Learning About the Effects of Alert Uncertainty in Attack and Defend Decisions via Cognitive Modeling

Palvi Aggarwal, Carnegie Mellon University, Pennsylvania, USA, **Frederic Moisan**, AIM Institute, EM Lyon Business School, GATE, **Cleotilde Gonzalez**, Carnegie Mellon University, Pennsylvania, USA, and **Varun Dutt**, Indian Institute of Technology Mandi, Himachal Pradesh, India

Objective: We aim to learn about the cognitive mechanisms governing the decisions of attackers and defenders in cybersecurity involving intrusion detection systems (IDSs).

Background: Prior research has experimentally studied the role of the presence and accuracy of IDS alerts on attacker's and defender's decisions using a game-theoretic approach. However, little is known about the cognitive mechanisms that govern these decisions.

Method: To investigate the cognitive mechanisms governing the attacker's and defender's decisions in the presence of IDSs of different accuracies, instance-based learning (IBL) models were developed. One model (NIDS) disregarded the IDS alerts and one model (IDS) considered them in the instance structure. Both the IDS and NIDS models were trained in an existing dataset where IDSs were either absent or present and they possessed different accuracies. The calibrated IDS model was tested in a newly collected test dataset where IDSs were present 50% of the time and they possessed different accuracies.

Results: Both the IDS and NIDS models were able to account for human decisions in the training dataset, where IDS was absent or present and it possessed different accuracies. However, the IDS model could accurately predict the decision-making in only one of the several IDS accuracy conditions in the test dataset.

Conclusions: Cognitive models like IBL may provide some insights regarding the cognitive mechanisms governing the decisions of attackers and defenders in conditions not involving IDSs or IDSs of different accuracies.

Application: IBL models may be helpful for penetration testing exercises in scenarios involving IDSs of different accuracies.

Keywords: cybersecurity, behavioral game theory, instance-based learning theory, alerts

Address correspondence to Palvi Aggarwal, Dynamic Decision Making Laboratory, Carnegie Mellon University, Pittsburgh, PA 15213, USA; e-mail: aggarwalpalvi12@gmail.com

HUMAN FACTORS
Vol. 00, No. 0, Month XXXX, pp. 1-17
DOI:10.1177/0018720820945425
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

There is a need for better decision support tools in an increasingly complex cybersecurity world, where large amounts of data are collected by network sensors while defender's cognitive abilities are limited (Gonzalez et al., 2014; Sawyer & Hancock, 2018; Sawyer et al., 2015). An important technology that may help support defender's detection of threats is the intrusion detection system (IDS; Bhatt et al., 2011; Mukherjee et al., 1994). IDSs provide the human defender with alerts that may signal potential cyberattacks. Although IDSs may reduce the workload of defenders in cyber threats detection, they are also prone to inaccuracies such as false alarms and misses (Jajodia et al., 2010; Roy et al., 2010). In fact, false and negative information from IDSs and the associated mental workload on defenders may severely influence their decision-making (Finomore et al., 2013; Mancuso et al., 2013).

Prior human factors research has found alerts, alarms, and decision aids to be an important component of decision-making processes (Meyer et al., 2014). People often trust alerts and rely and comply with them to make decisions (Meyer, 2004). Compliance is the tendency to respond as if there is a problem when an alert is generated, for example, defending the network when the IDS generates an alert for a cyber threat. However, reliance is a tendency to continue with the normal activity when an alert is absent, for example, not defending the network when the IDS does not generate any alerts. Sometimes, alerts may have inaccuracies such as misses and false alarms, which could affect reliance and compliance differently. Human factors research has investigated how inaccuracies in general alarm systems affects human reliance

and compliance on such systems (Chancey et al., 2017; Wiczorek et al., 2014; Wiczorek & Meyer, 2016). For example, Wiczorek and Meyer (2016) found that alarm systems' false alarms and misses create an asymmetry bias as false alarms affect both compliance and reliance; however, misses affect only reliance and not compliance.

In this research, we aim at advancing our understanding about the cognitive mechanisms underlying the decisions of attackers and defenders when relying on the alerts of an IDS-like system in a cybersecurity context. Specifically, we use computational cognitive models to emulate the process by which humans make decisions in the presence of alerts of diverse accuracy in a two-player cybersecurity game. The cognitive models developed help elucidate the mechanisms involved in trusting alerts while making attack or defend decisions in a cybersecurity setting.

BACKGROUND

The current research relies and advances the research from Dutt et al. (2016) (hereafter, DMG). DMG studied the impact of presence and accuracy of IDS alerts on attacker's and defender's decisions in a two-person game. DMG varied the availability of IDS alerts as well as the accuracy of IDS alerts when IDS was present. When IDSs were absent, there were no IDS alerts; whereas, when IDSs were present, they generated alerts against the hacker's actions (i.e., attack or not-attack). These alerts were of various accuracies, where IDSs correctly alerted defenders about cyberattacks or no-attacks on 10%, 50%, or 90% of the trials. DMG found that the proportion of defend actions was similar when IDS was absent and when it was only 50% accurate. However, the proportion of defend actions was reduced when the IDS was mostly inaccurate (10% accuracy) or very accurate (90% accuracy). The proportion of attack actions were similar across all conditions.

The human factors literature (e.g., Bliss et al., 2002; Meyer, 2004) would suggest that results from DMG show reliance and compliance with IDS alerts when they are accurate and show mistrust by taking the reverse action when IDS alerts are inaccurate. Contrary to defenders, attackers do not get IDS alerts, but they receive "feedback" regarding the success of their attacks and what an IDS is potentially reporting to the defender. In DMG's results, the attackers reduced their attack actions when the IDS alerts were highly accurate. Attackers were able to learn when the defender relied on highly accurate IDS alerts and acted accordingly to reduce their attack actions in the fear of being caught.

In the current research, we will try to explain and advance DMG's findings in three ways. First, we use a theoretical perspective of decisions from experience, building cognitive models that represent the attacker's and defender's decision processes; specifically, we aim at explaining why the proportion of defend actions were reduced when IDSs were highly accurate and IDSs were highly inaccurate, and why the attack proportions reduced when IDS alerts were highly accurate. We develop a cognitive model based on instance-based learning theory (IBLT; Gonzalez & Dutt, 2011; Gonzalez et al., 2003), a theory of decisions from experience, to help explain the decisions made by attackers and defenders, while accounting for the inaccuracies or unavailability of IDS alert systems.

Second, we use the data collected in Dutt et al. (2016) to calibrate the IBL models and their cognitive parameters and compare the model's attack and defend actions to those from the human participants in DMG. Third, to demonstrate that this is not simply an overfitting exercise of model to human data, we test the generalization of the model's predictions in a newly collected dataset. However, to address some of the limitations of the DMG design, our new experiment involved a slightly different scenario where the IDS was only available on 50% trials with different levels of accuracies.

THE IDS GAME

The goal of both attackers and defenders in the DMG's IDS game was to maximize their payoffs in the task. In this game, first, the attacker (hacker) made a choice to attack or not-attack a network. The attacker's choice triggered an IDS alert (when IDS is present) for

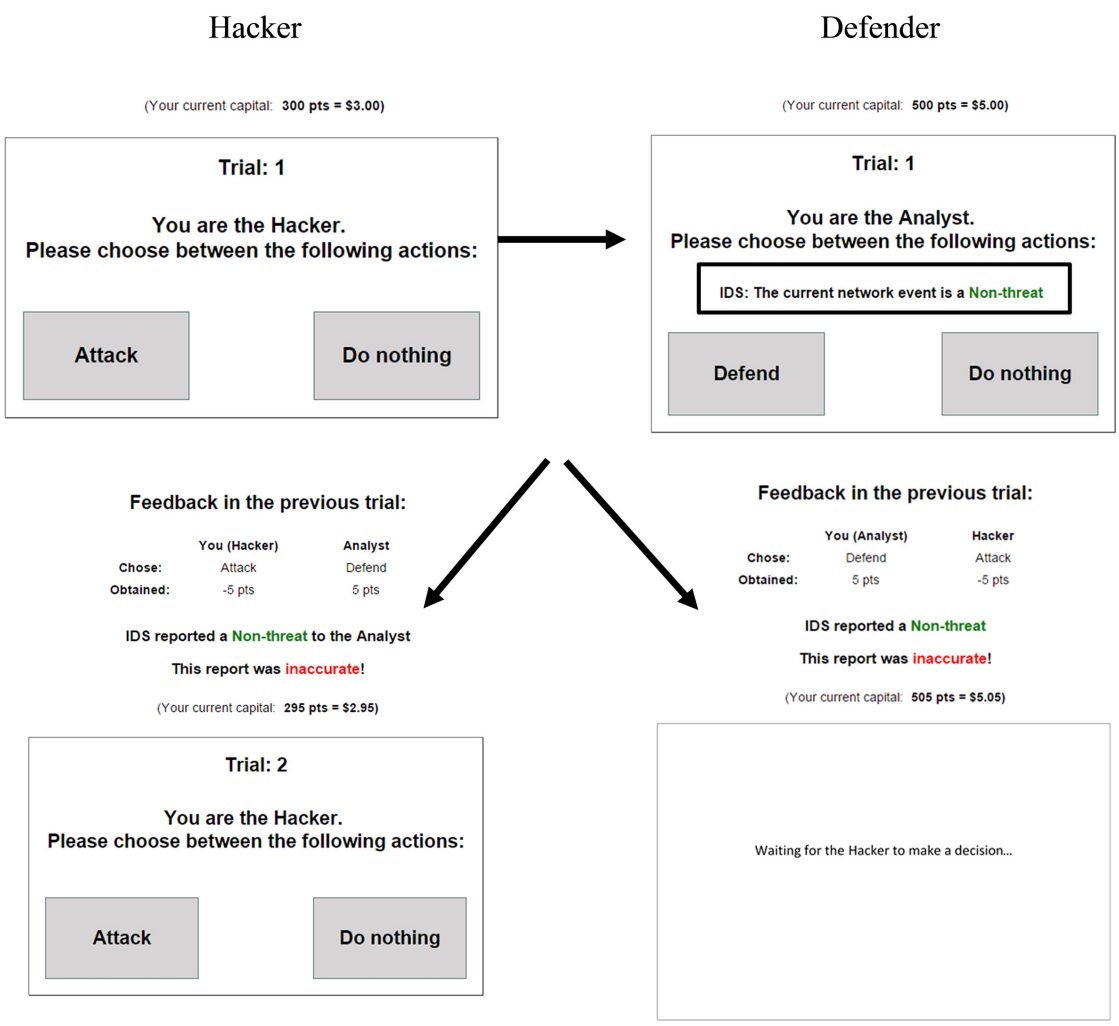

Figure 1. The dynamics of a trial in one of the IDS-present conditions in the cybersecurity game. The arrows indicate the sequence of events in a trial involving an attacker (hacker) and a defender (analyst). IDS = Intrusion Detection System. *Source*. Dutt et al. (2016).

defenders (analysts) indicating whether the network event was a cyber threat or not. Based on IDS alerts, defenders made a choice to defend or not defend the network. Once the defender had made her choice, both players were provided with information about the actions taken and payoffs obtained for both players, IDS alerts from the previous trial (if IDS was present), and the players' own cumulative payoff (opponent's cumulative payoff was not shown; Figure 1). The attackers knew whether the IDS was active or not and it was accurate or not accurate at the time of feedback.

When attackers and defenders took not-attack and not-defend actions respectively, then both players received 0 points. However, if attackers took attack actions while defenders took not-defend actions, then attackers received +10 points and defenders −15 points. In contrast, if defenders took defend actions and attackers took not-attack actions, then attackers received 0 points and defenders −5 points. Finally, if the

defenders took defend actions while attackers took attack actions, then attackers received −5 points and defenders +5 points (Figure 2).

TRAINING AND TEST DATASETS

The training dataset was collected by Dutt et al. (2016) using the IDS game above. We used their data set for model parameter calibration. The test dataset was newly collected in this research for the purpose of testing the calibrated models in novel IDS scenarios.

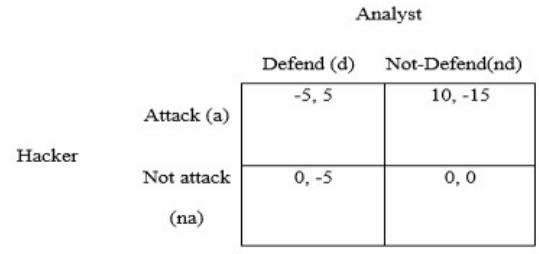

Figure 2. Payoff matrix in a security game played repeatedly between a hacker and an analyst.

Training Dataset

DMG experimentally investigated the impact of the presence and accuracy of IDS alerts on decision-making in a simulated two-player cybersecurity game. The presence of IDS was varied as IDS absent or IDS present and the accuracy was varied as 10% accurate, 50% accurate, and 90% accurate. In the experiment, participants were randomly paired together to act as attackers or defenders. They were assigned to one of the following four between-subject conditions over 100 trials: IDS-absent ($N = 20$ pairs) IDS-present (accuracy: 10 %, $N = 25$ pairs; 50%, $N = 29$ pairs; and 90%, $N = 26$ pairs) to make attack-and-defend decisions. Figure 3 reproduces the proportion of attack and defend actions by human participants and Nash equilibrium reported originally in DMG. In this figure, the dependent variable (y-axis) was computed by first coding the attack/defend action as 1.0 and not-attack/not-defend as 0.0 for each participant and each trial. Next, the proportion of attack/defend actions were computed by averaging the 1.0s and .0s across all

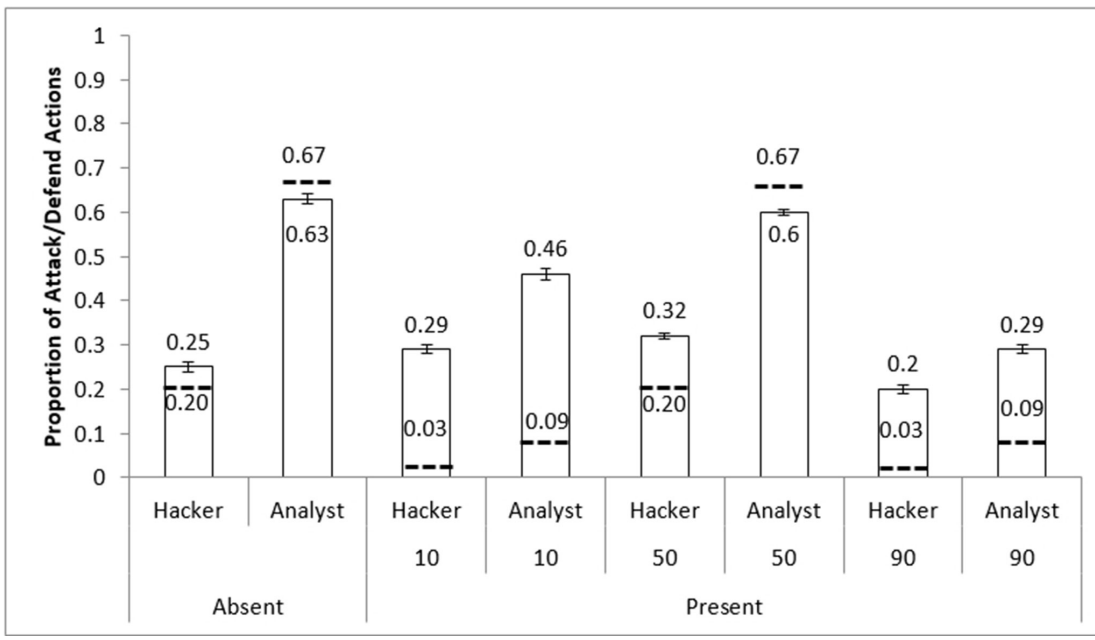

Figure 3. The average proportion of attack or defend actions in human data across different between-subject conditions. The dotted lines show the proportion of optimal Nash actions. *Source* Dutt et al. (2016).

trials and participants in each condition. Ideally, a smaller proportion of attack actions and a larger proportion of defend actions is desirable as this combination reduces cyberattacks.

As seen in Figure 3, the defend proportions were similar when the IDS was absent (0.63) and when it was 50% accurate (0.60). However, the defend proportions reduced when the IDS was present and when it was either inaccurate (10% accurate; 0.46) or accurate (90% accurate; 0.29) compared to when IDS was absent (0.63). The attack proportions were not influenced by the IDS's presence and accuracy. In a majority of the conditions, the human proportions deviated from the optimal Nash actions (dotted lines; reported in Aggarwal et al., 2018; Dutt et al., 2016).

We also analyzed the proportion of attack and defend actions as a function of IDS alerts (Figure 4). When IDS was inaccurate (10% accuracy), both attacker and defender took more (less) attack and defend actions when IDS reported a nonthreat (threat). As the accuracy of IDS increased, the proportion of defend and attack actions reduced (increased) when IDS alert reported nonthreat (threat).

Test Dataset

DMG only considered the situations where IDSs were either completely present or absent. We conducted a new study in this research where the IDS was present randomly in only 50% of the trials with an overall accuracy of 10%, 50%, and 90%. Other procedures were same as those in DMG. Thus, this dataset represents a mixture of two experimental training conditions, where the IDS was present in some trials and absent in others.

Experimental design and procedures. In the test dataset, 136 Amazon MTurk participants performed the task and were randomly paired together to act as attackers or defenders. They were assigned to one of the following three between-subjects IDS conditions over 100 trials: 10% accurate ($N = 22$ pairs), 50% accurate ($N = 21$ pairs); and 90% accurate ($N = 25$ pairs).

Across all conditions, the IDS was present in only 50% of the trials. In trials where the IDS was not present, the IDS alerts were not available to both players. The study was approved by respective committees at the Indian Institute

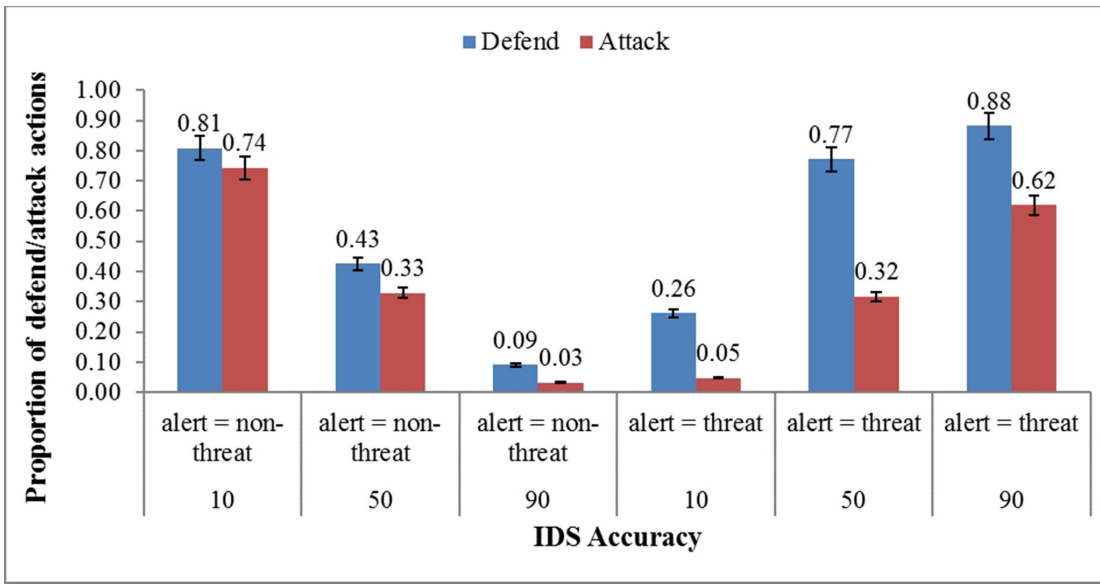

Figure 4. The proportion of defend and attack actions when the IDS issued an alert (threat) and when it did not (nonthreat) in the training dataset.

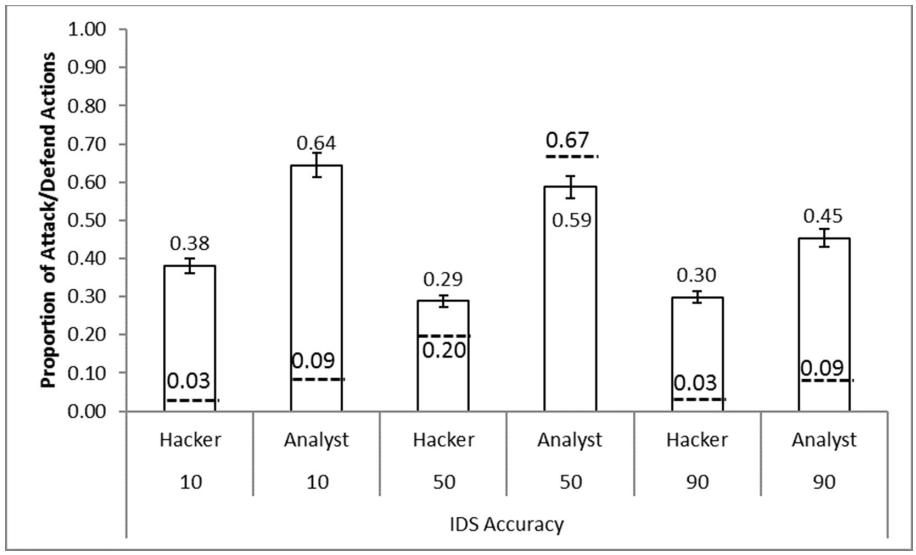

Figure 5. The average proportion of attack and defend actions in human data across different between-subject conditions in the test dataset. The dotted lines show the proportion of optimal Nash actions. The Nash proportion when the IDS was absent were 0.20 and 0.67 for hacker and defender, respectively. The error bars represent a 95% confidence interval around the mean value.

of Technology Mandi and Carnegie Mellon University. Participation was completely voluntary and participants signed a consent form before starting their study. All participants were from STEM backgrounds and their ages ranged from 18 years to 67 years (mean = 34 years; *SD* = 11 years).

Results

Similar to the results from DMG in the training dataset, we found that the proportion of attack actions for human participants were similar in all three IDS accuracy conditions in the test dataset (Figure 5). Similarly, the proportion of defend actions reduced when the IDS was 90% accurate compared to when it was 50% accurate. However, unlike the training dataset, the proportion of defend actions was similar in conditions when the IDS was 10% accurate and 50% accurate.

The proportion of attack and defend actions as a function of IDS alerts in test dataset showed similar results as the training data set (Figure 6). When the IDS was inaccurate (10% accuracy), both attacker and defender took more (less) attack and defend actions when the IDS alert reported nonthreat (threat). As the accuracy of IDS increased from 10% to 50% and 90%, the proportion of defend and attack actions reduced (increased) when IDS alert reported nonthreat (threat).

AN INSTANCE-BASED MODEL OF BINARY CHOICE

We developed cognitive models of defender and attacker using an IBL model of binary choice (Dutt et al., 2013; Gonzalez & Dutt, 2011; Lejarraga et al., 2012). IBL models use the formalization of the memory mechanisms from the adaptive control of thought-rational (ACT-R) cognitive architecture (Anderson & Lebiere, 1998) and the decision process from IBLT (Gonzalez et al., 2003). An instance in the IBL model is a unit of experience consisting of the situation (attributes of task), the decision made in the current situation, and the utility (the outcome of choosing an option in the current situation). In the IBL model, among all options, the option that has the highest expected utility (i.e., blended value) is chosen as a decision. The blended value $V_{k,t}$ of option k at trial t is computed using equation (1):

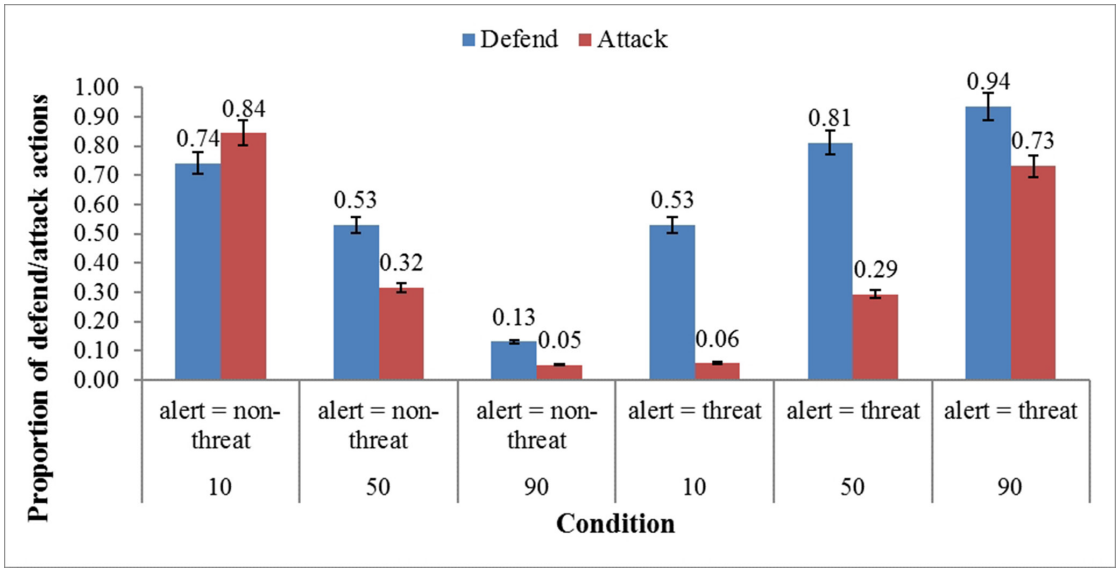

Figure 6. The proportion of defend and attack actions when the IDS issued an alert (threat) and when it did not (nonthreat) in the test dataset.

$$V_{k,t} = \sum_{i=1}^{n} P_{i,k,t} * X_{i,k,t} \quad (1)$$

where $x_{i,k,t}$ represents the outcome of an instance i for option k at trial t (outcome could be −5, 0, + 5, +10 shown in Figure 2) and $p_{i,k,t}$ represents the probability of retrieval of an instance i for option k at any trial t (value of k is either to attack/defend or to not-attack/not-defend).

The retrieval probability of an instance i is the ratio of activation of ith instance corresponding to the activation of all instances (1, 2, … n; *where n is total instances*) created within the option k at trial t. The retrieval probability is defined as

$$P_{i,k,t} = \frac{e^{A_{i,k,t}/\tau}}{\sum_{i=1}^{n} e^{A_{i,k,t}/\tau}} \quad (2)$$

Here, $\tau = \sigma * \sqrt{2}$ and σ is a free noise parameter. Noise captures the inaccuracy of remembering past experiences from memory. At each trial t, activation of an instance i on option k is computed as

$$A_{i,k,t} = ln\left(\sum_{t_p \in \{1,,t-1\}} (t - t_p)^{-d}\right) + s * ln\left(\frac{1-\gamma_{i,k,t}}{\gamma_{i,k,t}}\right) \quad (3)$$

Where d is decay parameter; $\gamma_{i,k,t}$ represents a random number drawn from a uniform distribution between 0 and 1; and t_p represents all the previous trials where the instance i was either created or its activation was reinforced due to its reoccurrence. The numbers of terms in summation correspond to the frequency of observations and the difference between two time periods correspond to the recency of outcomes. The activation of an instance increases with frequency and recency of outcomes. For larger values of d (>1.0), the model pays more attention to recent events compared to the smaller value of d (<1.0). The noise parameter s helps to capture the individual variability in human behavior.

IBL Models for Attackers and Defenders

We developed two IBL models: one model (NIDS) disregarded the IDS alerts and the second model (IDS) considered them in the instance structure. In the NIDS model, the situation part of instances consisted of only actions of attackers and defenders without consideration for IDS alerts. In the IDS model, the situation part of instances consisted of actions of attackers, defenders, and IDS alerts. The NIDS model was calibrated to data in the IDS-absent

condition and the IDS model was calibrated to data in the IDS present conditions.

The models were initialized with two prepopulated instances for each player (attacker: attack/not-attack; defender: defend/not-defend), and the IDS alert slot was initialized to no alert. The outcome of prepopulated instances was calibrated in models as free parameters. Both NIDS and IDS models used prepopulated instances in memory to take decisions in the first few rounds (Lejarraga et al., 2012).

Model fitting. IBL models were created in MATLAB® and possessed four free parameters per player: decay d, noise σ, prepopulated instance values for attack/defend actions, and not-attack/not-defend actions. The model used the same number of model agents as of human participants to play the role of attackers and defenders repeatedly for 100 trials across different between-subject conditions.

First, we found the best set of d and s parameters and prepopulated instance values in different models using the training dataset. The NIDS model was calibrated on IDS-absent condition and the IDS model was calibrated on IDS-present (10%, 50%, and 90% accuracy) conditions in the training dataset. To get the optimized parameters, we minimized the sum of mean-squared deviations (MSDs), which is the squared difference between human and model actions. MSDs was computed over the proportion of attack and defend actions separately using the following equation:

$$MSD = \frac{1}{100}\sum_{t=1}^{100}(Model\ Actions_t - Human\ Actions_t)^2 \quad (4)$$

Where, $Model\ Actions_t$ and $Human\ Actions_t$ refers to the average proportion of attack or defend actions from the model and human data, respectively, in trial t (total trials = 100). The average proportion of attack or defend actions for a trial were computed by averaging these actions across all participants for the trial. A genetic algorithm (Konak et al., 2006) was used to optimize the parameter values for both attacker and defender participants in the game. As done in prior research (Gonzalez & Dutt, 2011), the d and s parameters were varied between 0.0 and 10.0, and the prepopulated instances were varied between 0.0 and 15.0, where the upper bound chosen for prepopulated instances was much higher compared to the outcomes possible in the cybersecurity game. These ranges ensured that the optimization could capture the optimal parameter values with high confidence. The genetic algorithm had a population size of 20, a crossover rate of 80%, and a mutation rate of 1%. The algorithm stopped when any of the following constraints were met: stall generations = 200, function tolerance = 1×10^{-8}, and when the average relative change in the fitness function value over 200 stall generations was less than function tolerance (1×10^{-8}). These assumptions are like other studies in literature where models have been fitted to human data using the genetic algorithm (Aggarwal et al., 2017; Sharma & Dutt, 2017).

Furthermore, we generated predictions from the IDS models across different conditions in the test dataset, and we evaluated the goodness of predictions by comparing them to human data collected in the test dataset using the MSD. For generating predictions, we ran the models in the test dataset with matching parameter values that were determined in the training dataset. For example, the IDS model parameters obtained in the training dataset in the 10% accuracy condition were run in the 10% accuracy condition in the test dataset (similarly for the other two accuracies).

Hypotheses

Based on the IBLT, we propose the following hypotheses from the models:

H1: We expect the defender's model to take more defend actions compared to not-defend actions, except when guided by accurate IDS alerts. In addition, we expect the attacker's model to take less attack actions in all the conditions to avoid negative rewards.

H2: We expect the defenders to rely on their past experiences when IDS alerts are absent (i.e., because these players do not have any other source of information to rely upon in the absence of IDS alerts).

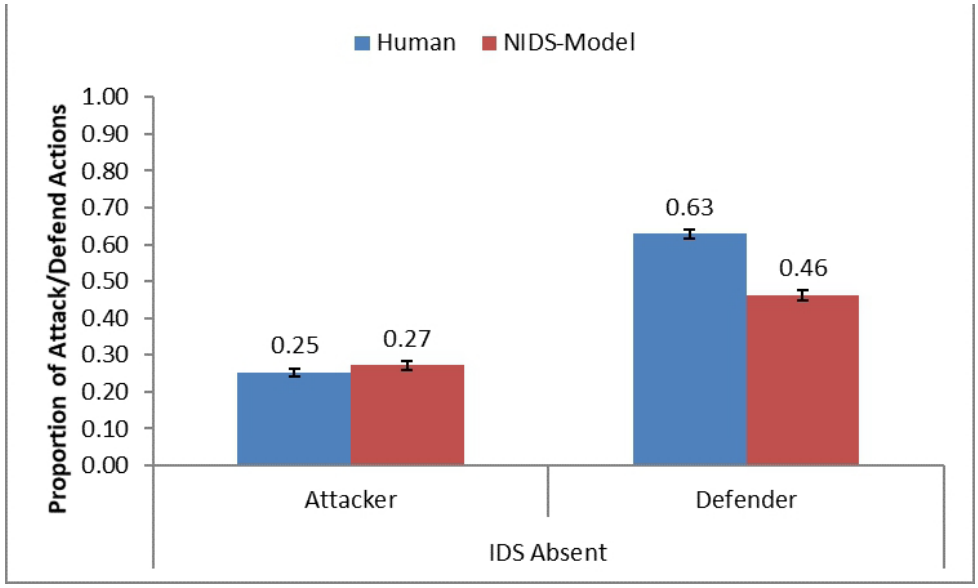

Figure 7. The proportion of defend and attack actions for the NIDS model and human participants in the IDS-absent condition. The error bars represent a 95% confidence interval around the mean.

Thus, the NIDS model of defender is expected to exhibit low memory decay (*d*) leading to high primacy effect.

H3: We expect the attackers to rely on recent actions of defenders to learn if they defend more often in the absence of IDS alerts (i.e., because attackers do not have information about IDS alerts during feedback). Thus, the NIDS model of attacker is expected to exhibit high memory decay (*d*) leading to recency effect.

H4: We expect the defender to rely more on recent information such as IDS alerts as the accuracy of IDS alerts increases. Thus, the IDS model for defender is expected to show higher value of decay (*d*) for IDS 50% and IDS 90% accuracy conditions compared to IDS 10% accuracy condition.

H5: We expect attackers to always rely on recent information to avoid losses. Thus, both NIDS and IDS models for attackers are expected to show high decay (*d*) value.

MODEL RESULTS

Training Results

NIDS model. Figure 7 shows the attack and defend proportions from the calibrated NIDS model and human participants in the IDS-absent condition. The model defenders showed high defend proportions and the model attackers showed less attack proportions. The model behaved in agreement with our hypothesis H1 and H2. In the absence of IDS alerts, defenders became risk averse and took high defend actions to secure the network. This behavior of defender made attackers cautious and thus the model attacker showed less attack actions. Overall, the NIDS model for attackers was more accurate in replicating human attacker actions (MSD = .0004) compared to human defender's decisions (MSD = .0279). The defend actions from NIDS model are comparatively less that human defend actions as shown in Figure 7.

Table 1 presents the calibrated parameters for attackers and defenders for the NIDS model. The *d* value was higher for attackers compared to defenders. These results suggest greater reliance on recent information for attackers

TABLE 1: NIDS Model Parameters in the IDS-Absent Condition

Condition	Attacker	Defender
IDS absent	d^1 = 3.10, s^2 = 7.10, H_A^3 = 6.80, H_{NA}^4 = 14.93, MSD = .0004	d = .03, s = 9.98, A_D^5 = 14.79, A_{ND}^6 = .07, MSD^7 = .0279

Note. [1]The decay parameter. [2]The noise parameter.
[3]Prepopulated instance value for attack actions.
[4]Prepopulated instance value for not-attack actions.
[5]Prepopulated instance value for defend actions.
[6]Prepopulated instance value for not-defend actions.
[7]MSD calculated as mean square deviation between the model action proportion and human action proportion.

compared to defenders. Attackers relied on recent actions of defender and defenders relied on their past experiences to defend the network. The *s* value was high for both defenders and attackers. Thus, both players showed more variability in their trial-to-trial decisions. These results agree with hypothesis H5.

IDS model. Next, we calibrated the IDS model in each of the three between-subject conditions in the training dataset, using the procedure described above. Figure 8 shows the attack and defend proportions from the IDS model and human participants in various conditions. As hypothesized in H1, the IDS model showed larger (smaller) defend proportions when IDS was 50% accurate (10% or 90% accurate). These results indicate that inaccurate and accurate IDS alerts both create similar results for the defender's actions. In case of accurate IDS alerts, defenders relied and complied with these alerts. However, in case of inaccurate IDS alerts, defenders mistrusted the alerts and took the opposite action. Furthermore, as hypothesized in H1, the attack proportions remained similar across all conditions. This behavior could be explained using cognitive parameters of IBL models.

Table 2 shows the MSDs from the IDS model in different conditions. The IDS model could accurately account for the decisions of both defenders and attackers in different IDS-present training conditions.

We also analyze the model to check if it could account for conditional attack and defend actions as reported in Figures 4 and 6 for the human data. The model was less accurate at accounting for conditional defend actions, that is, defend when IDS said threat and defend when IDS said nonthreat, in all conditions (overall MSD = .076) compared to the other results reported above. However, the model

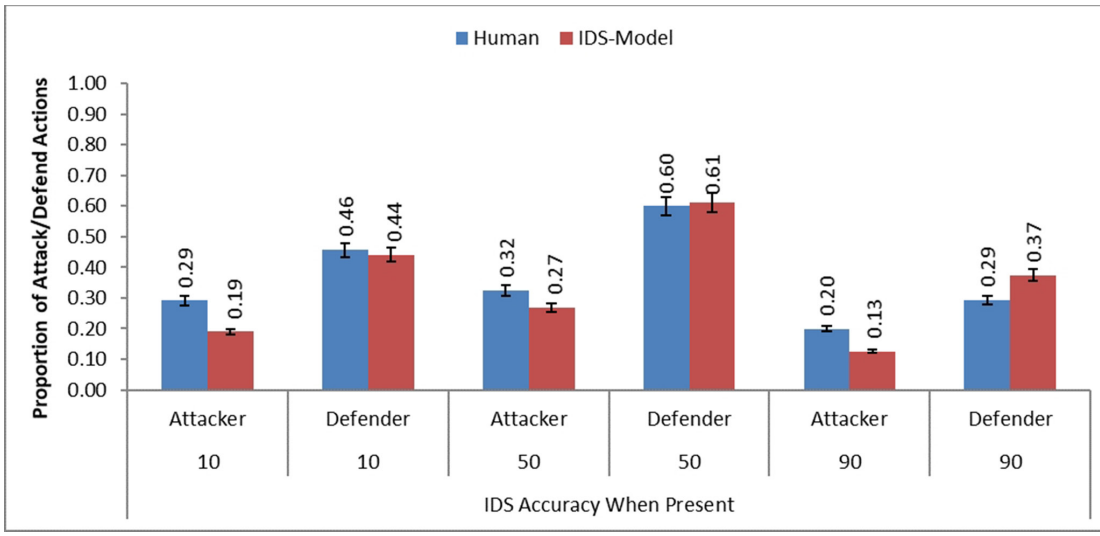

Figure 8. The proportion of defend and attack actions for IDS model and human participants in the IDS-present condition. The error bars represent a 95% confidence interval around the mean. .

TABLE 2: MSDs Obtained in the IDS Model in Different Conditions in the Training Dataset

Condition	Attacker	Defender
IDS present accuracy = 10%	*MSD* = .0103	*MSD* = .0003
IDS present accuracy = 50%	*MSD* = .0030	*MSD* = .0001
IDS present accuracy = 90%	*MSD* = .0055	*MSD* = .0067

Note. IDS = Intrusion Detection System; MSD = Mean-Squared Deviation.

could accurately account for conditional attack actions in all conditions (MSD = .005).

Table 3 shows the calibrated parameters of IDS model across different conditions in the training dataset. As stated in H4 and H5, the *d* value was higher for both attackers and defenders across all the conditions except in the 10% IDS accuracy condition for the defender. This signifies that the attacker relied on recent actions to decide whether to attack or not. The defender's reliance on recent information was high when the accuracy of IDS alerts was high. The *s* value was higher for both attackers and defenders in all conditions except for defender in the 10% IDS accuracy and for attacker in the 90% IDS accuracy.

TEST RESULTS

Figure 9 presents the generalization results from the IDS model for different IDS accuracy conditions in the test dataset. In addition, Table 4 presents the MSDs for the IDS model predictions using human data in the test dataset.

As seen in Table 4, the MSDs were very low when the IDS was 50% accurate. However, the MSDs were considerably higher with the IDS was 10% and 90% accurate. As seen in Figure 9, the model predicted the responses for the 50% condition with admirable accuracy. However, for the two other conditions, the model underestimated the attack and defend proportions consistently. In addition, the model was less accurate at accounting for conditional defend and attack actions across all conditions in the test dataset (conditional defend MSD = .053; conditional attack MSD = .061).

DISCUSSION AND CONCLUSIONS

We build on DMG's work and investigate the effect of different accuracies on the defender's reliance and compliance on the IDS. We found that with increased inaccuracy of the IDS, both the reliance and compliance increased. To understand the decision processes of human defenders and attackers, we use computational models and these models were used to predict human actions in novel situations.

The NIDS model was calibrated to IDS-absent data and the resulting parameters suggested two main effects. First, we found more reliance on recent actions for attackers compared to defenders. It is possible that defenders acted risk-averse and took more defend actions when the IDS was absent. When defenders defend excessively, attackers likely rely on recent information about their own and

TABLE 3: Model Parameters in the IDS Model in Different Conditions in the Training Dataset

Condition	Attacker	Defender
IDS present accuracy = 10%	$d^1 = 9.85$, $s^2 = 2.77$, $H_A{}^3 = .11$, $H_{NA}{}^4 = 12.83$	$d^1 = .19$, $s^2 = .62$, $A_D{}^5 = 12.18$, $A_{ND}{}^6 = 14.99$
IDS present accuracy = 50%	$d^1 = 9.69$, $s^2 = 2.34$, $H_A{}^3 = 8.51$, $H_{NA}{}^4 = 14.16$	$d^1 = 8.34$, $s^2 = 2.90$, $A_D{}^5 = 14.03$, $A_{ND}{}^6 = 13.18$
IDS present accuracy = 90%	$d^1 = 5.65$, $s^2 = .19$, $H_A{}^3 = 9.35$, $H_{NA}{}^4 = 2.52$	$d^1 = 9.90$, $s^2 = 3.65$, $A_D{}^5 = 14.81$, $A_{ND}{}^6 = 13.90$

Note. [1]The decay parameter. [2]The noise parameter. [3]Prepopulated instance value for attack actions. [4]Prepopulated instance value for not-attack actions. [5]Prepopulated instance value for defend actions. [6]Prepopulated instance value for not-defend actions.

TABLE 4: MSDs Obtained in the IDS Model in Different Conditions in the Test Dataset

Condition	Attacker	Defender
IDS present accuracy = 10%	*MSD* = .030	*MSD* = .023
IDS present accuracy = 50%	*MSD* = .000	*MSD* = .001
IDS present accuracy = 90%	*MSD* = .039	*MSD* = .025

Note. IDS = Intrusion Detection System; MSD = Mean-Squared Deviation.

their opponent's actions to decide between their attack or not-attack actions, explaining the higher recency for attackers compared to defenders. Second, we found excessive reliance on recency mechanisms in all conditions for both attackers and defenders except for the defender when the IDS was 10% accurate. This may be explained if attackers and defenders relied on the IDS alert information to increase their utility. When IDSs were inaccurate (10% accurate), the recent experience of IDS alerts was unlikely to be solely helpful for defenders to decide their actions. Thus, defenders also tended to rely on the historical IDS alerts in their memories and experience in order to decide their actions. However, for attackers, the recent actions of defenders, including the recent IDS alerts, likely helped determine their actions. When the IDS was accurate (90% accurate), defenders tended to rely on recent IDS alerts and acted according to these alerts. Similarly, the attackers also relied on IDS alerts to avoid negative rewards. In contrast, when the IDS was 50% accurate, then defenders and attackers likely relied on recent information about the actions of their opponents during feedback to decide their actions in the next round.

We also tested the model in a new dataset where IDS alerts were partially available. The IDS model could not capture the exact proportions in test data particularly for the 10% and 90% accuracy conditions. One likely reason is that the IDS model used the recency and noise parameters as those derived in the training conditions. Perhaps, the partial availability of the IDS in the test data decreased attacker's and defender's reliance on IDS information. Overall, the decay parameter suggests that the defenders would heavily rely on IDS alerts to take their action if the alerts are accurate and available.

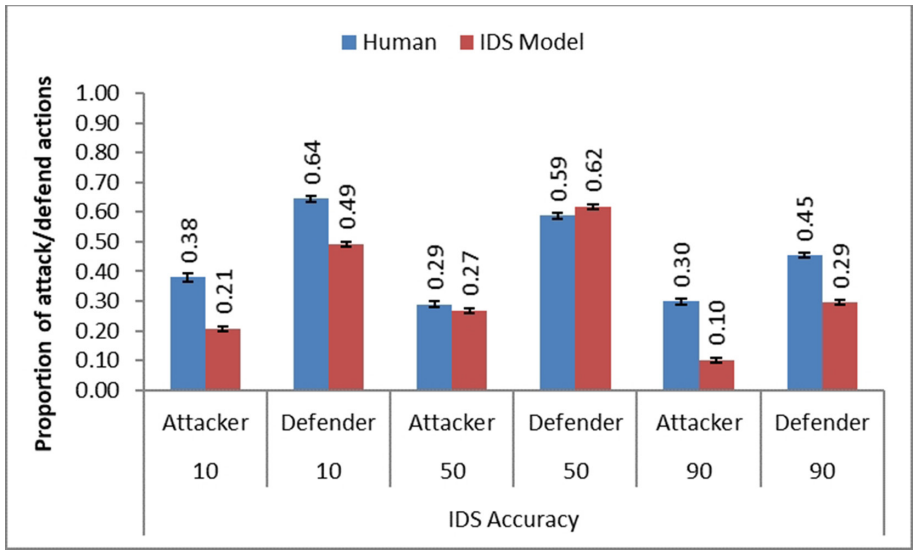

Figure 9. The prediction of attack and defend proportions from the IDS model and their validation from human data collected in the prediction dataset. The error bars represent a 95% confidence interval around the mean.

In addition, the IDS model could only capture for the proportion of attack actions as a function of IDS alerts in only training conditions. It could not capture the proportion of defend actions as a function of IDS alerts in both training and test conditions. This result is likely due to the way the model was fitted to the overall proportion of defend and attack actions in the training data and due to the model's instance structure that only considered the IDS alerts (where humans may also store an expectation about the accuracy of the IDSs in their memories). Overall, the models could be improved in future to account for these observations.

An important implication of this research is that cognitive agents may be used against penetration testing in networks. However, several improvements to the current models will need to be implemented. For example, the models could be modified to capture the overall accuracy of IDS alerts into miss rates and false alarm rates, as defenders and attackers might behave differently to such miss and false alarm rates. In addition to improvements in the model, we would also like to study how misses and false alarms affect defender's reliance or compliance on IDS.

Future research may also investigate how defenders react to different inaccuracies based upon different combinations of miss and false alarm rates. It would also be important to evaluate how the IBL model accounts for human decisions in scenarios where the IDS alerts could be verified with additional information checking, and how would defenders identify inaccuracies that are very rare (Sawyer & Hancock, 2018; Sawyer et al., 2015). One could evaluate the IDS model in environments where the IDS's accuracy dynamically varies between 50% accuracy and 100% accuracy. We are currently addressing some of these and other related ideas in our research program on behavioral cybersecurity.

APPENDIX

Nash Calculations

Dutt et al. (2016) generated Nash equilibria by the Gambit software (McKelvey et al., 2006). It is clear from the game in Figure 2 that there exists no Nash equilibrium in pure strategies (in each of the four possible outcomes, one player is better off deviating). As a result, the only equilibrium solution in this game is in mixed strategies (i.e., selecting each action with some probability), which specifies the following: the hacker attacks with $0.2(\frac{1}{5})$ probability, while the analyst defends with $0.66(\frac{2}{3})$ probability.

Next, Dutt et al. (2016) extend the definition of the above security game by introducing an IDS that can alert the analyst regarding the decision made by the hacker (thus, the analyst does not see the actions of the hacker directly; rather, she gets messages from the IDS based upon the hacker's decisions). The hacker first makes a choice, followed by the IDS that reports the existence/absence of an attack to the analyst. In the security game, we define *pa* as the probability of the IDS to accurately predict the hacker's choice (a wrong prediction therefore occurs with probability *1-pa*). The report from the IDS is determined through probability *pa* (e.g., if the hacker attacks by choosing *a*, IDS reports an attack with probability *pa* and nonattack with probability *1-pa*). After receiving the IDS recommendation, the analyst makes a choice.

Figure A1 lists the Nash equilibria in the cybersecurity game described above for *pa* = 10% (IDS is 10% accurate). The extensive form of the cybersecurity game in Figure A1 along with the Nash equilibria were generated by the Gambit software. As shown in Figure A1, first the hacker makes a choice to attack or not-attack the network, the IDS alerts the analyst by an "attack" or "not-attack" message and finally the analyst makes a choice.

Let p(a) be the probability (proportion) of attack actions; p(na) be the probability of not-attack actions; p(d) be the probability of defend actions; and p(nd) be the probability of not-defend actions. As shown in Figure A1, when the IDS is 10% accurate, the probability of attack is (0.027 ~ 03%).

Overall, the values obtained from Gambit in Figure A1 are:

p(a) = .03; p(na) = .97; p("a"|a)= .10; p("a"|na) = .90; p(d|"a") = 0; and, p(nd|"a") = 1, where, p("a"|a) is the probability of IDS to say "attack" given that the hacker attacks; p("a"|na) is the probability of IDS to say "attack" given

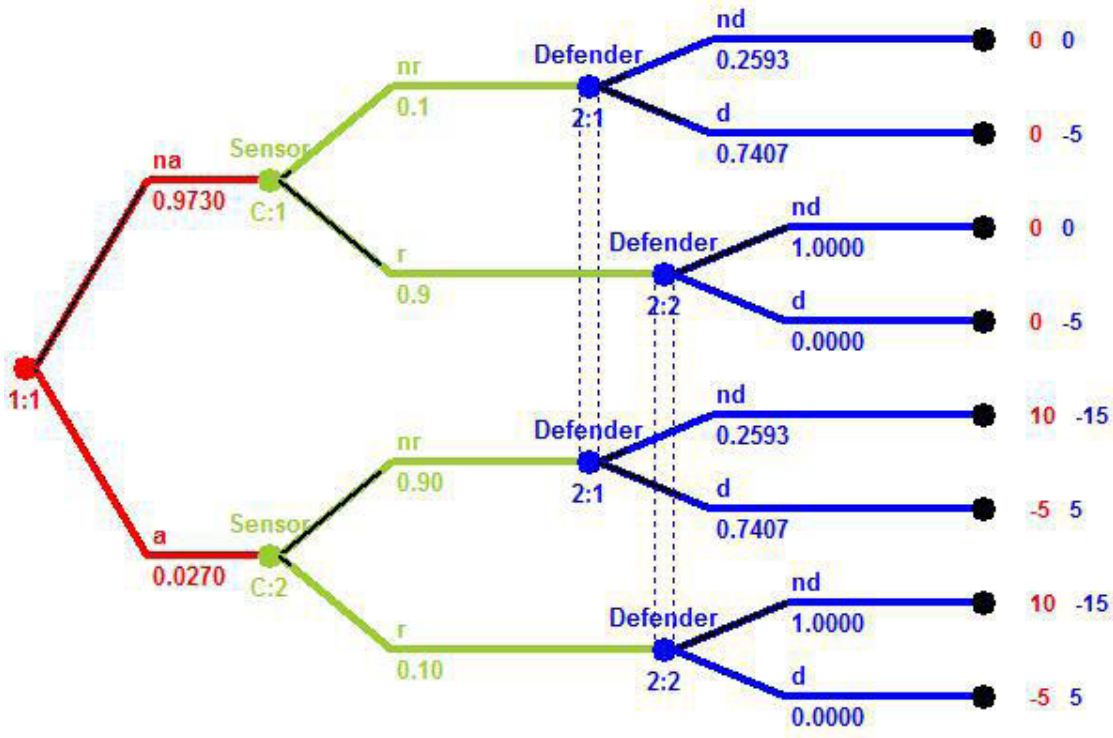

Figure A1. Cybersecurity game tree generated with gambit detailing nash equilibria when IDS is 10% accurate.

that the hacker does not attack; p(d|"a") is the probability of the analyst to defend the network given that the IDS say "attack"; and, p(d|"a") is the probability of the analyst to not defend the network given that the IDS say "attack."

Now, we apply the Bayes' rule to p("a"), that is, the probability of the IDS to say "attack" as per the following:

$$p(\text{"a"}) = p(\text{"a"}|\text{a}) * p(\text{a}) + p(\text{"a"}|\text{na}) * p(\text{na}) \quad (1)$$

Using the values of p("a"|a), p(a), p("a"|na), and p(na) from Figure A1 in (1), we get:

$$p(\text{"a"}) = .87 \text{ and } p(\text{"na"}) = .12 \quad (2)$$

Now, we apply the Bayes' rule to p(d), that is, the probability of the analyst to defend as per the following:

$$p(\text{d}) = p(\text{d}|\text{"a"}) * p(\text{"a"}) + p(\text{d}|\text{"na"}) * p(\text{"na"}) \quad (3)$$

Using the values of p("a") and p("na") from (2) in (3), we get:

$$p(\text{d}) = .09 \text{ and } p(\text{nd}) = .81 \quad (4)$$

Thus, the Nash proportion of attack and defend actions equaled to 3% and 9%, respectively, when the IDS was 10% accurate.

Using the same derivation, the Nash proportion of attack and defend actions equaled to 20% and 67%, respectively, when the IDS was 50% accurate. Also, the Nash proportion of attack and defend actions equaled to 3% and 9%, respectively, when the IDS was 90% accurate.

ACKNOWLEDGMENTS

We are grateful to Indian Institute of Technology Mandi and the Dynamic Decision Making Laboratory at Carnegie Mellon University for providing resources for this

project. Palvi Aggarwal was supported by Visvesverya Ph.D. Scheme for Electronics and IT (IITM/DeitY-MLA/ASO/77), Department of Electronics and Information Technology, Ministry of Communication & IT, Government of India. Cleotilde Gonzalez and Frederic Moisan were supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. Varun Dutt was supported by the Department of Science and Technology, Government of India award (Award number: IITM/DST-ICPS/VD/251).

KEY POINTS

- Two cognitive models, that is, NIDS and IDS model based on IBLT, were developed to explain the attacker's and the defender's behavior in IDS absence and IDS presence with different accuracies.
- Model defenders also showed higher proportion of defend actions when IDS alerts were either absent or 50% accurate compare to when IDS alerts were either 10% or 90% accurate.
- Model attackers showed similar proportion of attack actions across all IDS availability and accuracy conditions.
- Attackers exhibit high reliance on recent information to pay attention to defenders' actions.
- Defenders did not rely on recent information when IDS alerts were absent.
- Defenders' reliance on recent information was a function of accuracy of IDS alerts. The reliance on recent alerts increases as the alerts became more accurate.

ORCID iD

Palvi Aggarwal https://orcid.org/0000-0003-2488-8959

REFERENCES

Aggarwal, P., Gonzalez, C., & Dutt, V. (2017, June). Modeling the effects of amount and timing of deception in simulated network scenarios. In *2017 International conference on cyber situational awareness, data analytics and assessment (Cyber SA* (pp. 1–7). IEEE.

Aggarwal, P., Moisan, F., Gonzalez, C., & Dutt, V. (2018). Understanding cyber situational awareness in a cyber security game involving recommendations. *International Journal on Cyber Situational Awareness*, *3*, 11–38.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Erlbaum.

Bhatt, C., Koshti, A., Agrawal, H., Malek, Z., & Trivedi, B. (2011). Architecture for intrusion detection system with fault tolerance using mobile agent. *International Journal of Network Security & Its Applications*, *3*, 167–175. https://doi.org/10.5121/ijnsa.2011.3513

Bliss, J. P., Bowens, L., Krefting, R., Byler, A., & Gibson, A. (2002). Collective mistrust of alarms. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 46, pp. 1584–1588). SAGE Publications. https://doi.org/10.1177/154193120204601712

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the Compliance-Reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*, 333–345. https://doi.org/10.1177/0018720816682648

Dutt, V., Ahn, Y. S., & Gonzalez, C. (2013). Cyber situation awareness: Modeling detection of cyber attacks with instance-based learning theory. *Human Factors*, *55*, 605–618. https://doi.org/10.1177/0018720812464045

Dutt, V., Moisan, F., & Gonzalez, C. (2016). Role of intrusion-detection systems in cyber-attack detection. In *Advances in Human Factors in Cybersecurity* (pp. 97–109). Springer International Publishing.

Finomore, V., Sitz, A., Blair, E., Rahill, K., Champion, M., Funke, G., Mancuso, V., & Knott, B. (2013). Effects of cyber disruption in a distributed team decision making task. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 57, pp. 394–398). SAGE Publications. https://doi.org/10.1177/1541931213571085

Gonzalez, C., Ben-Asher, N., Oltramari, A., & Lebiere, C. (2014). Cognition and Technology. In C. Kott, A. Wang, & R. Erbacher (Eds.), *Cyber defense and situational awareness*. Springer International Publishing Switzerland.

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

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*, 591–635. https://doi.org/10.1207/s15516709cog2704_2

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

Konak, A., Kulturel-Konak, S., Norman, B. A., & Smith, A. E. (2006). A new mixed integer programming formulation for facility layout design using flexible bays. *Operations Research Letters*, *34*, 660–672. https://doi.org/10.1016/j.orl.2005.09.009

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*, 143–153. https://doi.org/10.1002/bdm.722

Mancuso, V., Funke, G. J., Finomore, V., & Knott, B. A. (2013). Exploring the effects of "Low and Slow" cyber attacks on team decision making. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 57, pp. 389–393). SAGE Publications.

McKelvey, R. D., McLennan, A. M., & Turocy, & T. L. (2006). Gambit: Software tools for game theory.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*, 196–204. https://doi.org/10.1518/hfes.46.2.196.37335

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*, 840–849. https://doi.org/10.1177/0018720813512865

Mukherjee, B., Heberlein, L. T., & Levitt, K. N. (1994). Network intrusion detection. *IEEE Network*, *8*, 26–41. https://doi.org/10.1109/65.283931

Roy, S., Ellis, C., Shiva, S., Dasgupta, D., Shandilya, V., & Wu, Q. (2010). A survey of game theory as applied to network security. In *43rd Hawaii international conference on system sciences* (pp. 1–10). IEEE.

Sawyer, B. D., Finomore, V. S., Funke, G. J., Matthews, G., Mancuso, V., Funke, M., & Hancock, P. A. (2015). Cyber vigilance. *American Intelligence Journal*, *32*, 151–159.

Sawyer, B. D., & Hancock, P. A. (2018). Hacking the human: The prevalence paradox in cybersecurity. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*, 597–609. https://doi.org/10.1177/0018720818780472

Sharma, N., & Dutt, V. (2017). Modeling choice variation in search strategies with multi-armed Bandit Problems. In *2017 International conference on machine learning and data science (MLDS)* (pp. 91–97). IEEE.

Wiczorek, R., Manzey, D., & Zirk, A. (2014). Benefits of decision-support by likelihood versus binary alarm systems: Does the number of stages make a difference? In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 58, pp. 380–384). SAGE Publications. https://doi.org/10.1177/1541931214581078

Wiczorek, R., & Meyer, J. (2016). Asymmetric effects of false positive and false negative indications on the verification of alerts in different risk conditions. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 289–292). SAGE Publications. https://doi.org/10.1177/1541931213601066

Palvi Aggarwal is a postdoctoral researcher at Carnegie Mellon University, USA. She completed her PhD from Indian Institute of Technology Mandi, India. Her current research areas include human factors in cyber security, cognitive modeling, and machine learning.

Frederic Moisan received his PhD in logic and experimental economics at the University of Toulouse, jointly working with the Institut de Recherche en Informatique de Toulouse (IRIT) and Toulouse School of Economics (TSE). Currently, he is a postdoctoral research associate at the Faculty of Economics, University of Cambridge.

Cleotilde Gonzalez is a research professor and director of the Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University. She is associate editor of the *Cognitive Science Journal* and part of the editorial board of multiple journals including the *Journal of Experimental Psychology*: *General*, *Human Factors Journal*, *Journal of Cognitive Engineering and Decision Making*, *System Dynamics Review*, *Decision*, and *Journal of Behavioral Decision Making*.

Varun Dutt is an associate professor and principal investigator at the Applied Cognitive Science Laboratory, School of Computing and Electrical Engineering, Indian Institute of Technology Mandi. His current research interests include cyber security, cognitive science, computational cognitive modeling, judgment and decision making, and artificial intelligence.

Date received: June 14, 2017
Date accepted: June 21, 2020