Accounting for Uncertainty in Deceptive Signaling for Cybersecurity

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

Han-Ching Ou
Harvard University
hou@g.harvard.edu

Cleotilde Gonzalez
Carnegie Mellon University
coty@cmu.edu

Milind Tambe
Harvard University
milind_tambe@harvard.edu

Christian Lebiere
Carnegie Mellon University
cl@cmu.edu

Abstract

Deceptive signaling has proven an effective method that can aid security analysists and deter attacks on unprotected targets by strategically revealing information to an attacker. However, recent research has shown that uncertainty in real-time information processing can have a negative impact on the effectiveness of the defense algorithm. The current research developed a new algorithm, dubbed Confusion Signaling, that aims to account for uncertainty in an abstracted insider attack scenario. The results of cognitive model simulations and a human behavioral experiment reveal interesting and unexpected reactions under uncertainty. We discuss the implications of these findings for signaling algorithms that aim to account for uncertainty using deceptive signaling for cybersecurity.

Keywords: deceptive signaling, uncertainty, insider attack, cognitive model, instance-based learning

1. Introduction

To protect a network from insider threats, a limited number of security analysts must monitor a larger number of potential targets. Therefore, to expand the perceived coverage of analysts, deceptive signaling for cybersecurity has been formally proven to be an effective strategy for mitigating attacks on unprotected targets against perfectly rational adversaries (Xu, Rabinovich, Dughmi, & Tambe, 2015). In addition, Cranford et al. (2021) showed that, while the strategy is less effective against boundedly rational human adversaries, it still proves better than not using deceptive signals at all. Signaling is a technique whereby information is revealed to an attacker regarding the protection status of a potential target, and the key is finding the correct balance between truthful and deceptive signals to maintain belief in the signal (Cooney et al, 2019). Recently, Bondi et al. (2020) showed that an important factor that affects this balance and influences attacker behavior is uncertainty in real-time information processing. For example, ignoring uncertainty in the adversaries' observation of a signal (i.e., missing a signal or perceiving a signal when there isn't one) can lead to unexpected adversary reactions that have a negative impact on the effectiveness of the signaling scheme. Therefore, the current research addresses these issues discovered by Bondi et al. to account for uncertainty when designing deceptive signaling schemes for cybersecurity.

Bondi et al. (2020) examined the effectiveness of deceptive signaling in a physical security setting, wildlife conservation for mitigating poaching. A security team (i.e., park rangers) patrols a set of geographic locations (i.e., targets), and drones are used to send signals to would-be attackers (i.e., poachers). The drones send a signal by flashing lights to indicate a ranger is nearby. If a poacher observes the signal, they are expected to flee the area or else be caught. The drones sometimes send deceptive signals indicating a ranger is nearby when in fact they are not. The goal for the security algorithm is to determine how often they can send deceptive signals while maintaining their effectiveness. Game-theoretic models are used to optimize the rate of signaling.

The scenario is modeled as a two-stage Stackelberg Security Game (SSG; Tambe, 2011; Xu et al, 2015) in which the first stage allocates defenders to targets and a second stage optimizes the rate of deceptive signaling. After the defender deploys a fixed strategy and allocates defenders to targets, the attacker may then observe the strategy and select a target to attack. The defender then detects intrusions and sends a signal or not, and the attacker reacts by continuing the attack or withdrawing. Theoretically, the algorithm (the Strong Stackelberg Equilibrium with Persuasion, peSSE; Xu et al., 2015) is predicted to reduce the defender's expected losses. However, when deployed in a real-world situation, the effectiveness of the algorithm is limited because it does not account for uncertainties in real-time information.

URI: https://hdl.handle.net/10125/102738
978-0-9981331-6-4

Bondi et al. (2020) presented an algorithm, Games with Uncertainty And Response to Detection with Signaling Solver (GUARDSS), that accounts for two types of uncertainty observed in the poaching domain: 1) defender uncertainty in detecting the attacker (henceforth *detection uncertainty*), and 2) attacker uncertainty in observing the signal (henceforth *signal uncertainty*). In a simulation experiment, Bondi et al. showed that the original peSSE algorithm results in significant decline in defender expected utility (i.e., the amount of loss incurred by the defender as a function of the number of attacks on uncovered targets) as the amount of uncertainty increases. Meanwhile, the GUARDSS algorithm maintains the effectiveness of deceptive signaling as uncertainty increases, with minimal decline in defender expected utility.

The present research aims to apply the insights and lessons learned from Bondi et al. (2020) to the domain of cybersecurity. Cranford et al. (2021), developed an abstraction of a cybersecurity task called the Insider Attack Game (IAG) to investigate attacker decision-making in an insider-attack scenario when faced with deceptive signals. In that study, human participants played the role of attackers, and their task was to first choose one of six targets to attack. They were then sent a message that either claims a target is monitored (signal) or is not monitored (no signal), after which the attacker must decide to continue their attack or withdraw. When the message claims a target is monitored, this signal is sometimes deceptive. The attackers made repeated decisions, and the system predetermines which targets to defend and which targets to send signals for each trial as determined by the peSSE. The results showed that the peSSE was more effective at mitigating attacks and reducing defender losses compared to never signaling or only truthfully signaling. However, the study was not designed to account for uncertainty in information processing because the presentation of signals was clear, and the environment controlled. It is likely that uncertainty would exist in a real-world situation. Therefore, in the present research we aim to investigate the impact of uncertainty on the peSSE, and to design a new algorithm that effectively accounts for uncertainty.

There are key differences between the poaching scenario and the insider attack scenario that influence the design of the new algorithm. In the IAG, there is no sensor that detects attacks, network security analysts have no constraints on which target they can travel to at each time point, and sending signals does not require a drone be present at a target. Instead, at each time point, we envision the defender allocating security analysts to monitor a set of targets and each target is assigned a signal or not. When an attacker selects a target to attack, if the signal is present, it can be observed (e.g., an icon in the taskbar could be red to indicate the computer is actively being monitored or green to indicate that it is not monitored). Given this type of scenario, the signaling algorithm does not need to detect attackers to be deployed. Therefore, for the present research we focus solely on accounting for signal uncertainty. We must note however that, for signaling algorithms that are adaptive and personalized to individuals, such as that proposed by Cranford et al. (2020a; 2020b), it is important to know which targets were attacked when an analyst is not present to detect them, and to account for detection uncertainty, to maintain accurate models and predictions of attacker behavior that can be used to adapt the signal and/or coverage appropriately.

The GUARDSS approach uses a branch and price algorithm to solve an exponentially large linear program. However, the approach is a bit overkill for our current problem because, in the IAG, there is no reliance on drones for detecting attacks or sending signals, constraints on movement of analysts is not limited by physical distances, and there is no subsequent reallocation of defenses to sensor locations. Because the structure of the IAG is substantially different from the poaching domain, our approach to accounting for signal uncertainty in the IAG is to remain closer to the original peSSE. As described in more detail later, we add a layer of confusion to the signaling state via a confusion matrix (i.e., whether a signal was perceived or not if presented or not) and likewise adjust the equation for determining signaling probabilities to account for said confusion. Hence, the algorithm is dubbed Confusion Signaling.

We conducted a human behavioral experiment to examine the effectiveness of the Confusion Signaling algorithm. Our results reveal surprising and interesting human responses to uncertainty. Particularly, that humans seem to mull on unexpected events brought on via signal uncertainty, but which impacts behavior in a different direction than predicted (Erev et al., 2010). These results led us to revise our cognitive model to make more accurate predictions, as will be discussed.

In what follows, we first describe the IAG, the peSSE, and the new Confusion Signaling algorithm that accounts for uncertainty in the IAG. We then present a cognitive model of human attack decisions in the IAG and model predictions of the impact of uncertainty on the effectiveness of deceptive signaling algorithms that do not account for signal uncertainty (i.e., the peSSE) and one that does (i.e., Confusion Signaling). We then present the results of a human-behavioral experiment that was designed to validate the model predictions, followed by new model predictions given revisions to the cognitive model. Finally, we discuss the results and their implications for signaling algorithms, including future research that aims to account for uncertainty in personalized, adaptive signaling algorithms.

2. Deceptive signaling for cybersecurity

For the present research, we designed a signaling algorithm that accounts for uncertainty in the IAG. We first describe the IAG and the original peSSE algorithm before describing the Confusion Signaling algorithm.

2.1. Insider Attack Game

The IAG is an abstraction of a cybersecurity task, designed to investigate the influence of deceptive signals on cyber-attacker decision making in an insider attack scenario (for brevity we summarize the procedure here, while complete details can be found in Cranford et al., 2021). Players take the role of an insider attacker and make repeated decisions of "hacking" computers to steal proprietary data. Attackers are presented six targets, of which two of the targets are monitored, or covered, by security analysts (i.e., simulated defenders controlled by a defense algorithm). Each target displays the reward earned if they successfully attack a target, the penalty imposed if the attack is unsuccessful, and the probability that the target is being monitored by an analyst. This information represents the full information of the payoff structure that could be acquired by an attacker by probing the network prior to an attack.

On each turn, attackers must first decide which of six targets to attack, with the goal of avoiding the two analysts. After selecting a target, the attacker is presented a message stating whether the computer is being monitored (i.e., signal) or not (i.e., no signal). In the instructions, the attacker is informed that the message is always truthful when claiming a target is not being monitored, but it is sometimes deceptive when claiming the target is being monitored. The attacker must then decide to either continue the attack or withdraw, after which they are provided feedback based on their decision. If they continue the attack, they are given a reward or penalty, depending on the true underlying coverage. If they withdraw then they are awarded zero points. Attackers are incentivized to earn as many points as possible across four rounds of 25 trials each, and a new set of targets are presented at the start of each round. Attackers play a practice round of five trials before proceeding with the main trials to become familiar with the interface and sequence of decisions.

2.2. Defense algorithm for deceptive signaling

The IAG is modeled as a Stackelberg Security Game (SSG; Tambe, 2011). SSGs model the interaction between an attacker and a defender under a game-theoretic framework. According to the SSG, a defender first plays a particular strategy (e.g., random patrolling of a set of targets) that is then observed by the attacker. The attacker then takes action by deciding which target to attack. Algorithms such as the Strong Stackelberg Equilibrium (e.g., Tambe, 2011) have been designed to optimize the allocation of limited defense resources and we have adapted it to the cybersecurity domain to assist network security analysts in defending a typically large number of targets given limited defense resources.

Xu et al. (2015) extended the SSG by incorporating elements of signaling, whereby defenders strategically reveal information about their strategy to the attacker to influence the attacker's decision making. For example, by using a combination of truthful and deceptive messages that claim a target is monitored, the defender can increase the perceived coverage of the targets. Xu et al.'s peSSE defense algorithm improves defense by finding 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 as to whether the target is monitored) so that a rational attacker would not attack in the presence of a signal.

In practice, the peSSE first allocates defenses proportionally across the set of targets so that the expected values of all targets are equal. The attacker then chooses a target to attack. A message is then sent to the attacker revealing the protection status of the target. When the message claims the target is not monitored, this is referred to as no signal, and is always truthful. When the message claims the target is monitored, this is referred to as the signal, which is sometimes deceptive. If signals were always truthful, the expected value of attacking a target given a signal would be negative and so a perfectly rational adversary should withdraw. However, due to limited defense resources, there would be many targets that are not signaled in which the attacker may attack with impunity. A signal is always sent when the target is truly monitored (i.e., $\mathbb{P}(signal|covered) = 1$), and the peSSE determines the optimal rate of sending deceptive signals, $\mathbb{P}(signal|uncovered)$, such that the expected value of attacking a target, given a signal, is equal to the expected value of withdrawing, or zero. Therefore, a perfectly rational adversary would continue to withdraw in the presence of a signal while the defender increases the perceived coverage of targets. The goal of the peSSE is to minimize the defender's loss, or defender expected utility (DEU), where the defender receives -1 point for every successful attack (i.e., attacks on uncovered targets), and is operationalized as:

$$DEU = -1 * \#\ attacks\ on\ uncovered\ targets \quad (1)$$

For the IAG, the coverage and signaling of targets is precomputed for each trial, and therefore each attacker experiences the same schedule of coverage and signaling, which controls for differences in behavior

that could be attributed to differences in coverage and signaling schedules.

2.3. Accounting for uncertainty in the IAG via Confusion Signaling

To account for uncertainty in the IAG, we add a confusion matrix to the algorithm that maintains the equilibrium for deceptive signaling so that the expected value of attacking given a signal remains at 0. As shown in Figure 1, the top portion depicts the peSSE without uncertainty, showing the average probability that a target is covered, in blue, and the average probability that a target is signaled given coverage or not, in red. The peSSE aims to maintain the equilibrium:

$$\mathbb{P}(covered|signal)U_{ac} = -\mathbb{P}(uncovered|signal)U_{au} \quad (2)$$

where U_{ac} is the utility of attacking a covered target (i.e., the penalty) and U_{au} is the utility of attacking an uncovered target (i.e., the reward). If reward and penalty are equal, then the corresponding signaling probabilities will be equal. More generally, the probability of signaling when the target is uncovered is:

$$\mathbb{P}(signal|uncovered) = -\frac{pU_{ac}}{(1-p)U_{au}} \quad (3)$$

where p is the coverage probability of the target.

As shown in the bottom portion of Figure 1, given signal uncertainty, a portion of the signals are not observed (i.e., in gray; *kappa*, κ), and a portion of the time when no signals are sent, they are falsely perceived (i.e., in green; *lambda*, λ). To maintain the equilibrium of Equation (2), and to account for signal perception, we must adjust the rate of signaling when a target is not covered and compute the new probability to send a signal given an uncovered target, $\mathbb{P}'(signal|uncovered)$. Therefore, we must consider the additional confusion matrix of uncertainty, shown in Table 1, when adapting the signaling algorithm.

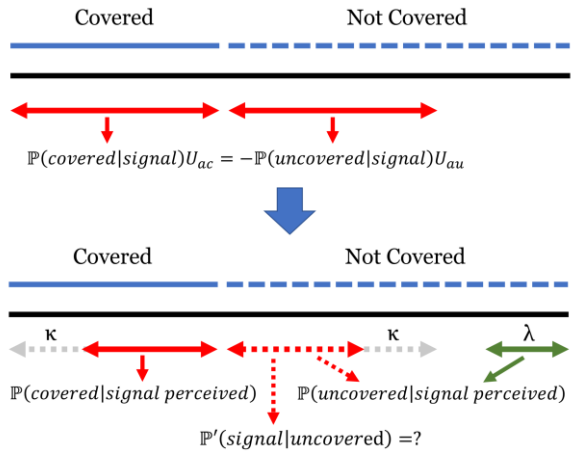

Figure 1. Depiction of signal uncertainties in the IAG.

Table 1. Confusion matrix for signal uncertainty

		Perceived
True	***no signal***	***signal***
no signal	$\mathbf{1-\lambda}$	$\boldsymbol{\lambda}$
signal	$\boldsymbol{\kappa}$	$\mathbf{1-\kappa}$

Given the confusion matrix in Table 1, we wish to maintain the equilibrium of Equation (2):

$$\mathbb{P}(covered|signal\ perceived)U_{ac} = -\mathbb{P}(uncovered|signal\ perceived)U_{au} \quad (4)$$

Therefore, we continue to always signal when a target is covered and keep the $\mathbb{P}'(signal|covered) = 1$, and solve for $\mathbb{P}'(signal|uncovered)$ by using Bayes' rule:

$$\mathbb{P}'(signal|uncovered) = max\left(0, min\left(1, -\frac{p(1-\kappa)U_{ac} + (1-p)\lambda U_{au}}{(1-p)(1-\kappa-\lambda)U_{au}}\right)\right) \quad (5)$$

where p is the probability a target is covered.

The new signaling equation is not without its limitations. When $\kappa + \lambda \geq 1$, the scheme will default to always signaling (i.e., the bottom right of the diagonal in the plots in Figure 2, described below). Additionally, when $\frac{\lambda}{1-\kappa} \geq -\frac{pU_{ac}}{(1-p)U_{au}}$, the scheme will default to never signaling. This creates a "safe zone" for using Confusion Signaling such that within the safe zone, the utility of attacking will remain constant and maintain the equilibrium while defender losses are expected to decrease as the number of false negatives increase.

3. Modeling uncertainty in the IAG

Cranford et al. (2021) created a cognitive model of attacker decision-making in the IAG to better understand how humans react and adapt to deceptive signals as assigned by the peSSE defense algorithm. For the present research we used the same model to make predictions against the Confusion Signaling algorithm.

3.1. Cognitive model of attackers in the IAG

We briefly summarize the cognitive model, while details can be found in Cranford et al. (2021). The cognitive model was developed in the ACT-R cognitive architecture (Anderson & Lebiere, 1998; Anderson et al., 2004) and decisions are made according to Instance-Based Learning Theory (IBLT; Gonzalez, 2013; Gonzalez, Lerch, & Lebiere, 2003). According to IBLT, decisions in dynamic environments are made by generalizing across past experiences, learned through feedback from repeated interactions. Experiences are encoded as instances in declarative memory, and represented by the contextual features of the situation, the decision/action made, and the outcome/utility of the

decision. Decisions are made by generating an expected utility of each possible action and selecting the action that maximizes that utility. The generation of expected utilities is made through a memory retrieval process called blending (Lebiere, 1999), which aggregates across past outcomes based on the contextual similarity of past instances, and their recency and frequency in memory, and is computed by the following equation:

$$V = \underset{V_t}{argmin} \sum_{i=1}^{n} P_i \times Sim(V_t, v_{it})^2 \quad (6)$$

The value, V, is the aggregate value based on matching chunks i, weighted by their retrieval probability P_i. The similarity function, $Sim(V_t, v_{it})$, is used to compare the outcomes in memory chunks v_{it} and candidate consensus values V_t, and is effectively the error function to be minimized. The similarity function returns a linearly scaled value, normalized between 0 and -1.0, computed as the absolute difference between V_t and v_{it}. In the simplest case, where the outcomes are numerical and the similarity function is linear, as is the case here, the process simplifies to a weighted average by the probability of retrieval. The retrieval probability is based on the activation strength of instances in memory, which is computed via standard ACT-R equations (see Anderson & Lebiere, 1998; Anderson et al., 2004).

The IBL model plays the IAG similar to humans. The model is run 1000 times to simulate a population of individuals and makes different decisions on each run based on stochasticity in the retrieval processes, and differences in initialization, which lead to unique trajectories of experiences. Each run of the model is initialized with a set of instances that represent knowledge gained from instructions and a simulated practice round. The "instruction knowledge" chunks include two instances that reflect information acquired from reading the instructions about the potential reward they could earn given the truthfulness of the messages: one represents the knowledge that attacking a target given no signal will always result in a reward (warning = absent, outcome = 10), and another represents the knowledge that attacking a target given a signal will only sometimes result in a reward (warning = present, outcome = 5). Five additional instances are added to represent experience during a practice round. Each instance represents an attack on a target during a practice trial and is selected uniformly at random. These initialized experiences, combined with stochasticity in retrieval, drives individual differences between runs.

On each trial the model first decides which target to select. The features include the reward [1, 10], penalty [-1, -10], and monitoring probability [0.0, 1.0]. The model generates an expected outcome for each target, via blending, and selects the target with the highest value. Next, the model generates a new expected value of attacking given only the signal feature [present, absent], and a straightforward decision rule is applied: if the value is greater than zero then the model attacks, else it withdraws. The model then receives feedback about its decision: zero if the decision was to withdraw, or the reward or penalty if the decision was to attack, given whether the target was uncovered or covered. Two instances are saved to memory each trial: one represents the expectation generated when deciding to attack or withdraw and the other represents the ground truth decision and feedback received. Storing expectations drives a confirmation bias in which the availability of additional positive instances in memory, created via positive expectations of attacking, perpetuates a bias to attack on future trials even after suffering a loss.

Cranford et al. (2021) showed that the model accurately predicts human performance in the IAG when playing against the peSSE algorithm. Not only does the model accurately predict the mean probability of attacking across trials, and the mean defender expected utility per round, it also accurately reflects the distribution of individual variances in overall attack probabilities. Thus, we use this model to make predictions about the impact of uncertainty in the IAG.

3.2. Expected impact of uncertainty on the effectiveness of deceptive signaling

The IBL cognitive model was used to make predictions regarding the impact of varying levels of both *kappa* and *lambda* signal uncertainty on the effectiveness of the peSSE. We note that amount of uncertainty in a real-world scenario is currently unknown, but we expect relatively small values for each and particularly small values for lambda. However, we make predictions across the full range of values for both the probability of kappa and the probability of lambda, from 0.0 to 1.0, in increments of 0.1. For each pair of values, the model was run 1000 times to simulate a population of individuals. On each trial, the signal/no-signal was misperceived based on the probability of uncertainty given by kappa/lambda. This was also done for initialized practice trials.

The top left panel of Figure 2 shows the overall mean probability of attack as a function of kappa and lambda uncertainties. The results show that as kappa increases, the probability of attack is expected to increase. As lambda increases, the probability of attack is expected to decrease.

The bottom left panel of Figure 2 shows the overall mean defender expected utility as a function of kappa and lambda uncertainties. The results mirror those of the probability of attack. As kappa increases, the defender's expected utility worsens, and as lambda increases, the defender's expected utility improves.

Taken together, we can expect that if we do not account for uncertainty in the IAG, the peSSE will perform increasingly worse as kappa increases, but performs better as lambda increases. However, as previously noted, we expect very little lambda uncertainty in the real world, and potentially greater kappa uncertainty. Kappa uncertainty means that sometimes when a target is covered the attacker will not perceive a signal, and if they attack then they will be caught (κ for covered targets in Figure 1). This should add confusion when no signal is experienced so an attacker should attack less often given no signal. However, because they also experience fewer signals when a target is not covered (κ for uncovered targets in Figure 1), the expected outcomes given no signal should remain positive. Therefore, as kappa increases, the attacker is given more opportunities to attack with impunity, even if they suffer a few losses. Meanwhile, because signals are always sent to covered targets, lambda only increases the probability of observing a signal when a target is uncovered (λ for uncovered targets in Figure 1). The effect is that the expected outcome of attacking given a signal should increase above zero, but as long as the attacker believes the signal, the result is fewer attacks overall.

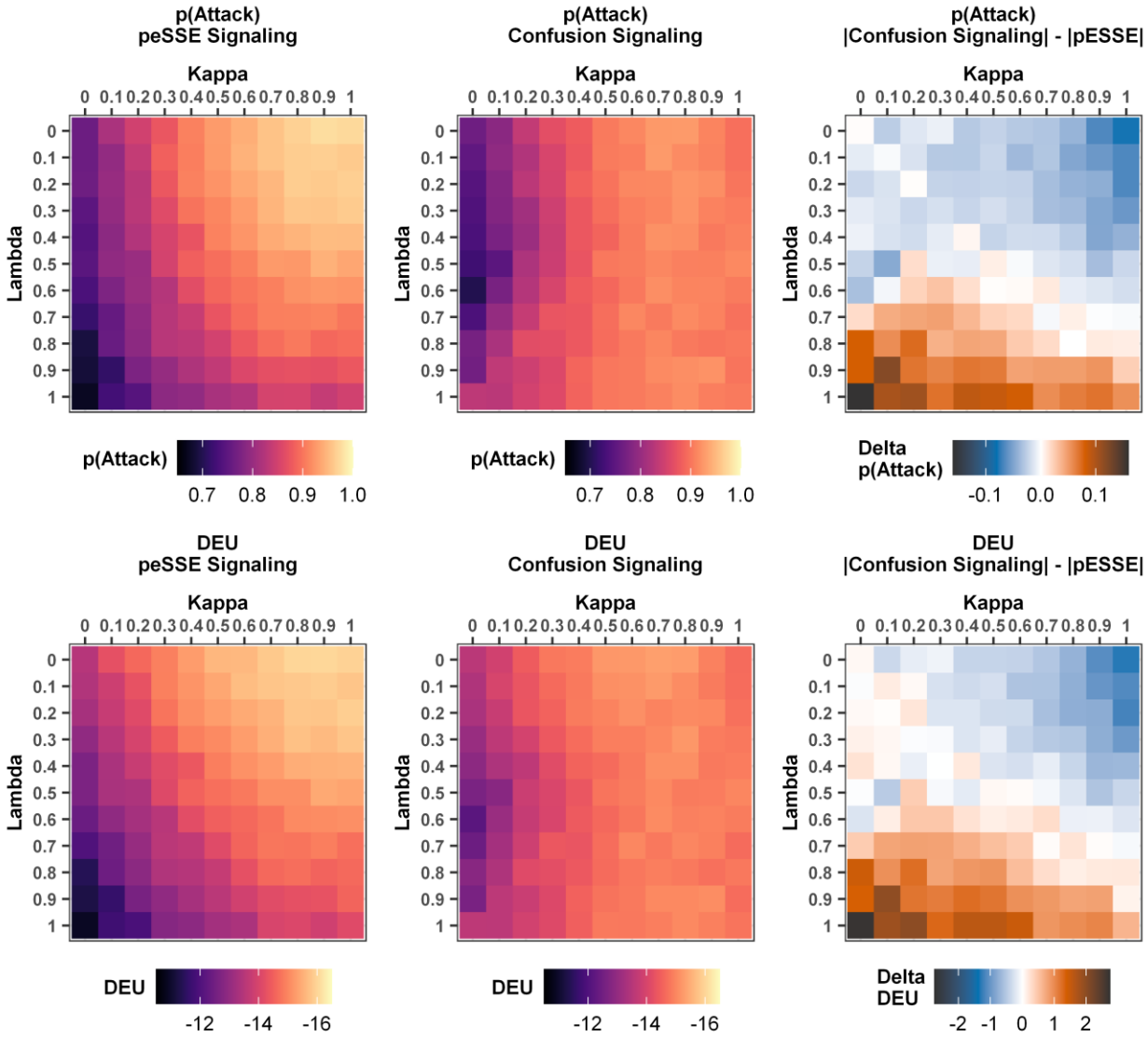

Figure 2. Cognitive model predictions of Confusion Signaling compared to the peSSE across varying levels of signal uncertainty. The bottom panel shows the difference between Confusion Signaling and peSSE.

3.3. Model predictions of Confusion Signaling

We used the cognitive model to make predictions of attacker behavior under the Confusion Signaling algorithm. The top center panel of Figure 2 shows the overall mean probability of attack as a function of kappa and lambda uncertainties, and the bottom center panel shows the overall mean defender expected utility. The panels on the right of Figure 2 show the expected change, delta Δ, from the Confusion Signaling algorithm compared to the peSSE for probability of attack (top-right) and DEU (bottom-right), respectively. Blue represents a lower probability of attack and higher defender expected utility for Confusion Signaling than the peSSE, indicating improved defenses under Confusion Signaling, whereas red indicates worse defense under Confusion Signaling.

Taken together the model predicts that the Confusion Signaling algorithm will largely improve defenses under signal uncertainty, especially as the amount of kappa uncertainty increases. At low levels of kappa, Confusion signaling is likely to reduce the overall number of attacks, while roughly maintaining the defender expected utility of the peSSE. However, Confusion Signaling is really expected to shine when kappa becomes larger than 0.2. In the real world, we can expect some level of kappa and lambda uncertainty, and more kappa than lambda, but we do not expect either value to be very large. Therefore, as a first test to validate the effectiveness of Confusion Signaling, we deployed a human behavioral experiment that simulates the effects of $\kappa = 0.3$ and $\lambda = 0.1$.

5. Validation of Confusion Signaling scheme in a human behavioral experiment

We conducted a human behavioral experiment to examine human decision making in the IAG under the Confusion Signaling algorithm. We recruited 111 participants via Amazon Mechanical Turk (all resided in the United States), of which 10 were removed from analysis due to data recording errors, resulting in a final sample size of 101. For completing the experiment and submitting a completion code, participants were paid a base of $1.50, and earned up to $4.50 in bonus payment at a rate of $0.01 per point accumulated in the game. For brevity, details of the experimental design can be found in Cranford et al. (2021).

The results are compared to human performance against the peSSE, as reported in Cranford et al. (2021), and to cognitive model simulations. The current study used the same targets and coverage schedule that was used in the Cranford et al. study. The only difference is the underlying signaling scheme. The human results are compared to three cognitive simulations. One is a model of the peSSE, another is a model of the peSSE that adds kappa and lambda signaling uncertainty, and the third is a model of the Confusion Signaling. The model was run 1000 times to simulate a population of attackers and to generate stable estimates of performance. The coverage and signaling schemes were the same for each participant or model agent within each experiment or simulation. Since observing the signal in the IAG is straightforward, to simulate kappa and lambda uncertainty in the experiment and simulations, the signals were changed up front according to the probabilities of kappa and lambda (i.e., 30% of the time a signal was scheduled to be sent, it was not sent, and 10% of the time no signal was scheduled, one was sent).

5.1. Refining the cognitive model

As will be described next in the results section, humans did not react to Confusion Signaling as expected. Attackers attacked less often than predicted, and to model this reduction in attacks, we implemented a mulling mechanism that would strengthen the activation of unexpected, anomaly events when users experienced a loss given no signal. The idea is that unexpected events trigger extra thought about them and so the activation trace in memory is stronger and has more of an impact on future retrievals (e.g., see Erev et al., 2010). Due to kappa uncertainty, such mulling would decrease the future probability of attack on un-signaled targets. However, instead we observed that attackers continued to attack un-signaled targets at a high rate, but the probability of attack on signaled targets was reduced. This result is possibly due to the fact that the instructions tell the participant that the message is always truthful when claiming a target is uncovered, which is true of the peSSE given no uncertainty, but when uncertainty exists participants will sometimes experience a loss when they do not perceive a signal. The key insight here is that humans seem to either recognize these anomalous events as something incorrect with their perception or something incorrect with the game, and so instead of encoding the instance as a "no signal" event, they encode it as a "signal" event. Therefore, when mulling, the instance encoded upon feedback and mulled on is encoded as if a signal was present, and thus the future probability of attack given a signal is reduced.

We added two parameters that govern mulling: 1) the number of events to mull on before ceasing (i.e., after a while, the events are no longer unexpected or anomalous, or possibly that attackers become fatigued), and 2) the number of references added on each event (i.e., the number of times the event is rehearsed in memory). We explored several parameter values, and

the best fit was found with 10 mull events and 1 reference added per event. For models with uncertainty, mulling is also added to the initialized practice trials when kappa uncertainty is experienced.

We also revised the model to generalize better across signaling conditions. The following changes are based on other research with the IAG. First, we modified the model initialization. The "instruction knowledge" instances were changed so that when a signal is absent the outcome is 5.33 (the average reward of all targets), and when a signal is present now two instances are encoded, one with a negative outcome of -5.33 and the other with a positive outcome of 5.33 (averaging to 0, or the expected value given a signal). Secondly, in a study on the effects of endowment in the IAG, Cranford et al. (2020c) observed that attackers are not affected very much by initial losses. Therefore, the penalties are reduced to be no greater than the total current points. The effect is an improved fit of the model in the initial few trials. Lastly, Cranford et al. (2020a) observed that some participants attack greater than 95% and these participants also reported that they ignored the signal. A model that removed the signal feature from the decision performed similarly. Therefore, a portion of model agents use a strategy of ignoring the signal. This value was set at 23% and is based on survey responses reported in Cranford et al. (2020a).

5.2. Results

To assess performance of humans and models, we analyzed the mean probability of attack per round, the distributions of overall probabilities of attack, and defender expected utilities. Figure 4 shows the mean probability of attack per round. Humans attack less often under Confusion Signaling than the peSSE. The models for the peSSE and Confusion Signaling fit very well to the human data; in fact, the correlation and RMSE of the mean probability of attack across trials is (r = 0.78, $RMSE$ = 0.037) and (r = 0.88, $RMSE$ = 0.051), respectively. Compared to the peSSE, when uncertainty is added to the peSSE, model predictions show a decrease in probability of attack. However, Confusion Signaling is expected to further reduce the probability of attack beyond that of the peSSE.

As can be seen in the histograms in Figure 5, Confusion Signaling results in many fewer participants that attack almost all the time. Many more participants cluster around attacking approximately 60% of the time. The cognitive models of the peSSE and Confusion Signaling match very well to the full distribution of human participants. Compared to the peSSE with uncertainty, Confusion Signaling is expected to reduce the number of participants that attack most of the time and shift them toward the middle part of the distribution.

The defender expected utility is shown in Figure 6. The human data shows that the Confusion Signaling results in lower defender expected utility than the peSSE, but most of the effect is in the first two rounds. Meanwhile, defender utilities are about the same in the last two rounds. The peSSE and Confusion Signaling models match well to the human data, but there is some underestimation of the peSSE model in rounds 3 and 4.

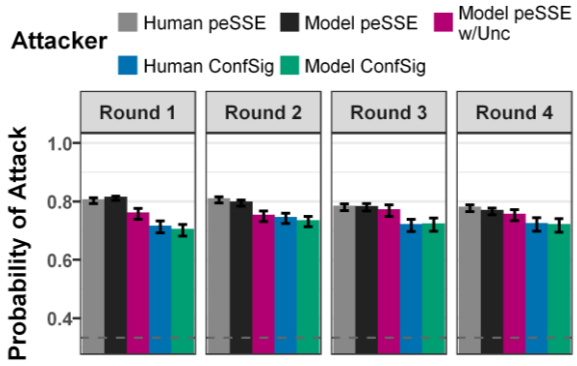

Figure 4. Mean probability of attack across trials comparing humans and models across signaling algorithms.

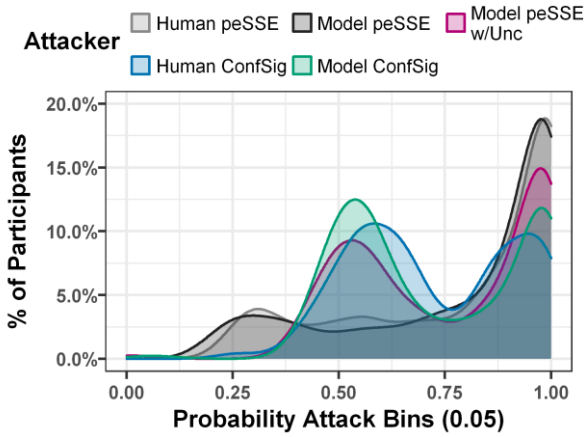

Figure 5. Distribution of overall probability of attack comparing humans and models across signaling algorithms.

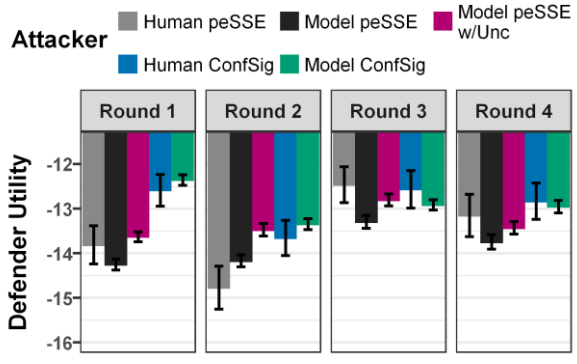

Figure 6. Mean defender expected utility per round comparing humans and models across signaling algorithms.

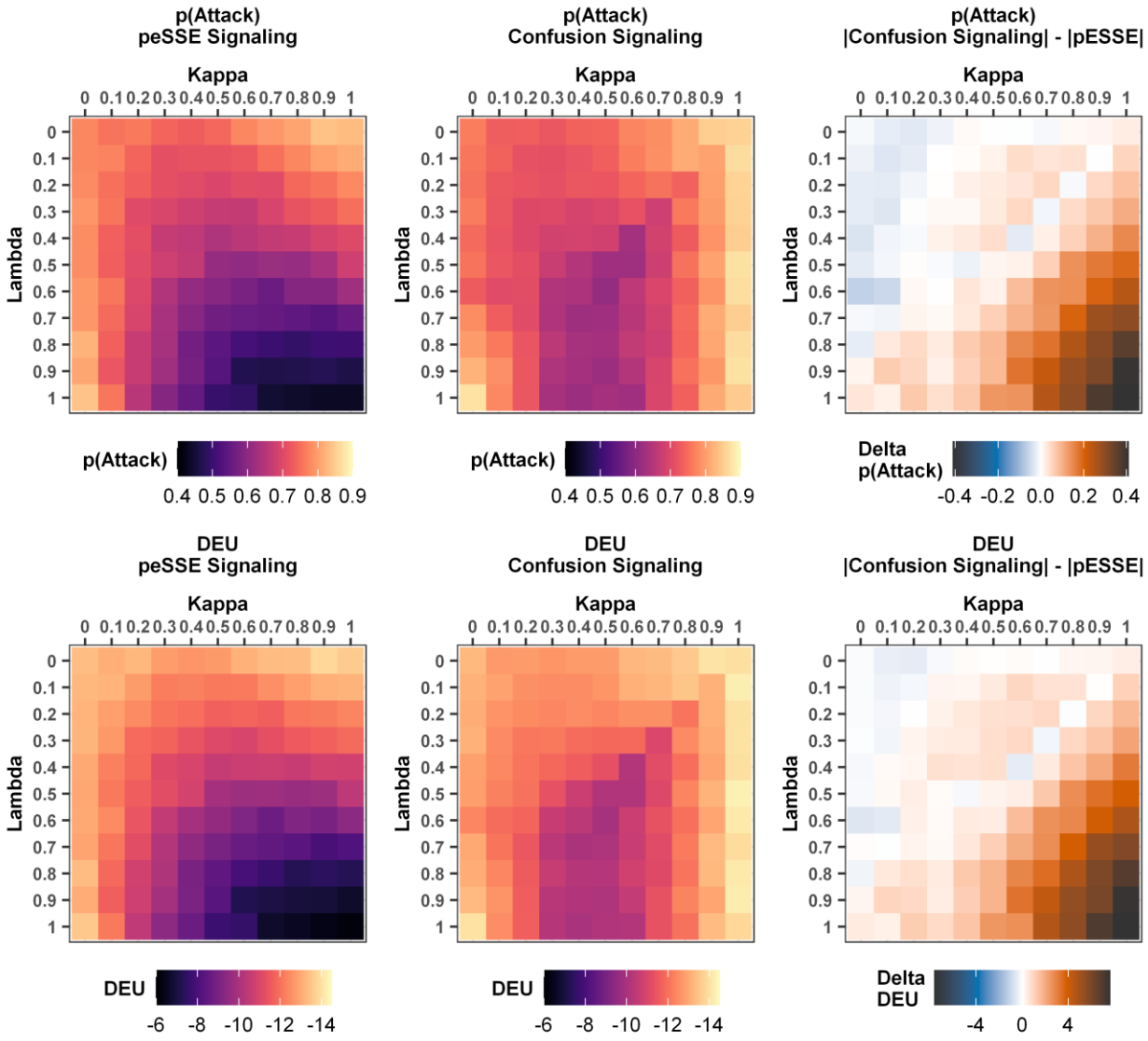

Figure 7. Revised cognitive model predictions of Confusion Signaling compared to the peSSE across varying levels of signal uncertainty. The rightmost figures show the difference between Confusion Signaling and peSSE.

5.3. Revised predictions of Confusion Signaling

Based on the observations of human performance, and given the model revisions described above, we tasked the model to make new predictions regarding the effects of uncertainty on attacker behavior in the IAG. As before, for each pair of values, the model was run 1000 times to simulate a population of individuals.

The results presented in Figure 7 show a different pattern than observed with the original cognitive model. Based on the insight that humans adapt to kappa uncertainty by overcorrecting given a signal via mulling, the peSSE model now attacks less often, resulting in lower defender expected utility, as kappa and lambda increase together. Within the safe zone, Confusion Signaling is expected to mostly reduce attacks and improve defender expected utility compared to the peSSE, but only maintains about the level of the peSSE as kappa uncertainty becomes large. Below the safe zone, Confusion Signaling tends to fail and the peSSE remains a better signaling algorithm.

6. Conclusion

The present research aimed to design a deceptive signaling algorithm for cybersecurity that effectively accounts for uncertainty in real-time information processing. Simulations showed that our solution, the

Confusion Signaling algorithm, should prove successful in reducing attacks and improving defender expected utility, or at least maintaining them, compared to the peSSE when signal uncertainty is present. The Confusion Signaling algorithm results in differences in attacker behavior across levels of uncertainty. This makes it clear that it is important to understand the amount of uncertainty in a particular environment. It also opens the possibility of directly manipulating uncertainty to induce desired levels of performance from the defense algorithm. The benefit of Confusion Signaling is currently limited to a safe zone where the combined probability of kappa and lambda uncertainty is less than 1. Future research will address this limitation by considering how to successfully adapt the signal in this zone. One method may be to mirror the signaling probabilities within the safe zone.

The human behavior study revealed interesting and unforeseen reaction to signal uncertainty. Particularly, given kappa uncertainty, users seem to mull on those events in a way that causes fewer attacks on signaled targets instead of fewer attacks on un-signaled targets. These findings have important implications for signaling algorithms that attempt to account for uncertainty. Instead of reducing the expected value of attacks on un-signaled targets, algorithms need to account for kappa uncertainty by reducing the expected value of attacks on signaled targets. Future research is aimed at revising the Confusion Signaling algorithm to account for this behavior of boundedly rational humans.

Another avenue of future research will be to investigate the ecological validity of the current findings. The abstracted IAG task does not necessarily reflect the scale of real-world rewards and penalties that professional attackers would experience, and the Mechanical Turk participants lack their decision-making expertise and motivation. It is possible that professional attackers would better account for, and remain robust in the face of, environmental uncertainty.

Finally, the present research only accounts for signal uncertainty because detection is not necessary for static signaling methods for cybersecurity. However, when personalized, adaptive methods are used (e.g., Cranford et al., 2020a; 2020b), detection is key to track attacker decision-making in order to adapt a model to an individual and provide informed recommendations for adapting the signaling scheme. Therefore, future research is aimed at exploring how to account for detection uncertainty in the IAG.

Acknowledgements

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

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. doi:10.1037/0033-295X.111.4.1036

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum. doi:10.4324/9781315805696

Bondi, E., Oh, H., Xu, H., Fang, F., Dilkina, B., & Tambe, M. (2020). To signal or not to signal: Exploiting uncertain real-time information in signaling games for security and sustainability. *Proceedings of AAAICAI*, *34*(02), 1369-1377. doi:10.1609/aaai.v34i02.5493

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., & Lebiere, C. (2020a). Adaptive cyber deception: Cognitively-informed signaling for cyber defense. *Proceedings of the 53rd HICSS* (pp. 1885-1894). doi:10.24251/HICSS.2020.232

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., & Lebiere, C. (2020b). Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science*, *12*, 992–1011. doi:10.1111/tops.12513

Cranford, E. A., Gonzalez, C., Aggarwal, P., Tambe, M., & Lebiere, C. (2020c). What attackers know and what they have to lose: Framing effects on cyber-attacker decision making. *Proceedings of the HFES*, *64*(1), 456–460. doi:10.1177/1071181320641102

Cranford, E. A., Gonzalez, C., Aggarwal, P., Tambe, M., Cooney, S., & Lebiere, C. (2021). Towards a cognitive theory of cyber deception. *Cognitive Science, 45:e13013*, 1–28. doi:10.1111/cogs.13013

Cooney, S., Vayanos, P., Nguyen, T. H., Gonzalez, C., Lebiere, C., Cranford, E. A., & Tambe, M. (2019). Warning time: optimizing strategic signaling for security against boundedly rational adversaries. In *Proceedings of the 18th* AAMAS (p. 1892–1894).

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S., Hau, R., Hertwig, R., Stewart, T., West, R., Lebiere, C. (2010). A choice prediction competition, for choices from experience and from description. *Journal of Behavioral Decision Making 23*(1): 15-47. doi:10.1002/bdm.683

Gonzalez, C. (2013). The boundaries of instance-based learning theory for explaining decisions from experience. *Progress in Brain Research*, *202*, 73–98. doi:10.1016/B978-0-444-62604-2.00005-8

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635. doi:10.1007/978-3-319-11391-3_6

Lebiere, C. (1999). A blending process for aggregate retrievals. *Proceedings of the 6th ACT-R Workshop*.

Tambe, M. (2011). *Security and game theory: Algorithms, deployed systems, lessons learned*. Cambridge University Press. doi:10.1017/CBO9780511973031

Xu, H., Rabinovich, Z., Dughmi, S., & Tambe, M. (2015). Exploring information asymmetry in two-stage security games. *Proceedings of the NCAI* (2, pp. 1057–1063). Austin, TX: Elsevier B.V.