Cyber Situation Awareness: Modeling Detection of Cyber Attacks With Instance-Based Learning Theory

Varun Dutt, Young-Suk Ahn and Cleotilde Gonzalez

Human Factors: The Journal of the Human Factors and Ergonomics Society 2013 55: 605 originally published online 6 November 2012

DOI: 10.1177/0018720812464045

Cyber Situation Awareness: Modeling Detection of Cyber Attacks With Instance-Based Learning Theory

Varun Dutt, Indian Institute of Technology, Mandi, India, and **Young-Suk Ahn** and **Cleotilde Gonzalez**, Carnegie Mellon University, Pittsburgh, Pennsylvania

Objective: To determine the effects of an adversary's behavior on the defender's accurate and timely detection of network threats.

Background: Cyber attacks cause major work disruption. It is important to understand how a defender's behavior (experience and tolerance to threats), as well as adversarial behavior (attack strategy), might impact the detection of threats. In this article, we use cognitive modeling to make predictions regarding these factors.

Method: Different model types representing a defender, based on Instance-Based Learning Theory (IBLT), faced different adversarial behaviors. A defender's model was defined by experience of threats: threat-prone (90% threats and 10% nonthreats) and nonthreat-prone (10% threats and 90% nonthreats); and different tolerance levels to threats: risk-averse (model declares a cyber attack after perceiving one threat out of eight total) and risk-seeking (model declares a cyber attack after perceiving seven threats out of eight total). Adversarial behavior is simulated by considering different attack strategies: patient (threats occur late) and impatient (threats occur early).

Results: For an impatient strategy, risk-averse models with threat-prone experiences show improved detection compared with risk-seeking models with nonthreat-prone experiences; however, the same is not true for a patient strategy.

Conclusions: Based upon model predictions, a defender's prior threat experiences and his or her tolerance to threats are likely to predict detection accuracy; but considering the nature of adversarial behavior is also important.

Application: Decision-support tools that consider the role of a defender's experience and tolerance to threats along with the nature of adversarial behavior are likely to improve a defender's overall threat detection.

Keywords: cyber situation awareness, Instance-Based Learning Theory, defender, adversarial behavior, experiences, tolerance

Address correspondence to Varun Dutt, School of Computing and Electrical Engineering, School of Humanities and Social Sciences, Indian Institute of Technology, Mandi, PWD Rest House 2nd Floor, Mandi - 175 001, H.P., India; e-mail: varun@iitmandi.ac.in.

HUMAN FACTORS
Vol. 55, No. 3, June 2013, pp. 605-618
DOI:10.1177/0018720812464045

Cyber attacks are the disruption in the normal functioning of computers and the loss of private information in a network due to malicious network events (threats), and they are becoming widespread. In the United Kingdom, organizers of the London 2012 Olympic Games believe that there is an increased danger of cyber attacks that could seriously undermine the technical network supporting everything, from recording world records to relaying results to commentators at the Games (Gibson, 2011). With the prevalence of "Anonymous" and "LulzSec" hacking groups and other threats to corporate and national security, guarding against cyber attacks is becoming a significant part of IT governance, especially because most government agencies and private companies have moved to online systems (Sideman, 2011). Recently, President Barack Obama declared that the "cyber threat is one of the most serious economic and national security challenges we face as a nation" (White House, 2011). According to his office, the nation's cybersecurity strategy is twofold: (1) to improve our resilience to cyber incidents and (2) to reduce the cyber threat. To meet these goals, the role of the security analyst (called "defender" onwards), a human decision maker who is in charge of protecting the online infrastructure of a corporate network from random or organized cyber attacks, is indispensable (Jajodia, Liu, Swarup, & Wang, 2010). The defender protects a corporate network by identifying, as early and accurately as possible, threats and nonthreats during cyber attacks.

In this article, we derive predictions about the influence of a simulated defender's experience and his or her tolerance to threats on threat detection accuracy for different simulated adversarial behaviors using a computational model. Adversarial behaviors are exhibited through different simulated attack strategies that differ in the timing of threat occurrence in a sequence of network events. We simulate a defender's awareness process through a computational model of dynamic decision making

based on the Instance-Based Learning Theory (IBLT) (Gonzalez, Lerch, & Lebiere, 2003) and derive predictions on the accuracy and timing of threat detection in a computer network (i.e., cyber situation awareness or cyberSA).

Generally, situation awareness (SA) is the perception of environmental elements with respect to time and/or space, the comprehension of their meaning, and the projection of their status after some variable has changed, such as time (Endsley, 1995). CyberSA is the virtual version of SA and includes situation recognition: the perception of the type of cyber attack, source (who, what) of the attack, and target of the attack; situation comprehension: understanding why and how the current situation is caused and what is its impact; and situation projection: determining the expectations of a future attack, its location, and its impact (Jajodia et al., 2010; Tadda, Salerno, Boulware, Hinman, & Gorton, 2006).

During a cyber attack, there could be both malicious network events (threats) and benign network events (nonthreats) occurring in a sequence. Threats are generated by attackers, while nonthreats are generated by friendly users of the network. In order to accurately and timely detect cyber attacks, a defender relies on highly sophisticated technologies that aid in the detection of threats (Jajodia et al., 2010). One of these cyber technologies is called an intrusion detection system (IDS), a program that alerts defenders of possible network threats. The IDS is not a perfect technology, however, and its predictions have both false positives and false negatives (PSU, 2011). Although there is ample current research on developing these technologies, and on evaluating and improving their efficiency, the role of the defender behavior, such as the defender's experience and tolerance to threats, is understudied in the cyber-security literature (Gardner, 1987; Johnson-Laird, 2006; PSU, 2011). In addition, it is likely that the nature of adversarial behavior also influences the defender's cyberSA (Gonzalez, 2012). One characteristic of adversarial behavior is the attacker's strategy regarding the timing of threats during a cyber attack: An impatient attacker might inject all threats in the beginning of a sequence of network events; however, a patient attacker is likely to delay this injection to the very end of a sequence (Jajodia et al., 2010). For both these strategies, there is prevailing uncertainty in terms of exactly when threats might appear in a cyber attack. Thus, it is important for the defender to develop a timely and accurate threat perception to be able to detect a cyber attack. Thus, both the nature of the defender's and adversary's behaviors may greatly influence the defender's cyberSA.

Due to the high demand for defenders, their lack of availability for laboratory experiments, and the difficulty of studying real-world cybersecurity events, an important alternative used to study cyberSA is computational cognitive modeling. A cognitive model is a representation of the cognitive processes and mechanisms involved in performing a task. An important advantage of computational cognitive modeling is to generate predictions about human behavior in different tasks without first spending time and resources in running large laboratory experiments involving participants. Of course, the model used would need to be validated against human data to generate accurate predictions with high confidence. The cognitive model of cyberSA that we present here relies on the Instance-Based Learning Theory, a theory of decisions from experience in dynamic environments that has demonstrated to be comprehensive and robust across multiple decision making tasks (Gonzalez & Dutt, 2011; Gonzalez, Dutt, & Lejarraga, 2011; Gonzalez et al., 2003; Lejarraga, Dutt, & Gonzalez, 2010).

In the next section, we explain the role that a defender's experience and tolerance to threats and the role that the nature of adversarial behavior might play in determining a defender's cyberSA. Then, we present IBLT and the particular implementation of a cognitive model of cyberSA. This presentation is followed by a simulation experiment where the model's experience, tolerance, and the exposure to different adversarial behaviors were manipulated. Using the model, we predict two measures of cyberSA: timing and accuracy. Finally, we conclude with the discussion of the results and their implications to the design of training and decision-support tools for improving a defender's detection of cyber attacks.

ROLE OF EXPERIENCE, TOLERANCE, AND ADVERSARIAL BEHAVIOR

A defender's cyberSA is likely to be influenced by at least three factors: the mix of threat and nonthreat experiences stored in memory; the defender's tolerance to threats, namely, how many network events a defender perceives as threats before deciding that these events represent a cyber attack; and the adversarial behavior (i.e., an attacker's strategy). The adversarial behavior is different from the first two factors. First, actions from the attacker are external or outside of the defender's control. Second, previously encountered adversarial behaviors might influence the defender's current experiences and tolerance to threats.

Prior research indicates that the defender's cyberSA is likely a function of experience with cyber attacks (Dutt, Ahn, & Gonzalez, 2011; Jajodia et al., 2010) and tolerance to threats (Dutt & Gonzalez, in press; McCumber, 2004; Salter, Saydjari, Schneier, & Wallner, 1998). For example, Dutt, Ahn, et al. (2011) and Dutt and Gonzalez (in press) have provided initial predictions about a simulated defender's cyber SA according to its experience and tolerance. Dutt and Gonzalez (in press) created a cognitive model of a defender's cyberSA based upon IBLT and populated the model's memory with threat and nonthreat experiences. The model's tolerance was determined by the number of events perceived as threats before it declared the sequence of network events to be a cyber attack. Accordingly, a model with a greater proportion of threat experiences is likely to be more accurate and timely in detecting threats compared with one with a smaller proportion of such experiences. That is because, according to IBLT, possessing recent and frequent past experiences of network threats also increases the model's opportunity to remember and recall these threats with ease in novel situations. However, it is still not clear how results in Dutt and Gonzalez (in press) were impacted by different adversarial behaviors.

Furthermore, recent research in judgment and decision making has discussed how experiencing outcomes, gained by sampling alternatives in a decision environment, determines our real decision choices after sampling (Gonzalez & Dutt, 2011; Hertwig, Barron, Weber, & Erev, 2004; Lejarraga et al., 2010). For example, having a greater proportion of negative experiences or outcomes in memory for an activity (e.g., about threats) makes a decision maker (e.g., defender) cautious about said activity (e.g., cautious about threats) (Hertwig et al., 2004; Lejarraga et al., 2010).

Prior research has also predicted that a defender's tolerance to threats is likely to influence his or her cyberSA. For example, Salter, Saydjari, Schneier, and Wallner (1998) highlighted the importance of understanding both the attacker and the defender's tolerance, and according to Dutt and Gonzalez (in press), a defender is likely to be more accurate when his or her tolerance is low rather than high. That is because possessing a low tolerance is likely to cause the defender to declare cyber attacks very early on, which may make a defender more timely and accurate in situations actually involving early threats. Although possessing low tolerance might be perceived as costly to an organization's productivity, it is expected to boost the organization's productivity.

The studies discussed previously provided interesting predictions about a simulated defender's experience and tolerance. However, these studies did not consider the role of adversarial behavior (i.e., attacker's strategies) and the interactions between a defender's behavior and an adversary's behavior. Depending upon adversarial behavior, threats within a network might occur at different times and their timing is likely to be uncertain (Jajodia et al., 2010). For example, an impatient attacker could execute all threats very early on in a cyber attack, whereas a patient attacker might decide to delay the attack and thus threats would appear late in the sequence of network events. There is evidence that the recency of information in an observed sequence influences people's decisions when they encounter this information at different times (early or late) in the sequence (Dutt, Yu, & Gonzalez, 2011; Hogarth & Einhorn, 1992). The influence of recency is likely to be driven by people's limited working memory capacity (Cowan, 2001), especially when making decisions from experience in emergency situations (Dutt, Cassenti, &

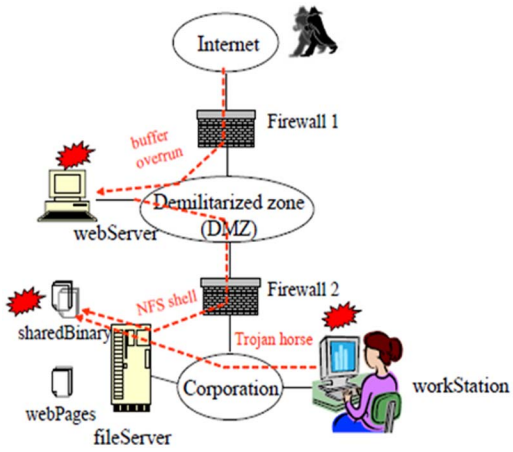

Figure 1. A typical cyber-infrastructure in a corporate network. The attacker uses a computer on the Internet and tries to gain access to the company's workstations through the company's webserver and fileserver.
Source. Adapted from Xie, Li, Ou, Liu, and Levy (2010).

Gonzalez, 2011; Dutt, Ahn, et al., 2011; Gonzalez, 2012). Given the influence of recency, we expect a defender with a greater proportion of threat experiences and a low tolerance to be more accurate and timely against an impatient attack strategy compared with a defender with fewer threat experiences and a high tolerance. However, we do not expect that to be the case for a patient attack strategy. That is because according to IBLT, a model (representing a defender) will make detection decisions by recalling similar experiences from memory. When the model has a greater proportion of threat experiences in memory, it is more likely to recall these experiences early on, making it accurate if threats occur early in an attack (i.e., generated by an impatient strategy). The activated threat experiences would be recalled faster from memory and would also increase the likelihood that the accumulation of evidence for threats exceeds the model's low tolerance. By the same logic, when threats occur late in cyber attacks (i.e., generated by a patient strategy), the situation becomes detrimental to the accuracy and timeliness of a model that possesses many threat experiences in memory and has a low tolerance. In summary, a model's experience of threats, its tolerance to threats, and an attack strategy may limit or enhance the model's cyberSA. These model predictions generate insights for the expected behavior of a defender in such situations.

CYBER INFRASTRUCTURE AND CYBER ATTACKS

The cyber infrastructure in a corporate network may consist of different types of servers and multiple layers of firewalls. We used a simplified network configuration consisting of a webserver, a fileserver, and two firewalls (Ou, Boyer, & McQueen, 2006; Xie, Li, Ou, Liu, & Levy, 2010). An external firewall ("firewall 1" in Figure 1) controls the traffic between the Internet and the Demilitarized Zone (DMZ; a subnetwork that separates the Internet from the company's internal LAN network). Another firewall ("firewall 2" in Figure 1) controls the flow of traffic between the webserver and the fileserver (i.e., the company's internal LAN network). The webserver resides behind the first firewall in the DMZ (see Figure 1). It handles outside customer interactions on a company's Web site. The fileserver resides behind the second firewall and serves as repository accessed by workstations used by corporate employees (internal users) to do their daily operations. These operations are made possible by enabling workstations to run executable files from the fileserver.

Generally, an attacker is identified as a computer on the Internet that is trying to gain access to the internal corporate servers. For this cyber-infrastructure, attackers follow a pattern of "island-hopping" attack (Jajodia et al., 2010, p. 30), where the webserver is compromised first, and then it is used to originate attacks on the fileserver and other company workstations.

A model of the defender, based upon IBLT, is exposed to different island-hopping attack sequences (depending upon the two adversarial timing strategies). Each attack sequence is composed of 25 network events (a combination of both threats and nonthreats), whose nature (threat or nonthreat) is not known to the model. However, the model is able to observe alerts that correspond to some network events (that

are regarded as threats) generated from the intrusion-detection system (Jajodia et al., 2010). Out of 25 events, there are 8 predefined threats that are initiated by an attacker (the rest of the events are initiated by benign users). The model does not know which events are generated by the attacker and which are generated by corporate employees. By perceiving network events in a sequence as threats or nonthreats, the model needs to identify, as early and accurately as possible, whether the sequence constitutes a cyber attack. In this cyber-infrastructure, we represented adversarial behavior by presenting event sequences with different timings for the 8 threats: an impatient strategy, where the 8 threats occur at the beginning of the sequence, and a patient strategy, where the 8 threats occur at the end of the sequence.

INSTANCE-BASED LEARNING MODEL OF DEFENDER'S CYBER SA

IBLT is a theory of how people make decisions from experience in dynamic environments (Gonzalez et al., 2003). Computational models based on IBLT have been shown to generate accurate predictions of human behavior in many dynamic decision-making situations similar to those faced by defenders (Dutt, Ahn, et al., 2011; Dutt, Cassenti, et al., 2011; Dutt & Gonzalez, in press; Gonzalez & Dutt, 2011; Gonzalez et al., 2011). IBLT proposes that every decision situation is represented as an experience called an *instance* that is stored in memory. Each instance in memory is composed of two parts: situation (S) (the knowledge of attributes that describe an event), a decision (D) (the action taken in such situation), and utility (U) (a measure of expected result of a decision that is to be made for an event). For a situation involving securing a network from threats, the situation attributes are those that can discriminate between threat and nonthreat events: the *IP* address of a computer (webserver, fileserver, or workstation, etc.) where the event occurred, the *directory* location in which the event occurred, whether the IDS raised an *alert* corresponding to the event, and whether the *operation* carried out as part of the event (e.g., a file execution) by a user of the network (which could be an attacker) succeeded or failed. In the IBL model of a defender, an instance's S part refers to the situation attributes defined previously, and the U slot refers to the expectation in memory that a network event is a threat or not. For example, an instance could be defined as [webserver, c:\, malicious code, success; threat], where "webserver," "c:\," "malicious code," and "success" constitute the instance's S part and "threat" is the instance's U part (the decision being binary: threat or not, is not included in this model).

IBLT proposes that a decision maker's mental process is composed of five mental phases: recognition, judgment, choice, execution, and feedback. These five decision phases represent a complete learning cycle where the theory explains how knowledge in instances is acquired, reused, and reinforced by human decision makers. Because the focus of this article is on cyberSA rather than on decision making, we only focus on the recognition, judgment, and choice phases in IBLT (not the execution and feedback phases). Among these, the IBLT's recognition and judgment phases accomplish the recognition and comprehension stages in cyberSA, and IBLT's choice phase is used to make a decision in different situations after recognition and comprehension has occurred. To calculate the decision, the theory relies on memory mechanisms such as frequency and recency. The formulation of these mechanisms has been taken from a popular cognitive architecture, ACT-R (Anderson & Lebiere, 1998, 2003) (the model reported here uses a simplified version of the activation equation in ACT-R).

Figure 2 shows an example of the processing of network events through the three phases in IBLT. The IBLT's process starts with the recognition phase in search for decision alternatives to classify a sequence of network events as a cyber attack or not. During recognition, an instance with the highest activation and closest similarity to the network event is retrieved from memory and is used to make this classification. For example, the first instance in memory matches the first network event because it is most similar to the event, and thus, it is retrieved from memory for further processing. Next, in the judgment phase, the retrieved instance is used to evaluate whether the network event currently being evaluated is

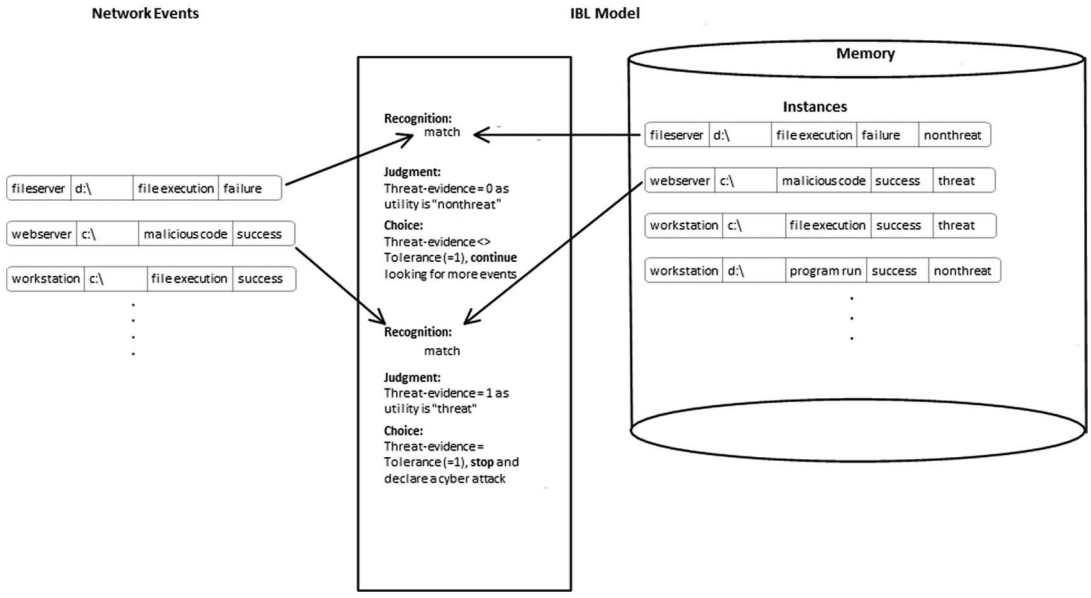

Figure 2. The processing of network events by the Instance-Based Learning model. The model uses recognition, judgment, and choice phases for each observed event in the network, and it decides to stop when the threat-evidence counter equals the tolerance parameter.

perceived as a threat or not. As seen in Figure 2, this evaluation is based upon the U part of the retrieved instance (as described previously, the instance's U part indicates the expectation whether the network event is a "threat" or "nonthreat"). Based upon the U part, a "threat-evidence" counter is incremented by one unit if the network event is classified as a threat; otherwise not. The threat-evidence counter represents the accumulation of evidence for threats and in a new network scenario, it starts at 0. For the first network event in Figure 2, the retrieved instance's U part indicates a nonthreat, so the threat-evidence counter is not incremented and remains at 0. For the second network event, however, the retrieved instance's U part indicates a threat, so the counter is incremented by 1.

In the choice phase, the model decides whether to classify a set of previously evaluated network events in the sequence as part of a cyber attack or to keep accumulating more evidence by further observing network events. In IBLT, this classification is determined by the "necessity level," which represents a satisficing mechanism used to stop search of the environment and be "satisfied" with the current evidence (e.g., the *satisficing strategy*; Simon & March, 1958). This necessity level is the mechanism used to simulate defenders of different tolerance levels. Tolerance is a free parameter that represents the number of network events perceived as threats before the model classifies the sequence as a cyber attack. Therefore, when the threat-evidence counter becomes equal to the tolerance parameter's value, the model classifies the sequence as a cyber attack. For example, for the first network event in Figure 2, the threat-evidence counter (=0) is less than the tolerance (=1) and the model continues to evaluate the next network event. For the second network event, however, the threat-evidence counter (=1) becomes equal to the tolerance (=1) and the model stops and classifies the entire sequence of events as a cyber attack. Translated to actual network environments, the convergence of the threat-evidence counter with tolerance would mean stopping online operations in the company. Otherwise, the model will keep observing more events and let the online operations continue undisrupted if it has not classified a sequence as a cyber attack.

An instance is retrieved in the recognition phase from memory according to an activation mechanism (Gonzalez et al., 2003; Lejarraga et al., 2010). The activation of an instance i in memory is defined using a simplified version of ACT-R's activation equation:

$$A_i = B_i + Sim_i + \varepsilon_i, \quad (1)$$

where i refers to the ith instance that is pre-populated in memory, and $i = 1, 2, \ldots$ constitutes the total number of pre-populated instances in memory; B_i is the base-level learning mechanism and reflects both the recency and frequency of use for the ith instance since the time it was created; and ε_i is the noise value that is computed and added to an instance i's activation at the time of its retrieval attempt from memory.

The B_i equation is given by

$$B_i = ln\left(\sum_{t_i \in \{1,\ldots,t-1\}} (t - t_i)^{-d}\right). \quad (2)$$

In this equation, the frequency effect is provided by $t - 1$, the number of retrievals of the ith instance from memory in the past. The recency effect is provided by $t - t_i$, the time since the tth past retrieval of the ith instance (in Equation 2, t denotes the current event number in the scenario). The d is the decay parameter and has a default value of 0.5 in the ACT-R architecture, and this is the value we assume for the IBL model.

Sim_i refers to the similarity between the attributes of the situation and the attributes of the ith instance. Sim_i is defined as

$$Sim_i = \sum_{l=1}^{k} P_l * M_{li} \quad (3)$$

The $\sum_{l=1}^{k} P_l * M_{li}$ is the similarity component and represents the mismatch between a situation's attributes and the situation (S) part of an instance i in memory. The k is the total number of attributes for a situation event that are used to retrieve the instance i from memory. The value of $k = 4$ as there are four attributes that characterize a situation in the network. As mentioned previously, these attributes are IP, directory, alert, and operation in an event. The match scale (P_l) reflects the amount of weighting given to the similarity between an instance i's situation part l and the corresponding situation event's attribute. P_l is generally a negative integer with a common value of –1.0 for all situation slots k of an instance i, and we assume this value for the P_l. The M_{li} or match similarities represents the similarity between the value l of a situation event's attribute and the value in the corresponding situation part of an instance i in memory. Typically, M_{li} is defined using a squared distance between the situation event's attributes and the corresponding instance's situation slots (Shepard, 1962). Thus, M_{li} is equal to the sum of squared differences between a situation event's attributes and the corresponding instance's S part. In order to find the sum of these squared differences, the situation events' attributes and the values in the corresponding S part of instances in memory were coded using numeric codes. Table 1 shows the codes assigned to the S part of instances and the situation events' attributes.

The noise value ε_i (Anderson & Lebiere, 1998; Lejarraga et al., 2010) is defined as

$$\varepsilon_i = s \times \ln\left(\frac{1-\eta_i}{\eta_i}\right), \quad (4)$$

where η_i is a random draw from a uniform distribution bounded in [0, 1] for an instance i in memory. We set the parameter s in an IBL model to make it a part of the activation equation (Equation 1). The s parameter has a default value of 0.25 in the ACT-R architecture, and we assume this default value in the IBL model. We use default values of d and s parameters.

EXECUTION AND RESULTS OF THE IBL MODEL

The IBL model representing a simulated defender was created using Matlab. The model runs over different network event sequences that represent the different timing strategies of attack as described previously. All sequences contained 25 network events. The model's memory was prepopulated with instances representing defend-

TABLE 1: The Coded Values in the S Part of Instances in Memory and Attributes of a Situation Event

Attributes	Values	Codes
IP	Webserver	1
	Fileserver	2
	Workstation	3
Directory	Missing value	–100[a]
	File X	1
Alert	Present	1
	Absent	0
Operation	Successful	1
	Unsuccessful	0

[a]When the value of an attribute was missing, then the attribute was not included in the calculation of similarity.

ers with different experiences, and the model used different levels of tolerance. The IBL model used Equations 1, 2, and 3 to retrieve an instance with the highest activation and made a decision about whether an event is a threat or nonthreat. The proportion of threat and nonthreat instances in the model's prepopulated memory classified it into two kinds: threat-prone model, whose memory consisted of 90% of threat instances and 10% of nonthreat instances for each network event in the sequence, and nonthreat-prone model, whose memory consisted of 10% of threat instances and 90% of nonthreat instances for each situation event in the sequence. Although we assumed that the 90% and 10% classification for threat-prone and nonthreat-prone models as extreme values, one could readily change this assumption to other intermediate values (between 90% and 10%) in the model.

After an instance was retrieved from memory, a decision was made to classify a sequence as a cyber attack or not depending upon the tolerance and the value of the threat-evidence counter. The tolerance level would classify the model into two kinds: risk-averse (model declares a cyber attack after perceiving one threat) and risk-seeking (model declares a cyber attack after perceiving seven threats). Based upon the aforementioned manipulations, we created four simulated model types: nonthreat-prone and risk-seeking, nonthreat-prone and risk-averse, threat-prone and risk-seeking, and threat-prone and risk-averse. Furthermore, adversarial behavior was simulated by considering different attack strategies about the timing of threats: patient (the last eight events in an event sequence were actual threats) and impatient (the first eight events in an event sequence were actual threats).

We had initially assumed 1,500 simulations in the model; however, we found that by reducing the number of simulations to one third (=500) of that, there was a minuscule change in values of our dependent variables. Thus, we decided to use only 500 simulations of the model as they were sufficient for generating stable model results. We ran 500 simulations (each simulation consisting of 25 network events), and the model's cyberSA was evaluated using its accuracy and detection timing in eight groups defined by: experience (threat-prone and nonthreat-prone), tolerance (risk-averse and risk-seeking), and attacker's strategy (impatient and patient). Accuracy was evaluated by computing the d′ (= Z[hit rate] – Z[false-alarm rate]), hit rate (= hits/[hits + misses]), and false-alarm rate (= false alarms/[false alarms + correct rejections]) (Wickens, 2001) over the course of 25 network events and averaged across the 500 simulations. The model's decision for each network event was marked as a hit if an instance with its U slot indicating a threat was retrieved from memory for an actual threat event in the sequence. Similarly, the model's decision was marked as a false alarm if an instance with its U slot indicating a threat was retrieved from memory for an actual nonthreat event in the sequence. Hits and false alarms were calculated for all events that the model observed before it declared a cyber attack and stopped or when all the 25 events had occurred (whichever came first). In addition to the hit rate, false-alarm rate, and d′, we also calculated the model's accuracy of stopping in different simulations of the scenario. Across the 500 simulations of the scenario, the simulation accuracy was defined as Number of scenarios with a hit/(Number of scenarios with a hit + Number of scenarios with a miss). The model's decision for each simulated scenario (out of 500) was marked as a "scenario with a hit" if the model stopped before observing all the 25 events; otherwise, if the model observed all the 25 events in the scenario, its decision was marked as a

"scenario with a miss." As both the patient and impatient scenarios were only attack scenarios where an attacker attacked the network, we were only able to compute the simulation accuracy in terms of scenarios with a hit or a miss. Furthermore, detection timing was calculated in each simulation as the "proportion of attack steps," defined as the percentage of threat events out of a total 8 that have occurred after which the model classifies the event sequence as a cyber attack and stops. Therefore, higher percentages of attacks steps would indicate the model to be less timely in detecting cyber attacks. Again, we expected a threat-prone and risk-averse model to be more accurate and timely against an impatient attack strategy compared with a nonthreat-prone and risk-seeking model; however, we don't expect that to be the case for a patient attack strategy.

RESULTS

Accuracy

As expected, the attack strategy interacted with the model's type (experience and tolerance) to influence its accuracy. This interaction is illustrated in Figure 3, which shows averages of d′, hit rate, and false-alarm rate, across the 500 simulated participants in each of the eight groups. For an impatient strategy, the d′ was higher for threat-prone models than the nonthreat-prone models, regardless of the risk tolerance (threat-prone risk-seeking: $M = 2.26$, $SE = .05$; threat-prone risk-averse: $M = 2.71$, $SE = .05$; nonthreat-prone risk-seeking: $M = -0.06$, $SE = .05$; nonthreat-prone risk-averse: $M = 0.33$, $SE = .05$). However, for the patient strategy the d′ was higher for the nonthreat-prone models than for the threat-prone models, again regardless of the risk tolerance (threat-prone risk-seeking: $M = -2.63$, $SE = .05$; threat-prone risk-averse: $M = -2.63$, $SE = .05$; nonthreat-prone risk-seeking: $M = -0.29$, $SE = .05$; nonthreat-prone risk-averse: $M = -0.35$, $SE = .05$). These results suggest that the nonthreat-prone model is unable to recognize threats from nonthreats for both patient and impatient attack strategies. In all cases of the nonthreat-prone models, the hit rates and false-alarm rates are very low. Similarly, in the patient strategy, the accuracy (d′) is very low. The models show very high false-alarm rates and very low hit rates. Also, as expected it is only when the attack strategy is impatient and the model has a threat-prone and risk-averse disposition that the d′ is the highest.

Furthermore, we compared the effect of model type and attack strategy on the model's simulation accuracy of stopping. The simulation accuracy was higher for the impatient strategy ($M = 60.70\%$, $SE = .01$) compared with the patient strategy ($M = 50.60\%$, $SE = .01$). Also, the simulation accuracy was higher was for threat-prone models (96.95%) compared to nonthreat-prone models (16.75%) and risk-averse models (59.55%) compared to risk-seeking models (54.15%). However, the attack strategy did not interact with the model type to influence the simulation accuracy. Thus, irrespective of the attack strategy (patient or impatient), the threat-prone and risk-averse models performed more accurately compared to nonthreat-prone and risk-seeking models.

Timing

Again as expected, the attack strategy interacted with the model type (experience and tolerance) to influence the proportion of attack steps. Figure 4 shows the nature of this interaction across 500 simulated participants in each of the eight groups. For the impatient strategy, it mattered whether models were threat- or nonthreat-prone, as well as whether they were risk-averse or risk-seeking, whereas for the patient strategy, it only mattered whether the models were threat- or nonthreat-prone, irrespective of whether they were risk-seeking or risk-averse. For the impatient strategy, the proportions of attack steps needed by threat-prone models (53.15%) were much less than those needed by nonthreat-prone models (92.60%). Also, for the impatient strategy, the proportions of attack steps needed by risk-averse models (50.90%) were much less compared with risk-seeking models (94.85%). For the patient strategy, however, although the proportion of attack steps needed by threat-prone models (10.10%) were much less than nonthreat-prone models (88.80%), there were no differences in the proportion of attack steps between risk-averse (48.80%) and risk-seeking (50.10%) models. In general, as expected, the threat-prone and risk-averse model used the least proportion of attack steps irrespective of the attack strategy, patient or impatient.

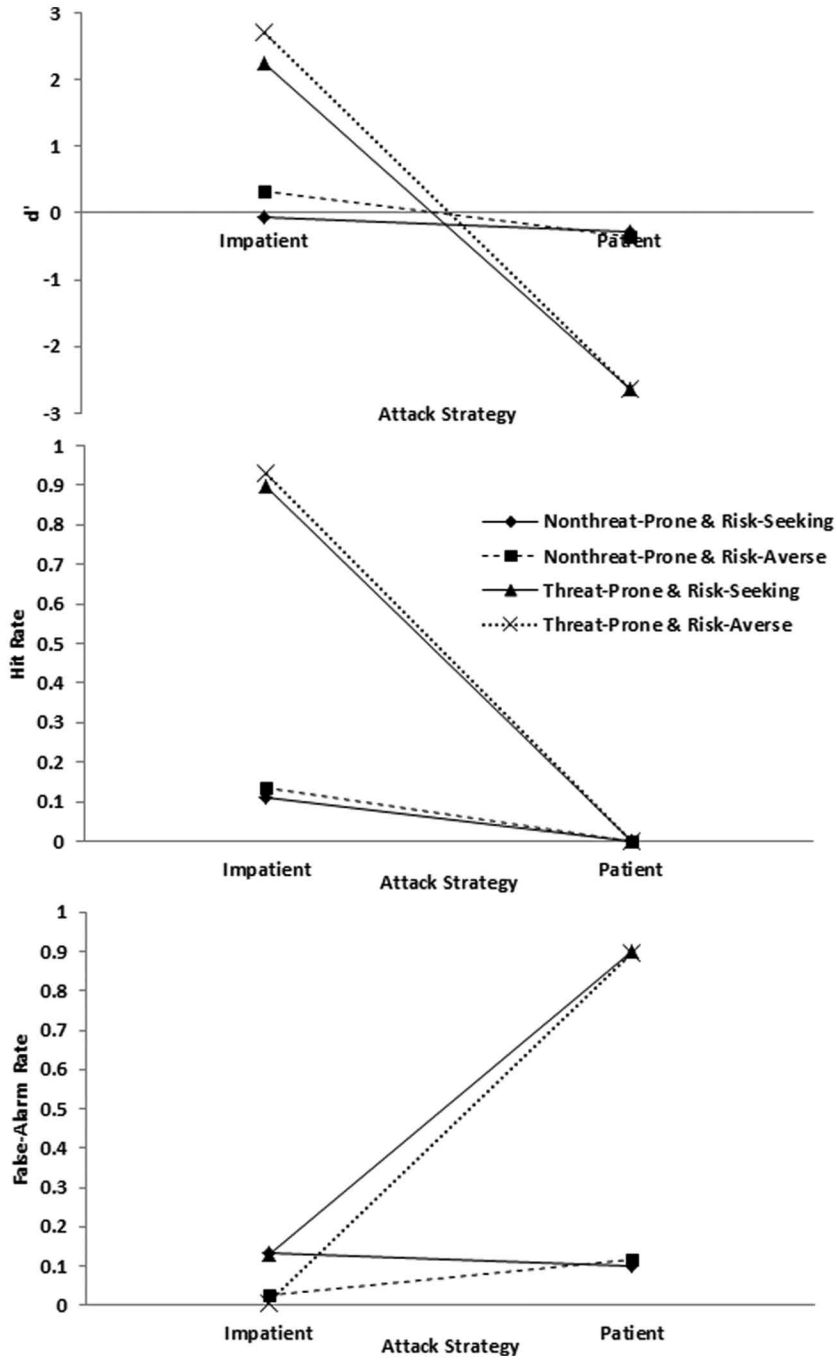

Figure 3. The influence of model type and the attack strategy on model's accuracy.

DISCUSSION

Cyber attacks are becoming increasingly common and they might cause major disruption of work and the loss of important information. Therefore, it is important to investigate defender and adversarial behaviors that influence the accurate and timely detection of network threats. In this endeavor, Instance-Based

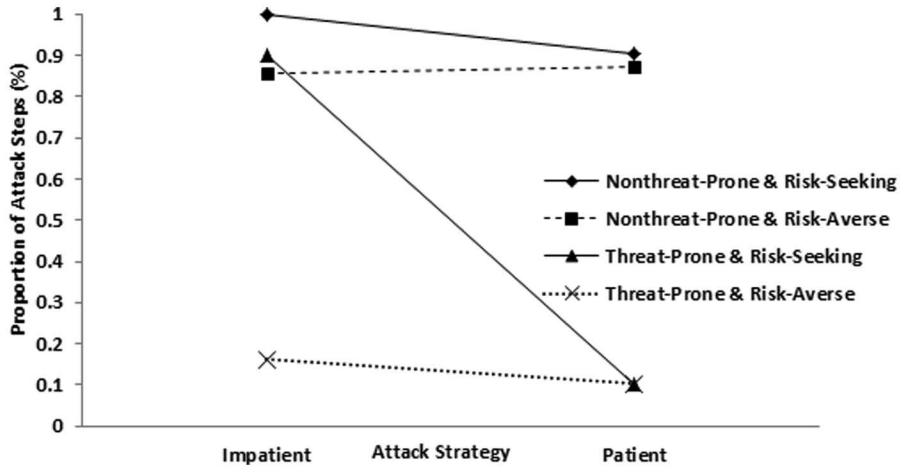

Figure 4. The influence of model type and the attack strategy on proportion of attack steps.

Learning Theory predicts that both defender and adversary behaviors are likely to influence the defender's accurate and timely detection of threats (i.e., cyberSA). Results from an IBL model predict that defender's cognitive abilities, namely, experience and tolerance, and the attacker's strategy about timing of threats together impact a defender's cyberSA.

First, we found that the model's accuracy (d′) is positive only when the attack strategy is impatient and the model has a threat-prone disposition regardless of the model's tolerance. This result is explained given the influence of recency of information on decisions in the model (Dutt, Ahn, et al., 2011; Gonzalez & Dutt, 2011). That is because an impatient strategy's early threats would increase the activation of threat instances in the threat-prone model's memory early on, and the early threats are also likely to increase the chances that the accumulation of evidence for threats would exceed the model's tolerance level, irrespective of whether it is risk-seeking or risk-averse. Therefore, both factors are likely to make the model perform more accurately against an impatient strategy when its cognitive disposition is threat-prone, irrespective of its risk tolerance.

Second, we found that the proportion of attack steps was influenced by both memory and tolerance against the impatient strategy, whereas only the memory seemed to influence the proportion of attack steps against the patient strategy. We believe the likely reason for this observation is the fact that when threats occur early, the accumulation of evidence builds up toward the tolerance level and it influences the timing of detection; however, when threats occur late, the accumulation of evidence might have already reached the tolerance level causing the model to stop much before encountering these late occurring threats. This observation is supported by increased number of false alarms, as well as the model needing lesser proportion of attack steps against the patient strategy (i.e., when the threats occurred late).

Also, we found an interaction between different attack strategies and the model's type: For an impatient attack strategy, possessing threat-prone experiences helped the model's accuracy (due to high hit rates), whereas possessing threat-prone experiences hurt the model's accuracy against a patient strategy (due to high false-alarm rates). This result is expected given that when threats occur early, possessing a majority of threat instances in the model increases the likelihood of detecting these threats early on. Moreover, based on the same reasoning, increasing the likelihood of detecting threats causes the model to detect these threats earlier, which hurts the accuracy when these threats actually occur late in the attack.

Another important observation is that the model possessing nonthreat-prone experiences seemed to show a close to zero d′ irrespective of the attack strategy. A probable reason for this

observation is the following: Having lesser proportion of threat experiences in memory would make it difficult for the model to retrieve these experiences whether the attack occurs early or late. Thus, the overall effect would be a decrease in ability to detect threats from nonthreats when the proportion of threat experiences in memory is low. Indeed we find that the hit rate in the model possessing nonthreat-prone experiences is very low.

Our results are clear in the one way to improve defenders' performance: It is important to train defenders with cases involving multiple threats that would results in a threat-prone memory and prepare them for impatient attackers. Furthermore, it is important to determine the defender's tolerance to risk, as that will determine how timely the defenders address an attack from impatient attackers. Unfortunately, as our results show, these two requirements would not be sufficient to prepare defenders for patient attacker strategies. Because the model's predicted accuracy (d′) is low in all cases in which the attacker follows a patient strategy, we would need to determine better ways to improve accuracy in these cases. Being trained with a threat-prone memory would not be enough in this case, given the high number of false alarms produced in this type of training, although fortunately only a small number of steps would be needed to determine an attack in these cases.

Although in our experimental manipulations we have simulated defenders with characteristics of memory and tolerance that varied at two opposite ends of the spectrum of several possibilities, one could easily modify our defender characteristics in the model to intermediate values. Thus, for example, the threat-prone and nonthreat-prone defenders could each have a 60%–40% and 40%–60% mix of threat and nonthreat instances in memory rather than the currently assumed 90%–10% and 10%–90% mix. Even if one changes this mix to intermediate values, we believe the direction of results obtained would agree with our current results. Second, recent research in cyber security has led to develop methods for correlating alerts generated by IDS sensors into attack scenarios and these methods seem to greatly simplify security defenders' job functions (Albanese, Jajodia, Pugliese, & Subrahmanian, 2011; Debar & Wespi, 2001; Ning, Cui, & Reeves, 2002). In future work, we plan to consider analyzing a defender's behavior with respect to these newer tools. Third, given the low d′ values we could attempt to improve the model's performance by using feedback for the decisions made. Feedback was not provided because defenders in the real world do not get this feedback during a real-time cyber attack (and might only learn about the attack after it has occurred). However, we do use a squared similarity assumption in the model and this assumption enables the model to observe the different attributes of network events. We believe that this similarity mechanism allows the model to produce some differences between hit and false-alarm rates on account of the memory and tolerance manipulations.

If our model's predictions on defender behavior (experiences and tolerance) are correct and the model is able to represent the cyberSA of human defenders, then it would have significant potential to contribute toward the design of training and decision-support tools for analysts. Based on our model predictions, it might be better to devise training and decision-support tools that prime analysts to experience more threats in a network. Moreover, our model's cyberSA was also impacted by how risk-seeking it was to the perception of threats. Therefore, companies recruiting analysts for network-monitoring operations would benefit by evaluating the defender's risk-seeking/risk-aversion tendencies by using risk measures like BART (Lejuez et al., 2002) or DOSPERT (Blais & Weber, 2006). Furthermore, although risk-orientation may be a person's characteristic (like personality), there might be training manipulations that could make defenders conscious of their risk-orientation or alter it in some ways.

At present, we know that predictions generated from the model in this article need to be validated against real human data; however, it is difficult to study real-world cyber-attack events because these occurrences are uncertain, and many attacks occur on proprietary networks where getting the data after they have occurred raises ownership issues (Dutt, Ahn, et al., 2011). Yet, as part of future research, we plan to run simulated laboratory studies assessing human behavior in situations involving different adversarial strategies that differ in the timing of threats. An experimental approach involving

human participants (even if not real defenders) will allow us to validate our model's predictions and improve its relevance and the default assumptions made with its free parameters. In these studies, we believe that some of the interesting factors to manipulate would include the threat/nonthreat experiences stored in memory. One method is to provide training to participants on scenarios that present them with a greater or smaller proportion of threats before they actually participate in detecting threats in island-hopping attacks (i.e., priming memory of participants with more or less threat instances as we did in the model). Also, we plan to record the participants' risk-seeking and risk-averse behavior using popular measures involving gambles to control for their tolerance level (typically a risk-seeking person is more tolerant to risks compared with a risk-averse person). Thus, our next goal in this research program will be to validate our model's predictions.

ACKNOWLEDGMENTS

This research was supported by the Multidisciplinary University Research Initiative Award on Cyber Situation Awareness (MURI; #W911NF-09-1-0525) from Army Research Office to Cleotilde Gonzalez. The authors are thankful to Noam Ben Asher and Hau-yu Wong, Dynamic Decision Making Laboratory, Carnegie Mellon University, for providing insightful comments on an earlier version of the article.

KEY POINTS

- Due to most corporate operations becoming online, the threat of cyber attacks is growing; a key element in keeping online operations safe is the cyber security awareness (cyberSA) of a defender, who is in charge of monitoring online operations.
- The defender's cyberSA is measured by his or her accurate and timely detection of cyber attacks before they affect online operations. It is likely influenced by the defender's behavior (experience and tolerance level) and adversary's behavior (strategies about different timing of threats).
- Defenders who are risk-averse and possess prior threat experiences are likely to improve their detection performance in situations involving impatient attackers; however, not in situations involving patient attackers.
- A cognitive model based on the Instance-Based Learning Theory (IBLT) represents a simulated defender. The model is simulated 500 times each for the different combination of the adversary's and defender's behaviors. This experiment generates predictions about the effects of those behaviors on the defender's cyberSA.
- Application of our results include the design of training tools that increase defenders' competency and the development of decision-support tools that improve their on-job performance in detecting cyber attacks.

REFERENCES

Albanese, M., Jajodia, S., Pugliese, A., & Subrahmanian, V. S. (2011). Scalable analysis of attack scenarios. In *Proceedings of the 16th European Conference on Research in Computer Security* (pp. 416–433). Leuven, Belgium: Springer-Verlag Berlin. doi:10.1007/978-3-642-23822-2_23

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., & Lebiere, C. (2003). The Newell test for a theory of mind. *Behavioral and Brain Sciences*, *26*(5), 587–639. doi:10.1017/S0140525X03000128

Blais, A.-R., & Weber, E. U. (2006). A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*(1), 33–47.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Debar, H., & Wespi, A. (2001). Aggregation and correlation of intrusion-detection alerts. In *Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection* (pp. 85–103). Davis, CA: Springer-Verlag. doi:10.1007/3-540-45474-8_6

Dutt, V., Ahn, Y. S., & Gonzalez, C. (2011). Cyber situation awareness: Modeling the security analyst in a cyber-attack scenario through Instance-Based Learning. *Lecture Notes in Computer Science*, *6818*, 280–292. doi:10.1007/978-3-642-22348-8_24.

Dutt, V., Cassenti, D. N., & Gonzalez, C. (2011). Modeling a robotics operator manager in a tactical battlefield. In *Proceedings of the IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support* (pp. 82–87). Miami Beach, FL: IEEE. doi:10.1109/COGSIMA.2011.5753758

Dutt, V., & Gonzalez, C. (in press). Cyber situation awareness: Modeling the security analyst in a cyber attack scenario through Instance-Based Learning. In C. Onwubiko & T. Owens (Eds.), *Situational awareness in computer network defense: Principles, methods and applications*. Hershey, PA: IGI Global.

Dutt, V., Yu, M., & Gonzalez, C. (2011). Deciding when to escape a mine emergency: Modeling accumulation of evidence about emergencies through Instance-based Learning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 841–845. doi:10.1177/1071181311551175

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factor*, *37*(1), 32–64. doi:10.1518/001872095779049543

Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution*. New York, NY: Basic Books.

Gibson, O. (2011, January 19). London 2012 Olympics faces increased cyber attack threat. *The Guardian*. Retrieved from http://www

.guardian.co.uk/uk/2011/jan/19/london-2012-olympics-cyber-attack

Gonzalez, C. (2012). Training decisions from experience with decision making games. In P. Durlach & A. M. Lesgold (Eds.), *Adaptive technologies for training and education* (pp. 167–178). New York, NY: Cambridge University Press.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating decisions from experience in sampling and repeated choice paradigms. *Psychological Review*, *118*(4), 523–551. doi:10.1037/a0024558

Gonzalez, C., Dutt, V., & Lejarraga, T. (2011). A loser can be a winner: Comparison of two instance-based learning models in a market entry competition. *Games*, *2*(1), 136–162. doi:10.3390/g2010136

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635. doi:10.1016/S0364-0213(03)00031-4

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. doi:10.1111/j.0956-7976.2004.00715.x

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55. doi:10.1016/0010-0285(92)90002-J

Jajodia, S., Liu, P., Swarup, V., & Wang, C. (2010). *Cyber situational awareness*. New York, NY: Springer.

Johnson-Laird, P. (2006). *How we reason*. London, UK: Oxford University Press.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2010). Instance-based learning: A general model of decisions from experience in repeated binary choice. *Journal of Behavioral Decision Making*, *23*, 1–11. doi:10.1002/bdm.722

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., & . . . Brown, R. A. (2002). Evaluation of a behavioral measure of risk-taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. doi:10.1037/1076-898X.8.2.75

McCumber, J. (2004). *Assessing and managing security risk in IT systems: A structured methodology*. Boca Raton, FL: Auerbach Publications.

Ning, P., Cui, Y., & Reeves, D. S. (2002). Constructing attack scenarios through correlation of intrusion alerts. In *Proceedings of the 9th ACM Conference on Computer & Communications Security (CCS '02)* (pp. 245–254). Washington, DC: ACM. doi:1-58113-612-9/02/0011.

Ou, X., Boyer, W. F., & McQueen, M. A. (2006). A scalable approach to attack graph generation. In *Proceedings of the 13th ACM Conference on Computer and Communications Security* (pp. 336–345). Alexandria, VA: ACM. doi:10.1145/1180405.1180446

PSU. (2011). Center for cyber-security, information privacy, and trust. Retrieved from http://cybersecurity.ist.psu.edu/research.php

Salter, C., Saydjari, O., Schneier, B., & Wallner, J. (1998). Toward a secure system engineering methodology. In *Proceedings of New Security Paradigms Workshop* (pp. 2–10). Charlottesville, VA: ACM. doi:10.1145/310889.310900

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, *2*, 125–140. doi:10.1007/BF02289630

Sideman, A. (2011). *Agencies must determine computer security teams in face of potential federal shutdown*. Retrieved from http://fcw.com/articles/2011/02/23/agencies-must-determine-computer-security-teams-in-face-of-shutdown.aspx

Simon, H. A., & March, J. G. (1958). *Organizations*. New York, NY: Wiley.

Tadda, G., Salerno, J. J., Boulware, D., Hinman, M., & Gorton, S. (2006). Realizing situation awareness within a cyber environment. In *Proceedings of SPIE Vol. 6242* (pp. 624204). Orlando, FL: SPIE. doi:10.1117/12.665763

White House, Office of the Press Secretary. (2011). *Remarks by the President on securing our nation's cyber infrastructure*. Retrieved from http://www.whitehouse.gov/the_press_office/Remarks-by-the-President-on-Securing-Our-Nations-Cyber-Infrastructure/

Wickens, T. D. (2001). *Elementary signal detection theory*. New York, NY: Oxford University Press.

Xie, P., Li, J. H., Ou, X., Liu, P., & Levy, R. (2010). Using Bayesian networks for cyber security analysis. In *Proceedings of the 2010 IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (pp. 211–220). Hong Kong, China: IEEE Press. doi:10.1109/DSN.2010.5544924

Varun Dutt received his PhD in engineering and public policy from Carnegie Mellon University in 2011. He is an assistant professor at the School of Computing and Electrical Engineering and School of Humanities and Social Sciences, Indian Institute of Technology, Mandi, India. Prior to this assignment, he worked as a postdoctoral fellow at the Dynamic Decision Making Laboratory, Carnegie Mellon University. He is also the knowledge editor of the English daily *Financial Chronicle*. His current research interests are in dynamic decision making and modeling human behavior.

Young-Suk Ahn is an MS (software engineering) student in the School of Computer Science at Carnegie Mellon University. He is also the CEO of Altenia Corporation, Panama. His current research interests focus on dynamic decision making and software engineering.

Cleotilde Gonzalez received her PhD in management information systems from Texas Tech University in 1996. She is an associate research professor and director of the Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University. She is affiliated faculty at HCII, CCBI, and CNBC. She is on the editorial board of *Human Factors* and is associate editor of the *Journal of Cognitive Engineering and Decision Making*.

Date received: October 31, 2011
Date accepted: August 19, 2012