Psychology*, 55, 2004, pp. 235-269.

Yamagishi, T. and M. Matsuda, "The Role of Reputation in Open and Closed Societies: An Experimental Study of Online Trading," Center for the Study of Cultural and Ecological Foundations of Mind, Working Paper Series 8, 2003.

Zviran, Moshe and William J. Haga, "Cognitive Passwords: The Key to Easy Access Control," *Computers and Security* 8(9), 1990, pp. 723-736.

This work was sponsored by grants from the Institute for Information Infrastructure Protection at Dartmouth College, under award number 2006-CS-001-000001 from the US Department of Homeland Security, National Cyber Security Directorate.