Adaptive Cyber Deception: Cognitively Informed Signaling for Cyber Defense

Edward A. Cranford
Carnegie Mellon University
cranford@cmu.edu

Palvi Aggarwal
Carnegie Mellon University
palvia@andrew.cmu.edu

Cleotilde Gonzalez
Carnegie Mellon University
coty@cmu.edu

Sarah Cooney
University of Southern California
cooneys@usc.edu

Milind Tambe
Harvard University
milind_tambe@harvard.edu

Christian Lebiere
Carnegie Mellon University
cl@cmu.edu

Abstract

This paper improves upon recent game-theoretic deceptive signaling schemes for cyber defense using the insights emerging from a cognitive model of human cognition. One particular defense allocation algorithm that uses a deceptive signaling scheme is the peSSE (Xu et al., 2015). However, this static signaling scheme optimizes the rate of deception for perfectly rational adversaries and is not personalized to individuals. Here we advance this research by developing a dynamic and personalized signaling scheme using cognitive modeling. A cognitive model based on a theory of experiential-choice (Instance-Based Learning Theory; IBLT), implemented in a cognitive architecture (Adaptive Control of Thought – Rational; ACT-R), and validated using human experimentation with deceptive signals informs the development of a cognitive signaling scheme. The predictions of the cognitive model show that the proposed solution increases the compliance to deceptive signals beyond the peSSE. These predictions were verified in human experiments, and the results shed additional light on human reactions towards adaptive deceptive signals.

1. Introduction

In cybersecurity, static defense strategies (e.g., intrusion detection, firewalls, anti-malware, or anti-virus) are effective front-line defenses that prevent many attacks. Despite their effectiveness, many attacks still succeed as adversaries continuously adapt to find and exploit new vulnerabilities. It is imperative to develop security defenses that thwart attacks before they occur and that adapt to ever-evolving adversaries. One way to actively prevent attacks is to employ signaling schemes based on game-theoretic algorithms.

Security analysists can actively monitor a network for fraudulent activity. However, resources are often limited, and a network cannot be fully monitored all the time, therefore signaling can aid in protecting unprotected resources. Signaling is a defense method whereby information is sent to an attacker that reveals the protection status of a potential target. Truthful signals can deter some attacks, but employing deceptive tactics can increase the perceived coverage of unprotected targets by finding the correct balance between truthful and deceptive signals [1].

Deception is a form of persuasion where one intentionally misleads an agent into a false belief, in order to gain an advantage over the agent and achieve one's goals [2]. Deception is often used for ill-gains, for example, in spear-phishing attacks or disinformation campaigns. However, it can also be used for good, to mitigate unwanted behavior or illegal activity, much like signage in a front lawn may deter would-be thieves even if no physical security system truly exists. In cybersecurity, deceptive signals can be used to deter attacks on uncovered systems beyond any capabilities of static defenses that do not use signaling or only use truthful signals.

Finding the right balance of deceptive signaling so that the attacker continues to believe the signal is crucial to the success of the strategy. Recently, game-theoretic research on deceptive signaling algorithms in Stackelberg Security Games (SSGs) has optimized the strategic allocation of limited defenses and the rate of deception so that a rational attacker would not attack when presented a signal [3]. However, this research optimized signaling for perfectly rational adversaries, and humans exhibit, at best, bounded rationality [4].

Deception is a tool used to trick the human mind, and as such, a better understanding of how would-be attackers react to and learn from deceptive tactics is important for developing effective cyber defenses. To these ends, we examined human behavior in a cyber-security game called the Insider Attack Game (IAG) that pits humans, who play the role of an inside-attacker, against cybersecurity analysts controlled by an algorithm. Results of laboratory experiments, and a cognitive model that accurately predicts human

URI: https://hdl.handle.net/10125/63971
978-0-9981331-3-3

performance in the IAG, show that humans behave far differently than predicted under assumptions of perfect rationality [1][5-7]. Humans exhibit nominally irrational behaviors that reflect capacity and information limitations, and the need to resort to heuristic strategies, that result in cognitive biases (e.g., confirmation bias). While signaling algorithms optimized for perfectly rational adversaries do improve defense compared to not signaling at all [3], they are less than effective against boundedly rational humans.

One reason for the algorithms' shortcomings is that they are static and not personalized to individuals. While humans are not perfectly rational, they learn quickly and can adjust behavior in real time. A signaling scheme that is adaptive to the individual can potentially outperform traditional signaling schemes. Based on our understanding of human behavior response to deceptive signaling in the IAG, through experimentation and cognitive modeling, we propose a signaling scheme that is adaptive to an individual's experience. The signaling scheme is designed to both exploit and maintain the attacker's belief in the signal.

In what follows, we first describe a signaling scheme that is optimized for and effective against perfectly rational adversaries, and how an approach based on cognitive modeling would differ. Next, we describe an online game that was developed to investigate human behavior response to deceptive signaling. Results from humans playing the game, and a cognitive model that accurately predicts their performance, provide key insights that lead to the design of a signaling scheme that is grounded in principles of human cognition. The scheme is predicted, via the cognitive model, to be more effective against boundedly rational humans than traditional schemes. The results of a laboratory experiment show that the signaling scheme is effective at increasing compliance with the signal compared to traditional schemes, but humans still attack more often than predicted by the model. The human behavior results are compared with those of the cognitive model to shed light on human response to deceptive signals, guide avenues of future research, and aid in the development of more effective signaling schemes.

2. Deceptive signaling for cybersecurity

In cybersecurity, deception has been adopted across many security techniques with much success, for example, in the strategic allocation of honeypots [8] and masking the properties of systems [9]. Using deceptive signals in Stackelberg Security Games also has great potential for use in cybersecurity.

SSGs model the interaction between an attacker and a defender using a game-theoretic framework. In the SSG, a defender plays a particular strategy (i.e., random patrolling of an airport terminal), the attacker observes the strategy, and then the attacker takes action. Under this framework, researchers have developed algorithms, such as the Strong Stackelberg Equilibrium (SSE), that optimally allocates limited defense resources across a set of targets [10]. These algorithms have been applied successfully across a number of physical security systems (e.g., protecting ports, scheduling air marshals, and mitigating poachers) [10-13]. Such security practices could be applied to the cyber realm, for example, in scheduling active monitoring of security systems by network administrators (e.g., security analysts).

Xu and colleagues [3] extended the SSG models by incorporating elements of signaling, in which a defender (sender) strategically reveals information about their strategy to the attacker (receiver) in order to influence the attacker's decision making [14-15]. Sending a message that reveals the protection status of target can influence attacker behavior. For example, a truthful message that reveals a target is monitored can deter attacks, but adversaries can attack with impunity when a message reveals the target is not monitored. However, defenders can use a combination of truthful and deceptive signals to help deter attacks on the unprotected resources. Xu et al.'s [3] solution, the Strong Stackelberg Equilibrium with Persuasion (peSSE), improves defense against a perfectly rational attacker compared to strategies that do not use signaling. For a given target, the peSSE finds the optimal combination of bluffing (sending a deceptive message that the target is monitored when it is not) and truth-telling (sending a truthful message that the target is covered) so that a rational attacker would not attack in the presence of a signal.

In practice, the SSE allocates defenses proportionally across the set of targets so that the expected values of all targets are equal. Once defenses are scheduled, the attacker can choose a target to attack. Then, as determined by the peSSE, the defender will send a signal to the attacker revealing the protection status of the target, which may sometimes be deceptive. Based on this information, the attacker can then choose to continue the attack or withdraw. If the attacker continues the attack, then they will receive a penalty if the target is truly monitored, but a reward if the target is open. The peSSE sends deceptive signals at a rate that makes the expected value of attacking a target, given a signal, equal to the expected value of withdrawing the attack, or zero. Therefore, under the assumption of perfect rationality, when presented with a signal an attacker will always break ties in favor of the defender and choose the safer option, to withdraw the attack.

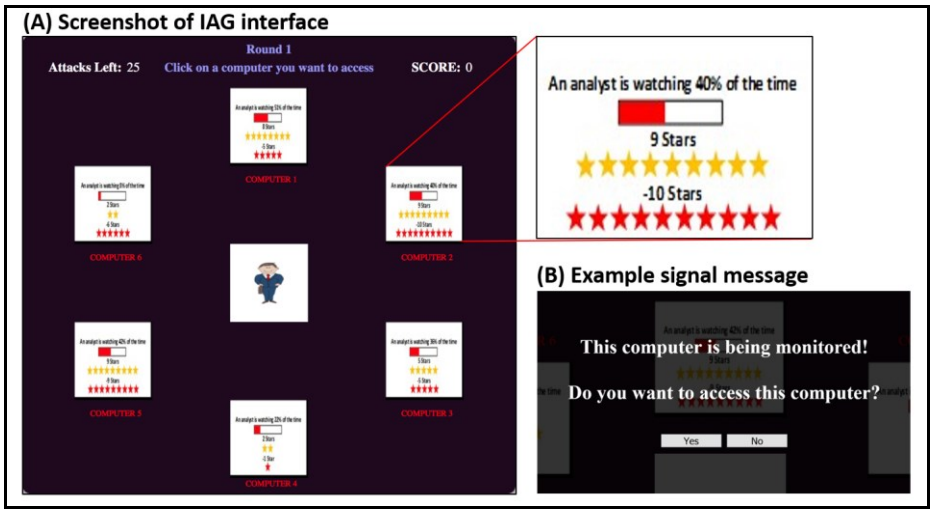

Figure 1. Screenshot of the IAG (A) and an example signal message (B).

Unfortunately, the peSSE is less than optimal against humans that do not always make the rational best decision [6-7]. The signaling scheme can be improved by taking into account cognitive dynamics and biases that influence decisions making. What is needed for a cognitive-based signaling scheme is a computational or mathematical model of human reactions in all circumstances. The adaptive scheme can then use that model to dynamically optimize signal presentation given previous events and responses.

3. Insider Attack Game

The Insider Attack Game (IAG) is an online game designed to investigate the interaction between an attacker and defender in a cybersecurity scenario, gain a better understanding of how humans react to deceptive signals, and assess the effectiveness of various signaling schemes [5][7]. Figure 1(A) shows a screenshot of the game interface. Players take the role of the attacker at the center of the screen (i.e., a company employee) and their goal is to score points by "hacking" computers to steal proprietary data. There are six computers from which to choose to attack, but only two security analysts (i.e., defenders controlled by a computer algorithm) that can monitor one computer each. If the player attacks a computer that is monitored, they lose points denoted by the number of red stars, but if the computer is not monitored then they win points denoted by the number of yellow stars. Each computer shows its reward for winning, penalty for losing, and the probability that the computer is being monitored (which reflects the SSE for the game).

Players make repeated attempts at attacking computers. On each turn, the player must first select a computer to attack. Then, the signaling algorithm determines whether to send a truthful signal or a deceptive signal. In the IAG with six targets and two analysts, the peSSE presents a signal every time a target is monitored, or 33% of trials on average. Additionally, the peSSE sends a signal half of the time when a target is not monitored, or 33% of trials on average. This means that, on average, a signal is deceptive half of the time. At this rate, the expected value of attacking given a signal is zero, the same expected value as withdrawing the attack. Therefore, a perfectly rational adversary that only attacks with a positive expected value (i.e., in the absence of a signal), is predicted to attack on 33% of trials on average (i.e., when a signal is *not* presented).

Figure 1(B) shows an example message signaling that a target is currently being monitored. If the computer is not being monitored, then the first line of the message is omitted. After reading the message, the player must decide whether to continue their attack or withdraw and earn zero points. Players play four rounds of 25 trials each (after an initial five trials of practice). The payoff structures and monitoring probabilities of the targets are different in each round. Coverage and signaling of targets were precomputed for each trial. Therefore, each individual player experiences the same coverage and signaling schedule.

3.1. Understanding human behavior in the IAG

Cranford et al. [6-7] presented the results of 100 human participants playing the IAG against the peSSE signaling scheme and a cognitive model of an attacker that accurately predicts human performance and helps explain human behavior.

Figure 2 shows the mean probability of attack across trials. The dashed line at the bottom of the graph shows the predicted probability of attack of a perfectly

rational adversary (33%). The results showed that humans attacked far more often than predicted, almost 80% of trials. Figure 3 displays the probability of attack on trials when a signal is presented, showing that humans attack more than 70% of trials while a perfectly rational adversary would never attack.

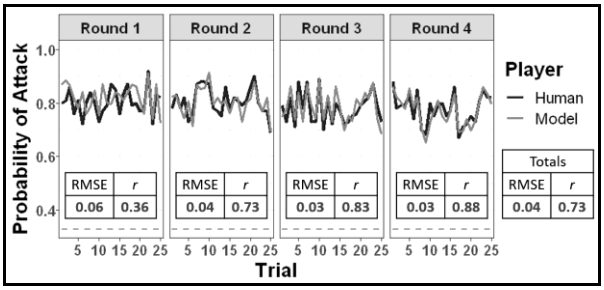

Figure 2. Mean probability of attack across trials and rounds in the IAG for humans compared to the model, playing against the peSSE.

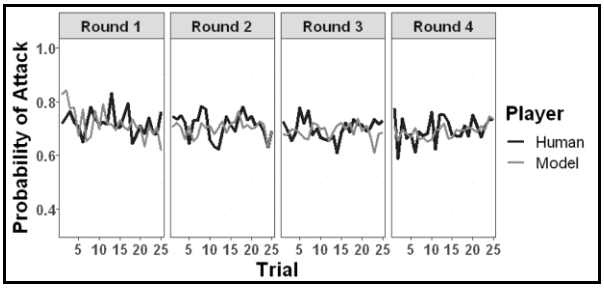

Figure 3. Mean probability of attack when a signal is present, comparing humans and model playing the IAG against the peSSE.

It is clear that humans do not make perfectly rational decisions. Instead, human behavior can be explained as decisions from experience [16]. To better understand the cognitive process underlying human decision making, a cognitive model was built in the ACT-R cognitive architecture [17-18] and decisions are made following instance-based learning theory (IBLT) [16]. According to IBLT, decisions are made by generalizing across past experiences, or instances, that are similar to the current situation. For the IAG, instances are represented by the features of the decision. This includes the context of the selected target, the decision, and the outcome. The context includes the monitoring probability [0.0, 1.0], reward [1, 10], and penalty values [-1, -10] associated with the selected target, and whether a warning signal was presented [present, absent]. The possible decisions are attack or withdraw, and the outcome is the reward or penalty based on the decision. In a given situation, for each possible decision, an associated utility is computed through blended memory retrieval weighted by contextual similarity to past instances. The decision with the highest expected utility is made. However, withdrawing always results in zero points. Therefore, the model only needs to determine the utility of attacking in order to make a choice.

In ACT-R, the retrieval of past instances is based on the activation strength of the relevant instance in memory and its similarity to the current context. The activation of an instance reflects the power law of practice and forgetting, and includes a partial matching process reflecting the similarity between the current context elements and the corresponding context elements for the instance in memory. A variance parameter *s* introduces stochasticity in retrieval. Similarities between numeric slot values are computed on a linear scale from 0.0, an exact match, to -1.0. Symbolic values are either an exact match or maximally different, -2.5, to prevent bleeding between memories for different actions and signal types.

A Boltzmann softmax equation determines the probability of retrieving an instance based on its activation strength. The IBL model uses ACT-R's blending mechanism [16][19] to calculate an expected outcome of attacking a target based on a consensus of past instances. The expected outcome is the value that best satisfies the constraints of all matching instances weighted by their probability of retrieval.

In summary, the outcomes of past instances are weighted by their recency, frequency, and similarity to the current instance to produce an expected outcome. If the value is greater than zero then the model attacks, else it withdraws.

For each trial, the model first selects a target with the highest expected outcome, generated via blending, and then decides whether to continue the attack or withdraw based on whether a signal was presented. For this decision, the model uses blending to generate an expected outcome for the given target, but only on the basis of the signal and ignores the values of the target context (i.e., the target information is occluded from the participants, so it is plausible that they do not consider the target information beyond deciding which target to select initially). An instance is then saved in memory that represents the model's expected outcome. Humans tend to remember not only the actual experience, but also their expectations prior to the experience [20]. This results in additional positive (or negative) instances, which in turn generates a confirmation bias whereby one's pre-conception of winning (or losing) perpetuates itself in future trials, even when it is actually disconfirmed. Based on the value of the expected outcome, a decision is made, and the action and outcome slots of the current instance are updated to reflect the action taken by the model and the ground-truth outcome. This final instance is saved in memory and thereby influences future decisions.

The model continues for four rounds of 25 trials each. The model behavior reflects its experiences. If an action results in a positive/negative outcome, then its future expectations will be increased/decreased, and the model will be more/less likely to select and attack that target in the future. Also, the impact of a particular past experience on future decisions strengthens with frequency and weakens with time.

The model was run 1000 times to simulate a population of individuals and to generate stable estimates of human performance. As shown in Figures 2 and 3, the model is highly accurate at predicting human performance (total RMSE = 0.04), even matching the trial-to-trial variations that reflect the underlying coverage and signaling schedules (total r = 0.73), and that accuracy increases over time. Not only does the model match the average human performance in the IAG, but it also matches well to the individual performance. Figure 4 shows the distribution of participants by their mean probability of attack. Like humans, some model simulations attack at a fairly low rate, while a large proportion attack 95% of the time or more. Figure 5 shows the distribution for when a signal present, and indicates that some participants comply with the signal, to a degree, while most do not.

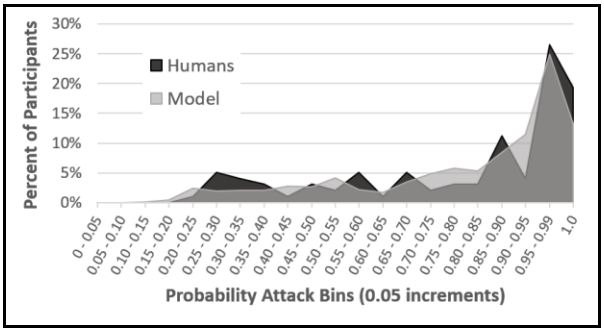

Figure 4. Distribution of participants by probability of attack for humans compared to the model playing the IAG against the peSSE.

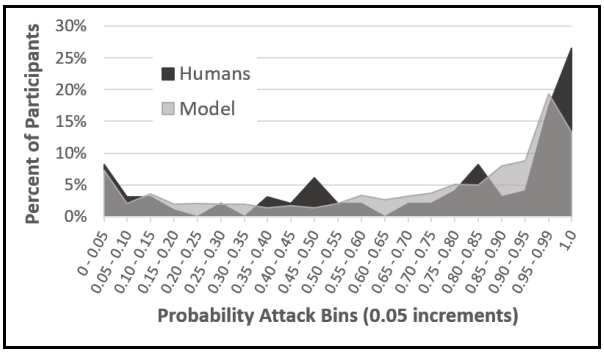

Figure 5. Distribution of participants by probability of attack when a signal is present, comparing humans and model playing the IAG against the peSSE.

Human decision making in the IAG is largely influenced by memory dynamics across past experiences. The peSSE suffers because human biases (e.g., recency, frequency, and confirmation) lead to overweighting of certain outcomes that, often, results in inflated expectations. Humans fail to fully comply with the signal because they are more likely to expect a positive outcome than a negative one as belief in the signal deteriorates. While deception is an effective tool for preventing malicious behaviors, the experience of successfully calling a bluff can reduce compliance with the signal. Regaining trust in the signal is difficult if not impossible to do under static signaling schemes. Therefore, an adaptive signaling scheme is needed that adjusts the rate of deception to dynamically balance (re)building trust in the signal and exploiting it, and thus optimizes compliance.

4. Cognitive signaling scheme for adaptive cyber defense

Individual attackers behave differently from one another, and each may learn and adjust behavior after repeated experience with deceptive signals. Therefore, an adaptive signaling scheme based on cognitive principles can be used to adjust the rate of deception, tailored to an individual's behavior, so as to maintain belief in the signal. Our initial solution towards this problem is to interleave blocks of trials with only truthful signals between blocks of trials with deceptive signals. The assumption is that experiences of rewards when a signal is present increases the probability of attacking in the future, while experiences of penalties given a signal reduces the probability of attacking in the future. Therefore, eliminating deceptive signals for a short period of time can help increase penalties and restore belief in the signal. The goal for the cognitive signaling scheme is to induce, and preserve, the belief that attacking given a signal will result in a loss.

Relying on the attacker's history of behavior, this new cognitive signaling scheme estimates the current probability of attack given a signal and judges whether the cost of issuing a truthful block outweighs the benefits of a deceptive block, to effectively reduce the future probability of attack given a signal. At the beginning of each block of trials, a closed form equation of the current probability of attack given a signal, reflecting the blending process used in generating expectations and the recency and frequency power laws in chunk activations, can be formulated based on the times t since past actual decisions made by the attacker, as:

$$P_{est}^{now}(A|S) = \frac{\sum_i^{wins} t_i^{-d} + \sum_j^{losses} t_j^{-d}}{\sum_i^{wins} t_i^{-d} + \sum_j^{losses} t_j^{-d} + \sum_k^{draws} t_k^{-d}} \quad (1)$$

Next, we estimate the change in probability of attack given a signal from a truthful block. Therefore, we need to make an additional assumption as to how wins and losses impact choice. We assume that the attacker will follow the same decision-making process, keeping the same format reflecting probability matching behavior:

$$P_{ass}^{now}(A|S) = \frac{\sum_i^{wins} t_i^{-d}}{\sum_i^{wins} t_i^{-d} + \sum_j^{losses} t_j^{-d}} \quad (2)$$

The impact of a truthful block of size b on $P_{ass}^{now}(A|S)$ results in a new estimate $P_{ass}^{then}(A|S)$ with an expected number $1/3 * b * P_{est}^{now}(A|S)$ of losses distributed randomly across the block, where 1/3 is the mean probability of sending a signal in a truthful block. For the present implementation, the block size b is set to 10. This value was chosen as a reasonable compromise that provides enough opportunities for switching blocks while allowing for enough experience within a block to impact behavior.

The adaptive cognitive signaling scheme is as follows: the next block will use a truthful signal if the following comparison of the cost in terms of additional attacks allowed in the next block is less than its benefits (i.e., the number of attacks saved in the remaining r trials during the rest of the experiment after that block):

$$\frac{1}{3} * b * [1 - P_{est}^{now}(A|S)] < \alpha * r * [P_{ass}^{now}(A|S) - P_{ass}^{then}(A|S)] \quad (3)$$

Where 1/3 is the difference in probability of a signal being generated between deceptive (66%) and truthful blocks (33%), and α is a discount parameter that can take any value between 0.0 and 1.0 (default is 1/3). The discount parameter is an assumption of how long the impact of the truthful block on the probability of attack given a signal will persist. If we assume that it will persist until the end and all future blocks will be deceptive blocks, then the right value would be 2/3 (i.e., the percentage of trials when a signal is generated). If it would persist indefinitely but all future blocks are truthful blocks, then that value would be 1/3. In practice, it will be somewhere between 1/3 and 2/3 depending of the mix of truthful and deceptive. The effect of the signal will dilute over time, so the minimum 1/3 is a reasonable default value.

In summary, the cognitive signaling scheme uses a closed form version of the model decision procedure to optimize the tradeoff between the cost of building trust in the signal using blocks of truthful signals, and the benefits of exploiting that trust in future blocks of deceptive signals.

4.1. Cognitive model predictions and human performance against cognitive signaling

The effectiveness of the cognitive signaling scheme was examined through cognitive model simulations and a human behavioral experiment. The cognitive model of the attacker presented above was run through 1000 simulations against the cognitive signaling scheme, and these predictions were then compared to performance of human participants. For the human experiment, 100 participants were recruited via Amazon Mechanical Turk. All participants resided in the United States. For completing the experiment and submitting a completion code, participants were paid $1 plus $0.01 per point earned in the game, up to a maximum of $5.50. One participant was removed from analysis because of incomplete data due to data recording errors, resulting in a final N of 99. For brevity, details of the experimental design can be found in Cranford et al. [7].

As an initial study, all players began with a block of truthful signals to establish baseline belief in the signal. As before, players played four rounds of trials each, with a different set of targets each round. Every 10 trials overall the algorithm determined whether to switch to a different type of block: either using only truthful signals or using deception according to the peSSE. Figure 6 shows the proportion of players that received a truthful block, across each of the 10 blocks in the game. The first block is always a truthful block. From there, depending on the individual's behavior, the cognitive signaling scheme assigns more truthful or deceptive blocks. The second block is always deceptive, and the third block is about evenly divided between truthful and deceptive. Over time, the proportion of truthful blocks declines because the estimated reduction in future probability of attack over the remaining blocks does not outweigh the near-term term costs of the truthful block. Overall, the probability of assigning truthful blocks is higher for humans than for the model, suggesting that humans are less trusting in the signal and more willing to attack.

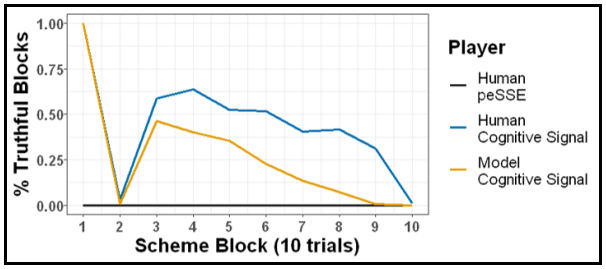

Figure 6. Proportion of truthful blocks assigned by the cognitive signaling scheme per block of 10 trials for humans compared to the model.

To assess human and model performance, the data was analyzed for the probability of attack across trials. The probability of attack was calculated as the proportion of players that continued the attack on a given trial. Figure 7 shows the probability of attack across trials for humans compared to the model when playing against the cognitive signaling scheme, which is compared to human performance when playing against the peSSE. Compared to the peSSE, the cognitive signaling scheme further reduces the probability of attack, but at the expense of giving up more attacks in the first block. Because all signals are truthful in the first block of the cognitive signaling condition, fewer signals are sent to deter attacks overall. The effect of an initial truthful block is immediately observable by a relatively lower probability of attack in trials 10 through 20 (which is always a deceptive block), and this trend continues through the game. The effect of the cognitive signaling scheme is more prominent in the model. As can be seen, humans attack more often than predicted by the model. Because humans tend to attack more than the model, the cognitive signaling scheme also presents more truthful blocks to humans (see Figure 6 above).

To assess the effectiveness of the signaling scheme, we examine defender utility. The defender is penalized one point every time the player attacks a target that is not monitored, and zero points otherwise (e.g., if a player attacks a target that is monitored, or does not attack). This means, the more often players attack in the face of a deceptive signal, the worse will be defender utility. Since targets are not monitored 66% of trials on average, a defender utility less than -17 (i.e., >2/3 of 25 trials) means the signaling scheme is better than a purely truthful signaling scheme, while a utility greater than -9 is ideal (i.e., <1/3 of 25 trials). While the cognitive signaling scheme reduces attacks, as displayed in Figure 8, defender utility is only marginally improved compared to the peSSE and much lower compared to model predictions. Compared to the model and the peSSE, more truthful signals were given to humans overall under cognitive signaling. This resulted in fewer signals sent to deter attacks on uncovered targets and consequently more free passes to attack with impunity, even though overall compliance with the signal is increased.

At first glance, these results indicate that the cognitive signaling scheme is not as effective as predicted. However, a closer inspection of the results revealed that the scheme is effective at influencing human behavior beyond the peSSE, for some humans. As shown in the histogram in Figure 9, the model fails to account for approximately 44% of participants that attacked at a rate of 95% or more. However, as shown in Figure 10, if we separate participants into two

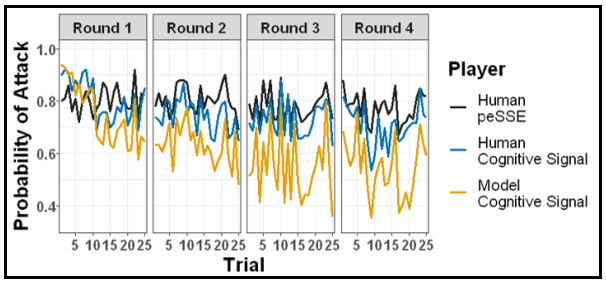

Figure 7. Mean probability of attack across trials and rounds in the IAG for humans and the model playing against the cognitive signaling scheme, which are compared to humans playing against the peSSE.

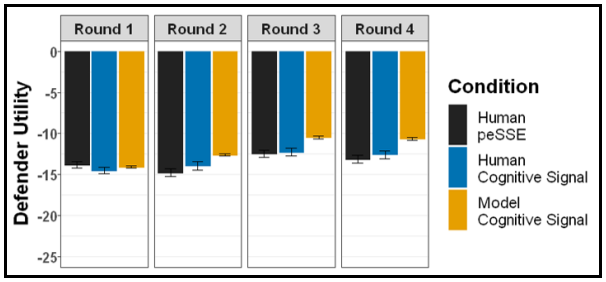

Figure 8. Defender utility for the cognitive signaling scheme compared to the peSSE.

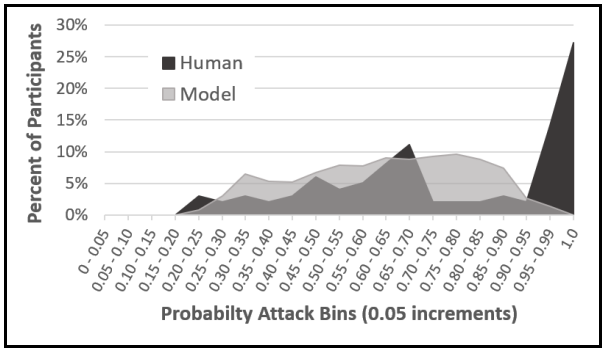

Figure 9. Distribution of participants by probability of attack for humans compared to the model playing the IAG against the cognitive signaling scheme.

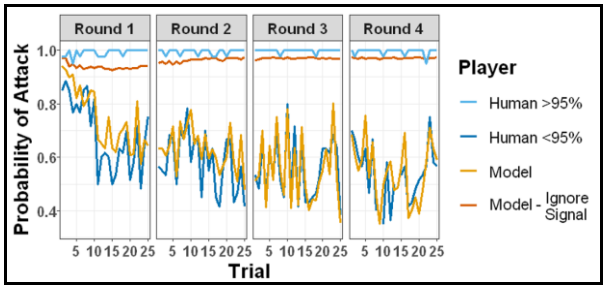

Figure 10. Mean probability of attack across trials and rounds in the IAG for humans compared to the model playing against the cognitive signaling scheme, separating humans that attack >=95% and <95%.

groups, the model is highly accurate at predicting performance of the approximately 56% of participants that attack at a rate less than 95%.

For the participants that attacked at a rate greater than 95%, the cognitive signaling scheme did not influence behavior even after giving these participants, almost exclusively, truthful blocks. Figure 11 shows the proportion of truthful blocks assigned per block of 10 trials for the two separate groups. The cognitive signaling scheme presented the same proportion of truthful blocks to the model as it did those participants that attacked less than 95% of the time. However, the scheme continued to present truthful blocks to the other group of participants because they continued attacking undeterred in the face of a signal.

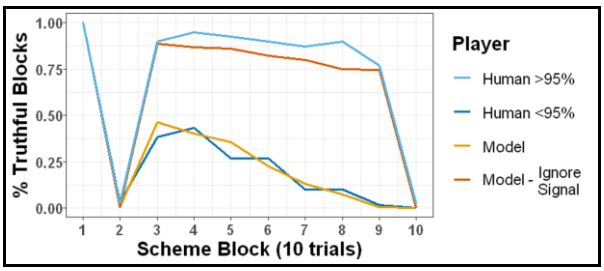

Figure 11. Proportion of truthful blocks assigned by the cognitive signaling scheme per block of 10 trials comparing the model to humans that attack >=95% and to those that attack <95%.

As shown in Figure 12, the cognitive signaling scheme provides better defense for a subset of humans, as indicated by low defender utility values that match what was predicted by the model. However, against some participants the scheme performs about as poorly as would be expected given no signals.

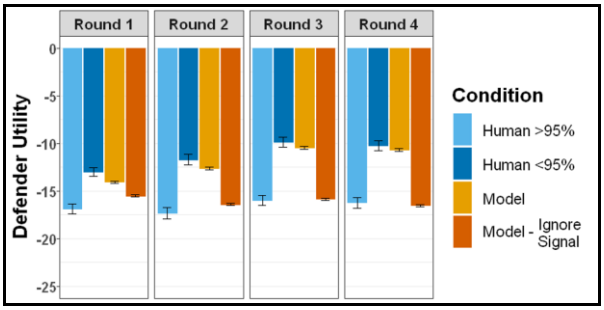

Figure 12. Defender utility for the cognitive signaling scheme, comparing the model to humans that attack >=95% and to those that attack <95%.

In fact, in a post-experiment survey that asked an open-ended question about what strategy participants used when faced with a signal, a majority of participants that attacked more than 95% responded that they ignored the signal. An informal analysis was conducted with two independent coders, and the responses were categorized based on the features in which decisions were based or the reported actions taken. Discrepancies between coders were resolved through discussion. The results are presented in Figure 13 comparing responses of participants that attacked greater than 95% to those that attacked less than 95%. For the former group, almost 23% reported that they ignored the signal while another ~10% reported that they always attacked. Approximately 10% reported that they stay and continue attacking the same target even after suffering a loss, while about 15% switch to another target and continue attacking. Meanwhile, for the latter group, none reported that they ignore the signal, while approximately 20% reported that they withdraw in the face of a signal, and ~12% withdraw if the monitoring probability was high. Overall, the survey results show that some participants ignore the signal and treat all instances equally. This means that the signaling scheme will not be effective against these participants because the expected value of attacking given a signal is combined with the expected value of attacking given no signal. Therefore, with only 2 analysts, the overall expected values would be positive, resulting in constant attacks.

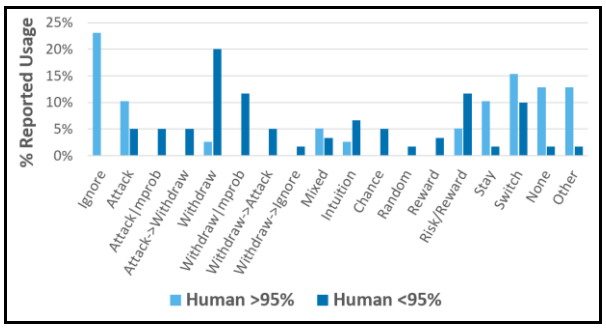

Figure 13. Distribution of reported attack strategies in the IAG for humans that attack >=95% compared to those that attack < 95%.

Based on these findings, we created a version of the cognitive model that does not consider the signal when generating an expected outcome of attacking the selected target. For this version, blending samples equally across past instances regardless of the signal, and so only recency and frequency of past instances play a role in decisions. The model attacks on 96.0% of trials (*SD* = 15.1%), with 54% of simulations attacking 100% of trials and 35.8% attacking greater than 95%, matching well to the distribution shown in Figure 9 of the humans that attack >=95%, and the other measures (see Figures 10-12). These results stress the importance of understanding the features that individuals consider in their decisions, since one's representation of the decision context strongly influences the chosen action.

5. General Discussion

In this paper, we improved upon traditional game-theoretic signaling schemes for cyber defense using a computational model of human cognition. The peSSE signaling scheme offers effective defense against boundedly rational human adversaries compared to not signaling. However, the algorithm optimizes the rate of deception for perfectly rational adversaries, which results in a static scheme that is not personalized to individual attackers. Through experimentation and cognitive modeling, we learned how humans respond to deceptive signals, and developed a cognitive signaling scheme that is adaptive and based on cognitive principles. Cognitive model predictions showed that the solution is promising at further influencing human behavior beyond the capabilities of the peSSE. These predictions were verified in human experiments, and the results helped shed additional light on individual differences in human behavior.

The cognitive model predicts human decisions are made by aggregated retrieval across past experiences based on the similarity to the current situation [16]. These decisions are influenced by frequency and recency of past experiences, cognitive biases, and representation of information in memory. These are the core assumptions for the cognitive signaling scheme.

Two key insights gleaned from the cognitive model regarding human behavior, are that: (1) decisions are highly affected by confirmation bias, and (2) it is important to consider what features the individual factors in their decision. The cognitive signaling scheme leveraged this information to induce bias and influence human behavior. Specifically, by relying on observations of actual human behavior, the cognitive signaling scheme estimated the probability of attack given a signal and, if it was too high, would send only truthful signals for a period of time in an attempt to rebuild trust in the signal and ultimately increase compliance. Continued attacks given truthful signals should strengthen the expectation that attacking in the future, given a signal, will result in a loss.

An open question for the cognitive signaling scheme is how long do we need to display truthful signals to regain trust, and thus compliance? Currently, the approach gives up some attacks early on with an initial truthful block, but this is done in order to increase belief in the signal for the rest of the experiment. The algorithm only determines whether to switch to a different type of signal after a block of 10 trials. Ten is a reasonable value, but the algorithm could be called as often as every trial. The implications of this are unclear at this point. It could result in too few truthful signals in a row to impact behavior, or it could help further personalize the scheme so that it is better adapted to the individual. Future research is aimed at exploring ways to optimize the proportion of truthful to deceptive signals over a period of time.

Cranford et al. [7] showed that humans seem to ignore the context of the selected target, and only consider the signal when making decisions of whether to continue to attack. This insight allowed us to simplify the cognitive signaling scheme and focus on reducing the overall probability of attack given a signal, and not need to take into account individual target values. Afterall, the SSE normalizes targets, so their expected values are equal [10].

An important observation from the human experiments was that the cognitive signaling scheme is only effective for some participants, while others seem to ignore the signal when making decisions. This further highlights the importance of accurately representing decision features. For participants that do not consider the signal, all targets are treated equally. Thus, trying to reduce the probability of attack given a signal by adjusting the rate of deception may prove fruitless when the overall expected values are positive for all targets. An alternative, method to combat such adversaries could be to shift coverage instead of, or in addition to, adjusting the rate of deception. For example, while it might be difficult or impossible to extract attack preferences to influence behavior, it might be possible to extract selection preferences and shift coverage to induce more experiences of loss given a signal. Driving the expected value of attacking to negative values could result in attackers starting to pay attention to the signal, which in turn would raise the effectiveness of cognitive signaling. Future research is aimed at exploring the potential of this method.

Another limitation of the current approach is that it relies only on deceiving when given a signal. Meanwhile, players can attack with impunity when no signal is presented. An alternative approach is to use deception two ways, when a signal is present and when it is absent. In this way, the attacker can lose points when a signal is absent, instilling further uncertainty in their decisions. In fact, recent research explored several game-theoretic algorithms that employ two-way deception that proved better than one-way deception against human participants [1]. Future research is aimed at exploring the potential of using two-way deception in the current cognitive signaling approach.

We have already used two-way deception in an alternative cognitive signal scheme, but it has not been tested against human participants [6]. In that scheme, the cognitive model is used to trace human behavior in real time to make predictions about the human's probability of attack given a signal, and determines on a trial-to-trial basis whether to give a signal based on the underlying coverage. The scheme shows potential,

and the use of two-way signaling is an enhancement over the current approach. Where the current approach stands out is in the fact that it is a closed-form solution that relies on a simplified version of the cognitive model to make predictions of individual behavior. However, there is room to refine the current cognitive signaling approach through the discount parameter, the size of the truthful block, and the assumptions concerning the likelihood of various coverage conditions. Future research will further explore the complexities of the cognitive signaling scheme.

One caveat to these approaches is that they rely on observing and tracking an individual's behavior. In the real world, it may prove difficult if not impossible to track all, or even some, of an adversary's actions. Luckily the methods are robust and can be tailored to a population, sub-group, or even a time-window of attacks. While not as effective as at the individual level, such a method could still reliably influence human behavior.

In conclusion, we have outlined an initial approach to deceptive signaling for cyber defense that relies on cognitive models of attacker behavior to balance the rate of deception in an attempt to keep belief in the signal high. The cognitive signaling scheme is adaptive and personalized, and can therefore be used to induce biases and influence attackers to comply with the signal beyond the capabilities of any static scheme. Future research is aimed at improving upon the current cognitive signaling scheme.

6. Acknowledgments

This research was sponsored by the Army Research Office and accomplished under MURI Grant Number W911NF-17-1-0370.

7. References

[1] Cooney, S., Wang, K., Bondi, E., Nguyen, T., Vayanos, P., Winetrobe, H., Cranford, E. A., Gonzalez, C., Lebiere, C., and Tambe, M. "Learning to Signal in the Goldilocks Zone: Improving Adversary Compliance in Security Games", In Proceedings of the ECML and PKDD, Wurzburg, Germany, 2019, 923.

[2] Rowe, N. C., and Rrushi, J., Introduction to Cyberdeception, Springer, Switzerland, 2016.

[3] Xu, H., Rabinovich, Z., Dughmi, S., and Tambe, M., "Exploring Information Asymmetry in Two-Stage Security Games", In Proceedings of the 29th AAAI CAI, Austin, TX, 2015, pp. 1057-1063.

[4] Simon, H. A., "Rational Choice and the Structure of the Environment", Psychological Review 63(2), APA, Washington, DC, 1956, pp. 129-138.

[5] Cranford, E. A., Lebiere, C., Gonzalez, C., Cooney, S., Vayanos, P., and Tambe, M., "Learning About Cyber Deception Through Simulations: Predictions of Human Decision Making with Deceptive Signals in Stackelberg Security Games", In Proceedings of the 40th Annual Conference of the CSS, Madison, WI, 2018, pp.258-263.

[6] Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., and Lebiere, C., "Towards personalized deceptive signaling for cyber defense using cognitive models", In Proceedings of the 17th Annual Meeting of the ICCM, Montreal, CA, 2019, 56.

[7] Cranford, E. A., Gonzalez, C., Cooney, S., Tambe, M., and Lebiere, C., "Constructing Deceptive Signaling Strategies for Cyber Defense Using Cognitive Models of Attacker Behavior", Cognitive Science, (under review).

[8] Kiekintveld, C., Lisy, V., and Pibil, R., "Game-theoretic foundations for the strategic use of honeypots in network security", In Cyber Warfare, Springer, Cham, 2015, pp. 81–101.

[9] Schlenker, A., Thakoor, O., Xu, H., Fang, F., Tambe, M., Tran-Thanh, L., Vayanos, P., and Vorobeychik, Y., "Deceiving Cyber Adversaries: A Game Theoretic Approach", In Proceedings of the 17th AAMAS, IFAAMAS, Stockholm, Sweden, 2018, pp. 892–900.

[10] Tambe, M., Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned, Cambridge University Press. Cambridge, UK, 2011.

[11] Pita, J., Jain, M., Ordónez, F., Portway, C., Tambe, M., Western, C., and Kraus, S., "ARMOR Security for Los Angeles International Airport", In Proceedings of the 23rd AAAI CAI, Menlo Park, CA, 2008, pp. 1884-1885.

[12] Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., and Meyer, G., "Protect: A Deployed Game Theoretic System to Protect the Ports of the United States", In Proceedings of the 11th AAMAS, IFAAMAS, Valencia, Spain, 2012, pp. 13-20.

[13] Sinha, A., Fang, F., An, B., Kiekintveld, C., & Tambe, M., "Stackelberg Security Games: Looking Beyond a Decade of Success", Proceedings of the 27th IJCAI, 2018, pp. 5494–5501.

[14] Battigalli, P., "Rationalization in signaling games: Theory and applications", International Game Theory Review 8(01), World Scientific, 2006, pp. 67–93.

[15] Cho, I.-K., and Kreps, D. M., "Signaling Games and Stable Equilibria", The Quarterly Journal of Economics, Oxford University Press, 1987, 102(2), 179–221.

[16] Gonzalez, C., Lerch, J. F., and Lebiere, C., "Instance Based Learning in Dynamic Decision Making", Cognitive Science 27(4), 2003, pp. 591-635.

[17] Anderson, J. R., and Lebiere, C., The Atomic Components of Thought, Erlbaum, Mahwah, NJ, 1998.

[18] Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y., "An Integrated Theory of the Mind", Psychological Review 111(4), APA, Washington, DC, 2004, pp. 1036-1060.

[19] Lebiere, C., "Blending: An ACT-R Mechanism for Aggregate Retrievals", In Proceedings of the 6th ACT-R Workshop. George Mason University, Fairfax, Va. 1999.

[20] Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., and Anderson, J. R., "A Functional Model of Sensemaking in a Neurocognitive Architecture", Computational Intelligence and Neuroscience, Hindawi Publishing Corp., London, UK, 2013, pp. 1-29.