COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e13013

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13013

Towards a Cognitive Theory of Cyber Deception

Edward A. Cranford,[a] Cleotilde Gonzalez,[b] Palvi Aggarwal,[b] Milind Tambe,[c,#] Sarah Cooney,[c] Christian Lebiere[a]

[a]*Department of Psychology, Carnegie Mellon University*
[b]*Social and Decision Sciences Department, Carnegie Mellon University*
[c]*USC Center for AI in Society, University of Southern California*

Received 25 June 2020; received in revised form 28 February 2021; accepted 7 June 2021

Abstract

This work is an initial step toward developing a cognitive theory of cyber deception. While widely studied, the psychology of deception has largely focused on physical cues of deception. Given that present-day communication among humans is largely electronic, we focus on the cyber domain where physical cues are unavailable and for which there is less psychological research. To improve cyber defense, researchers have used signaling theory to extended algorithms developed for the optimal allocation of limited defense resources by using deceptive signals to trick the human mind. However, the algorithms are designed to protect against adversaries that make perfectly rational decisions. In behavioral experiments using an abstract cybersecurity game (i.e., Insider Attack Game), we examined human decision-making when paired against the defense algorithm. We developed an instance-based learning (IBL) model of an attacker using the Adaptive Control of Thought-Rational (ACT-R) cognitive architecture to investigate how humans make decisions under deception in cyber-attack scenarios. Our results show that the defense algorithm is more effective at reducing the probability of attack and protecting assets when using deceptive signaling, compared to no signaling, but is less effective than predicted against a perfectly rational adversary. Also, the IBL model replicates human attack decisions accurately. The IBL model shows how human decisions arise from experience, and how memory retrieval dynamics can give rise to cognitive biases, such as confirmation bias. The implications of these findings are discussed in the perspective of informing theories of deception and designing more effective signaling schemes that consider human bounded rationality.

Keywords: Decision making; Cognitive model; Deception; Cybersecurity; Signaling; Instance-based learning theory; ACT-R; Stackelberg security game

[#] Milind Tambe is now at the Center for Research in Computation and Society, Harvard University.

Correspondence should be sent to Edward A. Cranford, Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA. E-mail: cranford@cmu.edu

> All warfare is based on deception…when [we are] far away, we must make [the enemy] believe we are near.
>
> –Sun Tzu, *The Art of War*

1. Introduction

In the surging world of cyber communication, deception thrives due to a lack of physical cues for detection. We are presented with an abundance of data through the internet and social media, where increasingly more human interactions take place. Thus, it is important to understand how deception influences human decision-making in the cyber domain. Throughout history, the psychology of deception has spanned the biological, cognitive, and social levels of cognition, with research examining physiological reactions, cognitive biases, and other effects on decisions, social interactions, as well as societal implications (for review, see Hyman, 1989; Newell, 1990). While there is much research in the psychology of deception, most of what is currently known consists of the use of verbal and nonverbal physical cues, including appearance, gestures, and descriptions, and the role of these attributes in the context of social interaction (Bond & DePaulo, 2008; Morgan, LeSage, & Kosslyn, 2009; Riggio & Friedman, 1983). For example, most studies frame the study of deception as it relates to the body, face, and the cues that may be leaked through gestures and words (Riggio & Friedman, 1983). In other words, most of what we know about the psychology of deception relies on the physical observation of behavior, and little is known regarding the psychology of deception in cyber domains.

Deception typically involves one agent (the sender) presenting truthful or false information (a signal) to an opponent (the receiver) in order to gain an advantage over the opponent. For example, a poker player with a weak hand may make a high raise (a false signal) in an attempt to intimidate their opponent into thinking they have a strong hand. If the opponent believes the signal, then the deception is successful and the opponent will fold; otherwise, the deception fails. Formally, deception has been defined as a form of persuasion where one intentionally misleads an agent into a false belief, in order to gain an advantage over the agent and to achieve one's goals (Rowe & Rrushi, 2016).

A deception succeeds through the exploitation of human processing constraints and perceptual, cognitive, and social biases (Mokkonen & Lindstedt, 2016). For example, magicians use sleight-of-hand by exploiting perceptual biases and limitations of the visual and attention systems (Ekroll & Wagemans, 2016). Other research showed that basketball referees had a tendency to refrain from calling a penalty on the offense for charging into a defender when the defender did not fall down, which can be explained through the representative heuristic that defenders typically fall down on illegal charges (Morgulev, Azar, Lidor, Sabag, & Bar-Eli, 2014). The defenders then exploited this bias and tried to deceive the referees by intentionally falling even when there was no penalty committed in order to draw a call. However, due to the high rate of deception from defenders, the referees responded by calling fewer penalties when the defenders fall from a legitimate contact. While this resulted in fewer incorrect calls (i.e., lower false alarm rate), fewer legitimate calls were made (i.e., lower hit rate), aligning with utility-maximizing decision-making and ecological rationality (Morgulev et al., 2014).

In the physical world, there is an abundance of cues that a deceiver must conceal from the adversary to avoid detection, including physiological, verbal, and nonverbal cues (Rowe & Rrushi, 2016). For example, our poker player tries to keep a “straight face” to avoid leaking any physical cues to their deception. On the other hand, in the cyber world, deception is highly successful given that there are very few physical cues available, making it easier to conceal. Cyber attackers use deception in the form of phishing emails, fake websites and malware, social engineering, misinformation campaigns, and so on. Defenders also employ deception using tactics such as honeypots, which are “fake” nodes in a network designed to appear attractive to attackers (Gonzalez, Aggarwal, Cranford, & Lebiere, 2020). In these examples, only a few verbal cues may be leaked through text (especially so for phishing emails and other deceptions involving text or dialogue), and nonverbal cues only include inconsistent interactions with technology (e.g., the system responds too fast/slow; Rowe & Rrushi, 2016). Therefore, most cyber deceptions focus on exploiting social and cognitive biases (Almeshekah & Spafford, 2016).

Stech, Heckman, and Strom (2016) propose that cyber deception for defense could be achieved by revealing or concealing the facts and fictions about the system. Deception techniques such as masking and repackaging conceal the facts to hide the true state of the network (Aggarwal et al., 2020). In contrast, techniques such as decoying and mimicking reveal the fictitious state of the network to mislead attackers (Aggarwal, Gonzalez, & Dutt, 2016). In this paper, we use signaling as a deception technique that combines revealing the facts and fiction about the true state of the network. The focus of the present research is on how deception using signaling influences human decision-making, and what biases influence those decisions when there are no observable cues for detecting deception. When humans are aware of the possibility of deception but have no cues to detect deception in the current context, the situation requires decision-making under uncertainty. In these situations, human vulnerabilities can make them susceptible to deception. Almeshekah and Spafford (2016) suggest that the same cognitive biases that influence human decision-making under physical deception can be exploited with cyber deception, including representativeness, availability, anchoring heuristics, and confirmation bias, to name a few.

A useful framework for studying decision-making under deception is with economic signaling games (Jenkins, Zhu, & Hsu, 2016; Moisan & Gonzalez, 2017). Signaling games are two-player games involving a sender and a receiver with incomplete information (Pawlick, Colbert, & Zhu, 2019). The sender has information unknown to the receiver and strategically reveals this information (either truthfully or deceptively) to the receiver in order to influence their decision-making (Battigalli, 2006; Cho & Kreps, 1987). According to Jenkins et al. (2016), after being presented a signal, the receiver’s decision-making process includes: (1) perceiving and comprehending the signal, (2) anticipating the outcome resulting from the signal, (3) evaluating the expected utility of possible actions, and (4) selecting the best action. This process is consistent with research on how humans make decisions under uncertainty. People rely on their own experience (e.g., in accordance with instance-based learning theory (IBLT); Gonzalez, Lerch, & Lebiere, 2003). Specifically, IBLT’s cognitive processes include information search, recognition and similarity processes, integration and accumulation of information, feedback, and learning (Gonzalez, 2013). These processes involve the estimation and anticipation of outcomes for possible actions and their probabilities based on prior observations and learning through feedback. The signaling framework and the underlying theories of decisions from experience provide a foundation for investigating deception as an adaptive process and how it influences dynamic decision making.

This paper studies the behavior of humans (in the role of attackers), who make decisions under various types of deceptive signaling defense strategies. We use the cybersecurity domain because it presents particular challenges regarding the non-physical cues for studying decisions under uncertainty and in the face of deception. Through laboratory experiments and the development of cognitive models that mimic human behavior, we aim at informing theories of deception in general and improving the design of cyber-defense mechanisms.

2. Game-theoretic models, signaling, and security

Recent work using game-theoretic models (Al-Shaer, Wei, Hamlen, & Wang, 2019), in particular, the research program on Stackelberg Security Games (SSGs; Sinha, Fang, An, Kiekintveld, & Tambe, 2018) has greatly improved physical security systems. Examples range from protecting ports and airports, scheduling air marshals, and mitigating poacher attacks. SSGs model the interaction between a defender and an adversary as a leader-follower game (Tambe, 2011) and are used to develop algorithms that optimally allocate limited defense resources over a set of targets (Pita et al., 2008; Shieh et al., 2012; Sinha et al., 2018; Tambe, 2011). Such algorithms could prove useful for cybersecurity, where it is often the case that organizations have limited defense resources to actively monitor a large number of computers on a network. These defense-allocation methods could be cost-effective if they use deceptive techniques to fool the attacker into thinking the defender is monitoring a target when in fact they are not and more generally reduce the attacker's assessment of his prospects by increasing the perceived coverage.

Typically, SSGs do not involve deception (Abbasi et al., 2016). However, Xu, Rabinovich, Dughmi, and Tambe (2015) extended the SSG by incorporating *signaling*, in which a defender (sender) deliberately reveals information about their strategy to the attacker (receiver) in order to influence the attacker's decision-making (Battigalli, 2006; Cho & Kreps, 1987). The attacker selects a target they wish to attack, then the defender sends a signal that divulges the protection status of a target (i.e., the target is monitored or not monitored), and finally the attacker decides whether to continue the attack or withdraw. Adopting this approach, truthful signals can deter some attacks on protected targets, but if a target is unprotected, then the attacker can attack with impunity. To help protect the unprotected resources, defenders can use a combination of truthful and deceptive signals to increase the perceived coverage of the targets. Xu et al.'s (2015) solution, the Strong Stackelberg Equilibrium with Persuasion (peSSE), determines the optimal combination of bluffing (sending a message that the target is covered when it is not) and truth-telling (sending a truthful message that the target is covered or not) so the attacker continues to believe the signal.

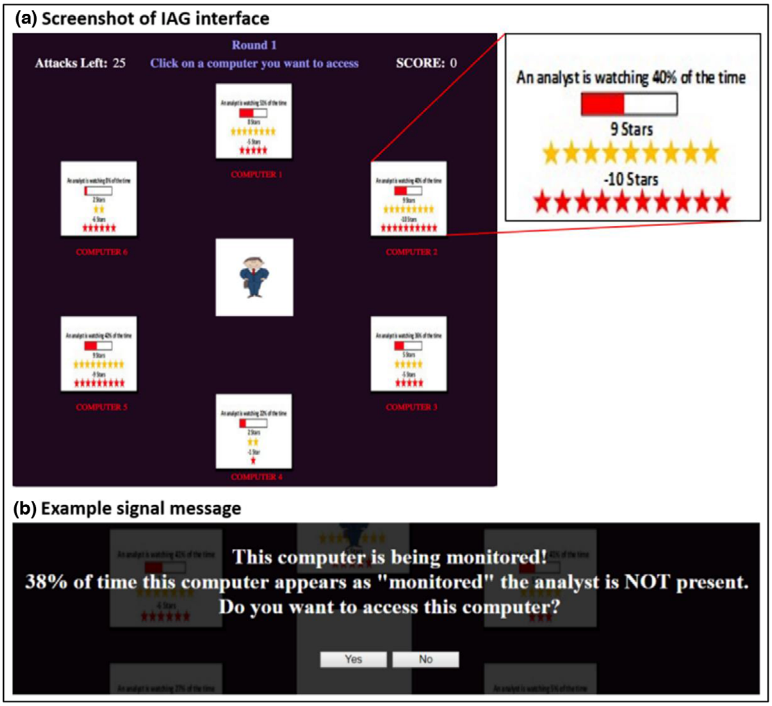
The peSSE has been formally proven to improve defender utility against a perfectly rational attacker (i.e., one that has complete information and unlimited processing capacity to make decisions that maximize reward), compared to strategies that do not use signaling (Xu et al., 2015). However, humans exhibit, at best, bounded rationality (Simon, 1956) and may interpret and react to deceptive signals in uncertain ways. From a rational analysis perspective (Anderson, 1991), the peSSE is optimized for environments in which there is no uncertainty because decision-makers are perfectly rational. In environments where signals are always truthful, humans may make decisions that look perfectly rational because the available information provides certainty. However, for environments in which the signals are sometimes deceptive, the assumptions of certainty may not be optimal for boundedly rational humans. While a perfectly rational adversary, with unlimited resources, can make optimal decisions in such environments, a boundedly rational human will make decisions based on the limited information that is often incomplete or incorrect (i.e., retrieved incorrectly from memory). Therefore, we can expect humans to make decisions that include bias, emerging from decision heuristics and memory retrieval processes over past experiences that reflect the probabilistic nature of cues in the world.

To gain a better understanding of how deceptive signals influence human decision-making, we pit humans against the peSSE in a cybersecurity game called the *Insider Attack Game* (IAG). Our methods involve a combination of game theoretical defense and signaling algorithms, cognitive modeling for representing attacker's behavior, and human laboratory studies to gain a better understanding of attacker behavior playing against the different defense strategies (Gonzalez et al., 2020). The present research advances the psychology of deception through understanding the cognitive processes involved in making decisions in the face of deceptive signals and the emergent biases that affect those decisions, leading to better signaling algorithms for defense against boundedly rational humans (see Cooney et al., 2019).

In what follows, we first describe the IAG. Then, we introduce a cognitive model of attacker behavior, constructed according to IBLT of decisions from experience (Gonzalez et al., 2003), and implemented in the Adaptive Control of Thought-Rational (ACT-R) cognitive architecture (Anderson & Lebiere, 1998; Anderson et al., 2004). This model is pitted against various defense algorithms to provide predictions of human behavior and the effectiveness of deceptive signals. These predictions are then tested against human performance in the IAG obtained from laboratory experimentation. Our analyses reveal human reactions and biases to deceptive signals. We discuss the implications of our findings for the psychology of deception, key insights and lessons learned through cognitive modeling efforts, and the potential of cognitive models for the design of novel signaling schemes that account for boundedly rational human adversaries. The experimental materials, data, analysis scripts, model code, and other supplemental material can be accessed at https://osf.io/jn69t.

3. IAG

We developed an abstract game for which we can carefully examine the basic underlying cognitive processes of decision-making in the presence of deceptive signals. The SSG is a generic algorithm that has been applied in naturalistic settings (e.g., Tambe, 2011), and we adapted it to an insider attack scenario. In the IAG, players take the role of employees at a


(a) Screenshot of IAG interface

Round 1

Attacks Left: 25 Click on a computer you want to access SCORE: 0

An analyst is watching 40% of the time

9 Stars

-10 Stars

(b) Example signal message

This computer is being monitored!
38% of time this computer appears as "monitored" the analyst is NOT present.
Do you want to access this computer?

Fig. 1. (a) Screenshot of the insider attack game. The attacker is represented in the center surrounded by six computer targets. Each target shows the probability of the computer being monitored displayed as a percentage in text and represented in a fillable gauge (the red bars), the payment one would receive if they attacked an uncovered target (yellow stars), and the penalty one would receive if they attacked a covered target (red stars). (b) Example message for the *Strong Stackelberg Equilibrium with Persuasion (peSSE)-Full Information* (*FI; peSSE-FI*) condition when a signal is present. The *peSSE* condition omits line two from the message. When the signal is absent (and for all messages in the *NoSignal* condition), lines one and two are omitted.

company, and their goal is to maximize their score by “hacking” computers to steal proprietary information (i.e., human players take the role of “attackers”). However, two security analysts (i.e., “defenders”) monitor the computers. Attackers can earn points if they avoid the defenders but lose points if they are caught. From the defenders’ perspective, the game is a two-stage SSG. As in classic single-stage SSGs, the first stage involves allocating the defenders. The allocation of the defenders is optimized by computing the Strong Stackelberg Equilibrium (SSE), which provides the monitoring probability (m-prob) of each computer based on their reward and penalty values (Tambe, 2011). An attacker then makes a move by selecting a computer to attack. In the second stage, after a computer is selected, defenders can take advantage of deceptive signaling techniques by strategically revealing potentially deceptive information to the attacker about whether the computer is being monitored (Xu et al., 2015). The attacker follows by deciding whether to continue the attack or withdraw.

A screenshot of the task interface is shown in Fig. 1(a). Attackers perform four rounds of 25 trials each, following an initial practice round of five trials. For each round, attackers are

Table 1
Attribute values (reward, penalty, monitoring probability) for each computer target in each round

Round	Target 1	Target 2	Target 3	Target 4	Target 5	Target 6
Round 1	[2, −1, 0.22]	[8, −5, 0.51]	[9, −9, 0.42]	[9, −10, 0.40]	[2, −6, 0.08]	[5, −5, 0.36]
Round 2	[5, −3, 0.41]	[8, −5, 0.48]	[7, −6, 0.41]	[8, −9, 0.37]	[5, −7, 0.27]	[2, −4, 0.05]
Round 3	[3, −3, 0.30]	[9, −4, 0.60]	[6, −6, 0.40]	[5, −8, 0.29]	[3, −6, 0.20]	[2, −2, 0.20]
Round 4	[4, −3, 0.37]	[6, −3, 0.51]	[7, −7, 0.40]	[5, −10, 0.24]	[5, −9, 0.26]	[3, −4, 0.23]

presented with six new computer targets, each with a different payoff (reward/penalty) structure. On a given trial, the two defenders monitor one computer each. Attackers can view information describing each target's reward and penalty values, as well as the m-prob (represented as a percentage and described as the "average amount of time that the target is monitored"). This information is provided to participants because an assumption of perfect rationality requires that the agent knows all relevant information when making a decision (i.e., we assume a pre-attack reconnaissance phase in which the attacker can observe and gain as much information as needed regarding the defender's strategy).

For each trial, attackers first select one of the targets to attack. After selection, they are presented with possibly deceptive information about whether the computer is being monitored (Fig. 1(b)). If the message claims that the computer is monitored, then the signal is deemed present, else it is absent. The attacker must then decide to either continue or withdraw the attack. An attack on a computer that is monitored results in losing points equal to the penalty associated with the target, whereas if the computer is not monitored the attacker gains points equal to the reward associated with the target. If the attacker chooses to withdraw the attack, they receive zero points.

Each round consists of a different set of computers with different payoff structures, which results in a different allocation of defense resources. Table 1 shows the rewards, penalties, and m-probs for each computer in each round. The m-probs for each target are derived by computing the SSE, which allocates defenses across in such a manner that the expected value of attacking each computer is positive and all equal. Each attacker experiences the same schedule of coverage and signaling throughout the game. That is, the SSE allocates defenses across the 25 trials for each round, and so predetermines which targets are monitored during each trial.

The peSSE determines which targets should present warning signals if selected. If a computer is being monitored, then the attacker will always receive a truthful signal claiming that the computer is currently monitored. However, if a computer is not being monitored, then sometimes the attacker will not receive a signal, which truthfully indicates that the computer is unmonitored, and other times the attacker will receive a deceptive signal claiming that the computer is being monitored. The conditional probability of sending a deceptive signal given a target is unmonitored is optimized by computing the peSSE (Xu et al., 2015):

$$p(signal|unmonitored) = \frac{(|penalty| \times m{-}prob)\,/reward}{1 - m{-}prob} \tag{1}$$

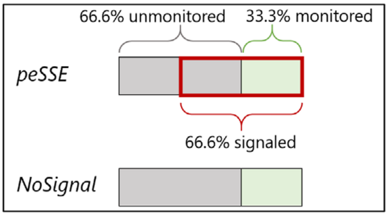


Fig. 2. Average probability of signaling given a target is monitored or not monitored. According to the *peSSE* algorithm, on average, a target is monitored one-third of the time and unmonitored two-thirds of the time. Signals are presented two-thirds of the time: 100% of the time when a target is monitored and 50% of the time a target is unmonitored, on average, so that the expected value of attacking given a signal is equal to zero (and equal to the expected value of withdrawing). The *peSSE* is compared to a condition that presents no signals, called *NoSignal*.

This value is the one that makes the expected value of attacking the computer, given a signal, equal to zero. In other words, as depicted in Fig. 2, given two defenders for six targets, a target is monitored one-third of the time on average and the peSSE will send additional signals on unmonitored targets until the expected value of attacking given a signal reaches zero.[1] This value is equal to the expected value of withdrawing the attack, and based on game-theoretical assumptions of perfect rationality, it is more advantageous (from an economical perspective) for an adversary to break ties in favor of the defender and withdraw their attack when presented a signal.

3.1. *Experimental conditions: Defense signaling schemes*

The IAG was deployed under several experimental conditions in order to assess the effectiveness of deceptive signals on attackers' decision-making. The baseline condition of the game is called the *peSSE* signaling scheme, which presents signals with probabilities calculated according to the peSSE algorithm as described above. In the *NoSignal* condition, a signal is never presented to the attackers, regardless of whether a defender is present or absent (i.e., no deception was used). The *peSSE-Full Information (FI)* signaling scheme is similar to the *peSSE* except that, upon seeing a signal, attackers are also provided the percent of time that the signal is deceptive for that target (see Fig. 1(b); e.g., signal messages for each condition are also provided in the Experimental Methods section). That is, in the *peSSE-FI* condition, we extend the assumptions of perfect rationality to ensure that all attackers have full knowledge of the probabilities of deception available to them, in addition to the m-probs.

4. An IBL model of attackers in the IAG

Cranford et al. (2018) created an IBL cognitive model of an IAG attacker to make predictions about how human participants would perform in the various experimental conditions. That model was later simplified to better represent human behavior when playing the IAG. These modifications generally involved replacing overly complex and implausible mechanisms, memories, and procedures and reverting parameter setting to their default values. The details and implications of these changes are discussed in the online Supplemental Material available at https://osf.io/tpfnv/. The cognitive model is implemented in the ACT-R cognitive architecture (Anderson & Lebiere, 1998; Anderson et al., 2004), and decisions are made following the methodology of IBLT (Gonzalez et al., 2003). A model based on mechanisms of the ACT-R architecture limits free parameters and constrains assumptions of representation and processes. More broadly, the model helps understand the interaction between cognitive processes and deceptive signals, particularly how a boundedly rational agent engages with deceptive interactions. IBLT has been used to model decision-making processes across several tasks with much success (Hertwig, 2015). Applications of IBL models include supply chain management (Gonzalez & Lebiere, 2005), social dilemmas (Gonzalez, Ben-Asher, Martin, & Dutt, 2015; Juvina, Saleem, Martin, Gonzalez, & Lebiere, 2013; Lebiere, Wallach, & West, 2000), two-person games (Sanner, Anderson, Lebiere, & Lovett, 2000; West & Lebiere, 2001), repeated binary-choice decisions (Gonzalez & Dutt, 2011; Lebiere, Gonzalez, & Martin, 2007), and classical single-stage SSGs (Abbasi et al., 2016).

In IBLT, decisions are made by generalizing across past experiences, or instances, that are similar to the current situation. Typically, experiences are encoded as chunks in declarative memory that contain the attributes describing the context in which each decision is made, the decision itself, and the outcome of that decision. This is consistent with the no-magic doctrine of cognitive modeling (Anderson & Lebiere, 1998) in which only the information directly available to the subject is represented in the cognitive model, and no additional knowledge constructs or preprocessing stages are assumed. In the attacker's model of the IAG, the context attributes include the probability that a computer is being monitored (m-prob; range 0 to 1.0), the value of the reward (range 0 to 10), the value of the penalty (range 0 to -10), and whether or not a signal is presented (present or absent). The possible decisions are attack or withdraw, and the outcome is the actual points received based on the action.

In a given situation, for each possible decision, an expected outcome is generated from memory through a retrieval mechanism called *Blending*. The decision with the highest expected outcome is made. In the present game, withdrawing always results in zero points. Therefore, the model only needs to determine the expected outcome of attacking in order to make a choice. For each decision, the model takes the description of each target and generates an expected outcome of attacking that target by retrieving similar past instances. In ACT-R, the retrieval of past instances is based on the activation strength of the relevant chunk in memory and its similarity to the current context. The activation A_i of a chunk i is determined by the following equation:

$$A_i = \ln \sum_{j=1}^{n} t_j^{-d} + MP * \sum_{k} Sim(v_k, c_k) + \varepsilon_i \tag{2}$$

The first term provides the power law of practice and forgetting, where t_j is the time since the jth occurrence of chunk i and d is the decay rate of each occurrence which is set to the default ACT-R value of 0.5. The second term reflects a partial matching process, where $Sim(v_k,c_k)$ is the similarity between the actual memory value and the corresponding context element for chunk slot k and is scaled by the mismatch penalty, which was set to 1.0. The term ε_i represents transient noise, a random value from a logistic distribution with a mean of zero and variance parameter s of 0.25 (common ACT-R value, e.g., Lebiere, 1999), to introduce 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 values.

The activation of a particular chunk determines the probability of retrieving that chunk according to the softmax equation, also known as the Boltzmann equation, reflecting the ratio of each chunk activation A_i and the temperature t, which was set to a neutral value of 1.0:

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}} \tag{3}$$

The IBL model uses ACT-R's blending mechanism (Gonzalez et al., 2003; Lebiere, 1999) to calculate an expected outcome of attacking a target based on past instances. Blending is a memory retrieval mechanism that returns a consensus value across all memories rather than a specific memory as computed by the following equation:

$$\underset{V}{\operatorname{argmin}} \sum_i P_i \times (1 - Sim\,(V, V_i))^2 \tag{4}$$

The value V is the one that best satisfies the constraints among actual values in the matching chunks i weighted by their probability of retrieval P_i. That objective is defined as minimizing the dissimilarity between the consensus value V and the actual answer V_i contained in chunk i. For the simple case of continuous values such as real numbers, this equation effectively specifies a weighted averaging process. To generate the expected outcome of a decision, the model matches memories to the current decision context and intended decision and uses blending to return the expected outcome. That value is not the true expected outcome but instead reflects the individual's limited experience as well as various statistical biases (e.g., recency, frequency, etc.; Lebiere et al., 2007; 2013).

4.1. The IBL model procedure

To begin the IAG, the model is initialized with seven instances: five represent a simulated practice round similar to that experienced by the human participants (i.e., for each practice trial, the model randomly selects one of the targets, uniformly distributed, and always decides to attack; a chunk is stored that represents the target context, the action to attack, and the signal and outcome values based on the coverage and signaling schedule of the practice round), and two represent knowledge gained from instructions (i.e., one instance had a signal value of absent and an outcome of 10, representing that attacking when a signal is absent will always result in a reward; another instance had a signal value of present and an outcome of 5, representing that attacking when a signal is present could result in either a penalty or a reward). This method ensures each run of the model begins with a different set of initial experiences, much like individual humans. The model requires these initial instances to begin or else it would fail to retrieve a projected outcome. These initial instances are quickly overwhelmed

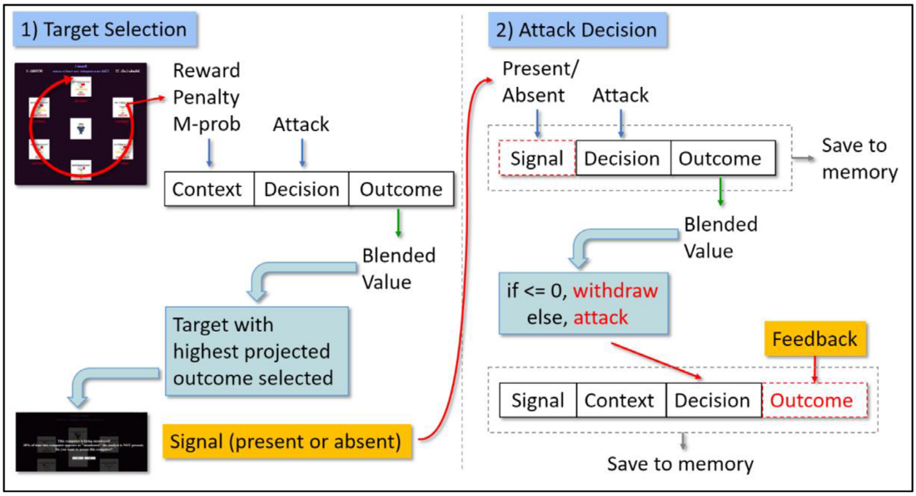


Fig. 3. The instance-based learning (IBL) model procedure.

by actual experience as the activation strengths of these chunks decay and are never reinforced. Therefore, they do not play a large role past the initial decisions due to their low probabilities of retrieval.

Fig. 3 shows how the IBL model of the attacker in the IAG operates on a given trial. In general, reflecting a bounded rationality approach, the model selects a computer with the highest projected outcome and then decides to attack if the projected outcome, given the signal value, is greater than zero. The first step is to select a computer to attack (Fig. 3, left panel). The model iterates through each of the six computers and generates a projected outcome of attacking each computer through blending retrieval as described above. At this point, there is no signal information, so the signal slot is ignored. The model keeps track of the computer with the highest projected outcome of attacking and then selects that computer to attack. Target selection is a process driven by generated expectations of ultimate outcomes (payoff or penalty) rather than past selection decisions; therefore, no instances of selection decisions are saved in memory since they would not influence future decisions.

Next, the model is presented a signal, or not, based on the computer that was selected. The context is augmented with a signal slot representing whether a signal is present or absent. In the *peSSE-FI* condition, when a signal is present, the context is further augmented with an additional slot representing the probability the signal is deceptive as presented to the human participants.

In the second step of the decision process (Fig. 3, right panel), the model retrieves a blended value representing the updated projected outcome of attacking the selected computer, given the value of the signal. As can be seen in Fig. 1(b), the pop-up signal message occludes all information about the selected target, so we inferred that humans base their decisions only on the value of the signal and ignore, or forget, the occluded target information. In other words, the target information is unlikely to be used in the decision unless the players explicitly maintain that information in working memory. Therefore, the similarity of the selected target's context to past instances is based solely on the value of the signal (i.e., m-prob, reward, and penalty values are ignored in step 2). Because humans tend to remember not only the actual experience but also their expectations about the experience (Gonzalez et al., 2003), a new instance is saved in declarative memory that reflects the model's expectations. This instance includes only slots for the signal value, decision to attack, and the projected outcome generated via blending. Therefore, these expectations do not play a role in future target selection because they lack the full decision context but only during future attack decisions. Finally, if the projected outcome is less than or equal to zero, then the model withdraws the attack, the decision slot is set to "withdraw," and the outcome slot is updated with the value zero. If the projected outcome is greater than zero, then the model attacks, the decision slot is set to "attack," and the outcome slot is updated to reflect the ground-truth outcome observed after the choice is made (feedback). This final instance, which includes the signal value, target context, ground-truth decision, and ground-truth outcome, is then saved in declarative memory.

In the *NoSignal* condition, because warning signals are never presented, the signal slot is removed entirely from the context representation. Therefore, in step 2, the model bases its decision solely on the action to attack. Additionally, the initialized instance that represents knowledge from instructions when a signal was present was modified so that the outcome was set to 10, matching the instance that represents knowledge from instructions when a signal is absent. This modification was necessary to keep the model exploring in early trials; otherwise, it could get trapped in a cycle of not attacking because the initial expectations are negative and never get a chance to increase. This is a common practice for triggering human-like exploration in early trials of a task without explicitly modeling strategies or other metacognitive processes (e.g., Lebiere et al., 2007).

The model's behavior reflects its experiences. If an action results in a positive/negative outcome, then the model's future expectations will be increased/lowered and it will be more/less likely to select and attack that computer in the future. Also, the impact of a particular past experience on future decisions strengthens with frequency and weakens with time. Finally, experiences are generalized across targets reflecting their attributes (payoff, penalty, probability). Due to the stochastic nature of the model, it was run through 1000 simulations in each of the experimental conditions of the IAG in order to generate stable predictions of human behavior. Of emphasis, is that the model behaves differently on each run and can therefore represent a population of human attackers without the need to parameterize for individual differences. The differences in behavior observed between runs are triggered by the stochastic component in memory retrieval, which underlies the generation of expectations but is especially amplified by the different experiences resulting from that stochasticity, which serves as the basis for future expectations. The model's performance was compared to that of humans.

5. Human experiment

A between-subjects experiment was conducted with human players to examine the effectiveness of the two deceptive signaling conditions, compared to the no signaling condition.

5.1. Method

5.1.1. Participants

One hundred thirteen participants participated in the *peSSE* condition, 108 in the *peSSE-FI* condition, and 107 in the *NoSignal* condition. All participants were recruited via Amazon Mechanical Turk (mTurk) and had a 90% or higher approval rate with at least 100 Human Intelligence Tasks (HITs) approved, resided in the United States, and had not participated in other conditions. For completing the experiment and submitting a completion code, participants were paid \$1 plus \$0.02 per point earned in the game, up to a maximum of \$4.50 in additional bonus pay. It is well known that mTurk participants are driven by maximizing payout, and the low base pay and high potential for bonus meant their potential payout was directly tied to their points earned in the game. This ensured that the defense algorithm, whose aim is to reduce attacker's utility by minimizing their points earned, was aligned with the goals of mTurk participants and thus appropriate for present investigations.

In the *peSSE* condition, 11 participants did not complete the experiment or submit a completion code, two participants restarted the experiment after partially completing it, one participant had incomplete data due to a recording error, and one participant previously participated in another condition (final $N = 98$). In the *peSSE-FI* condition, eight participants did not complete the experiment and four participants had data recording errors (final $N = 96$). In the *NoSignal* condition, six participants did not complete the experiment, one participant restarted the experiment, and five participants had data recording errors (final $N = 96$). These participants were removed from the analysis. Among the final sample in the *peSSE* condition, 56 were male, 41 were female (one participant did not specify), and the mean age was 34.64 (range: 21–68). In the *peSSE-FI* condition, 52 were male, 44 were female, and the mean age was 35.44 (range: 19–65). In the *NoSignal* condition, 57 were male, 39 were female, and the mean age was 35.65 (range: 21–73).

5.1.2. Design

The design was a 3 (signaling scheme: peSSE, peSSE-FI, and *NoSignal*) by 4 (round: 1 through 4) mixed-effects design. The signaling scheme was a between-subjects factor and the round was a within-subjects factor. Each of the three signaling scheme conditions was run separately, at different times.

5.1.3. Procedure

The experiment was conducted via mTurk. The experiment was advertised as "A fun game of decision making to help keep our systems safe!!" Participants clicked the link of one of the experimental conditions and were first presented with a consent form and asked a few demographic questions. After providing informed consent, participants were presented with instructions for how to play the game. Participants were told that they would be taking the role of an employee in a company, and their goal was to steal proprietary information by attacking computers. They could receive points for attacking computers that were not monitored by one of the two defenders, as denoted by the number of yellow stars displayed on the targets (see Fig. 1(a)), but could lose points for attacking computers that were monitored by a defender as denoted by the number of red stars displayed. Participants were informed that they would earn $1 for completing the game and the questionnaire and would earn an additional $0.02 per point accumulated throughout the game up to a maximum of $4.50. After reading the instructions, participants answered a few questions to test their knowledge of how to play the game, and they were provided feedback about the correctness of their answers. After receiving the feedback, they could proceed to the game.

Participants played a practice round of five trials to become familiar with the interface and then played four rounds of the game for 25 trials per round. The targets changed for each round as defined in Table 1. The location of the targets within the display was randomly assigned between participants but did not change within a round. Participants began a round by pressing a "continue" button indicating they were ready to begin. For each trial, participants began by selecting one of the six targets with the click of a mouse. After clicking the target, one of two messages were displayed depending on the coverage and signaling schedule defined for the experimental condition (e.g., see Fig. 1(b)). One message read, "This computer is being monitored! Do you want to access this computer?" if the computer was monitored or presented with a deceptive signal. The other message read, "Do you want to access this computer?" if the computer was not monitored (in the *NoSignal* condition, this message was displayed every time regardless of coverage, and participants were never warned that the computer was being monitored). Participants responded by either clicking a "yes" or a "no" button. If participants responded "yes" and continued the attack, then they received the number of points denoted by the yellow stars if the target was not monitored but lost the number of points denoted by the red stars if the target was monitored. The total points earned in a round are displayed in the top right of the interface. If they responded "no" and withdrew their attack, then they received zero points. In the *peSSE-FI* condition, for the second message above, participants were also told, "X% of time this computer appears as 'monitored' the analyst is NOT actually present." Where X was replaced with the percent of time the signal is deceptive for that target.

After completing 25 trials, participants were provided feedback regarding their score for the round and their cumulative score across rounds. At the end of the fourth round, participants were provided their final score and then pressed an "ok" button to continue to a 10-question, post-game survey. This data was not analyzed and is not further discussed. After completing the survey, participants were thanked for their participation and given a completion code. Participants had to return to the experiment website at Mechanical Turk and enter their completion code to claim their reward. Participants were paid the $1 base rate plus their earned bonuses within 24 h of completing the experiment.

5.2. Human and IBL model results

In the sections below, we first analyze human behavior/performance in the IAG by examining the probability of attack across rounds. This data is then compared to the performance of the IBL model. Next, we examine the number of points earned across rounds, comparing human and model performance. Finally, we also examined the target selection behavior of

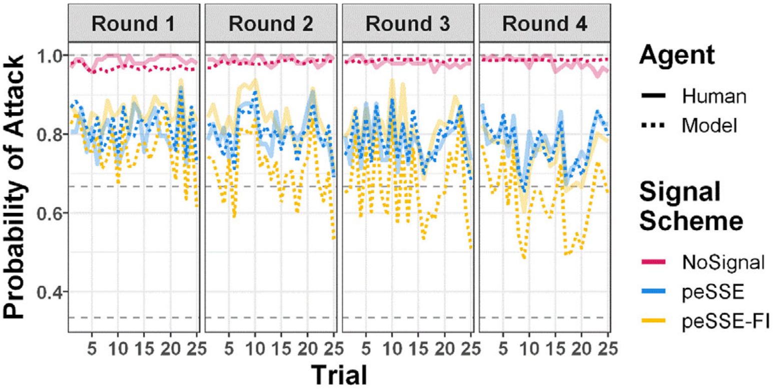


Fig. 4. Probability of attack across trials and rounds for humans, compared to the IBL model. The flat, dashed lines represent the predicted behavior under assumptions of perfect rationality.

humans, compared to the model to shed additional light on how deception influences human decisions in the IAG.

5.2.1. Probability of attack

Fig. 4 shows the mean probability of attack across 25 trials in each of the four rounds. The probability of an attack was calculated as the proportion of participants (or simulated attackers, for the IBL model) that continued the attack on a given trial. The results of the three experimental conditions are compared against three baselines that were built upon the assumption of a perfectly rational player (p.r.p.). In Fig. 4, these baselines are shown as flat, dashed lines over the 25 trials for each of the four rounds. First, because the expected value of attacking each target is positive (ignoring the signal), a p.r.p. is expected to attack every time (pAttack = 1.0). Second, if all signals were truthful and only presented if a target was truly covered, and a p.r.p. never attacks when a signal is present, then a p.r.p. is expected to attack at a rate equal to the average probability a target is not monitored (i.e., pAttack = 0.67; optimal truthful baseline). Third, if signals are sometimes deceptive at a rate determined by the peSSE, and a p.r.p. never attacks when a signal is present, then a p.r.p is expected to attack at a rate equal to the average probability a signal is not presented (pAttack = 0.33; optimal deceptive baseline). In the *NoSignal* condition, like p.r.p.'s, humans attack almost all of the time because the expected value of attacking is positive for all targets. In the *peSSE* and *peSSE-FI* conditions, although signaling reduces attacks, humans attack far more often than predicted by a p.r.p. and also more than a p.r.p. would if all signals were truthful.

An examination of the human data reveals that participants attacked on an average of 79.4% (*SD* = 24.7%) trials in the *peSSE* condition, 81.1% (*SD* = 22.5%) trials in the *peSSE-FI* condition, and 98.4% (*SD* = 5.1%) trials in the *NoSignal* condition. A mixed-factors ANOVA revealed the probability of an attack was significantly different between signaling schemes, $F(2, 287) = 28.04$, $p < .001$, being lower in the *peSSE* and *peSSE-FI* conditions, compared to the *NoSignal* condition, both ts > 7.35, ps $< .001$. There were no differences between

Table 2
RMSE and correlations between human and model data for each signaling scheme and round

	Round 1		**Round 2**		**Round 3**		**Round 4**		**Total**	
Signaling Scheme	**RMSE**	**Corr.**	**RMSE**	**Corr.**	**RMSE**	**Corr.**	**RMSE**	**Corr.**	**RMSE**	**Corr.**
peSSE	0.06	0.32	0.04	0.73	0.03	0.81	0.03	0.87	0.04	0.72
*peSSE**	0.05	0.53	0.04	0.85	0.04	0.86	0.04	0.93	0.04	0.80
peSSE-FI	0.10	0.54	0.13	0.87	0.15	0.84	0.12	0.92	0.13	0.85
NoSignal	0.02	−0.01	0.01	0.00	0.01	−0.28	0.02	−0.19	0.02	−0.19

Abbreviations: *peSSE*, Strong Stackelberg Equilibrium with persuasion; *peSSE-FI*, peSSE-full information (FI).

*This compares the instance-based learning model in the *peSSE* condition to the humans in the *peSSE-FI* condition.

peSSE and *peSSE-FI* conditions. In addition, humans showed no evidence of learning as the probability of attack was quite stable across rounds. In the *NoSignal* condition, the probability of attack remains near ceiling across rounds, all ts < 0.59, ps $> .555$. In the *peSSE* and *peSSE-FI* condition, there is a slight downward trend across rounds, but no significant differences were detected, all ts < 1.66, ps $> .098$.

As can be used seen in Fig. 4, the IBL model accounts very well for the overall probability of attack in the *peSSE* and *NoSignal* conditions, as well as the trial-by-trial fluctuations in the *peSSE* condition. Table 2 shows the Root Mean Square Error (RMSE) and correlations between the human data and IBL model data for each signaling scheme and round. In the *peSSE*, the RMSEs across rounds are very low (< 0.06), and the overall correlation is very high (0.72). The IBL model is less accurate in Round 1 but quickly aligns with human performance by Round 2. Importantly, the IBL model in the *peSSE* condition is more accurate of humans in the *peSSE-FI* condition (max RMSE = 0.05) than the IBL model in the *peSSE-FI* condition (min RMSE = 0.10), which underestimates the probability of attack. This suggests humans do not know how to make use of the probability in the full information condition. In the *NoSignal* condition, the total RMSE is very low (0.02), but the correlation is small and negative (−0.19). The fluctuations across trials can be mainly attributed to the schedule of signaling.

Not only does the model match well with the mean probability of attack across trials, it also accounts well for the full range of human behavior as can be seen in the histograms in Fig. 5. The histograms show the distribution of participants (or model simulations) by the mean probability of attack. In the *NoSignal* condition, the model shows a similar distribution as humans, with approximately 60% of simulations attacking 100% of trials and about 25% attacking between 95% and 99% of trials. In the *peSSE* and *peSSE-FI* conditions, like humans, a large proportion of model simulations attacked on more than 95% of trials, while the distribution tails off to about 20% at minimum. As can be seen in the bottom right panel of Fig. 5, the model in the *peSSE* condition matches better with the humans in the *peSSE-FI* condition than the model in the *peSSE-FI* condition, another indicator of human's lack of use of the additional probability information.

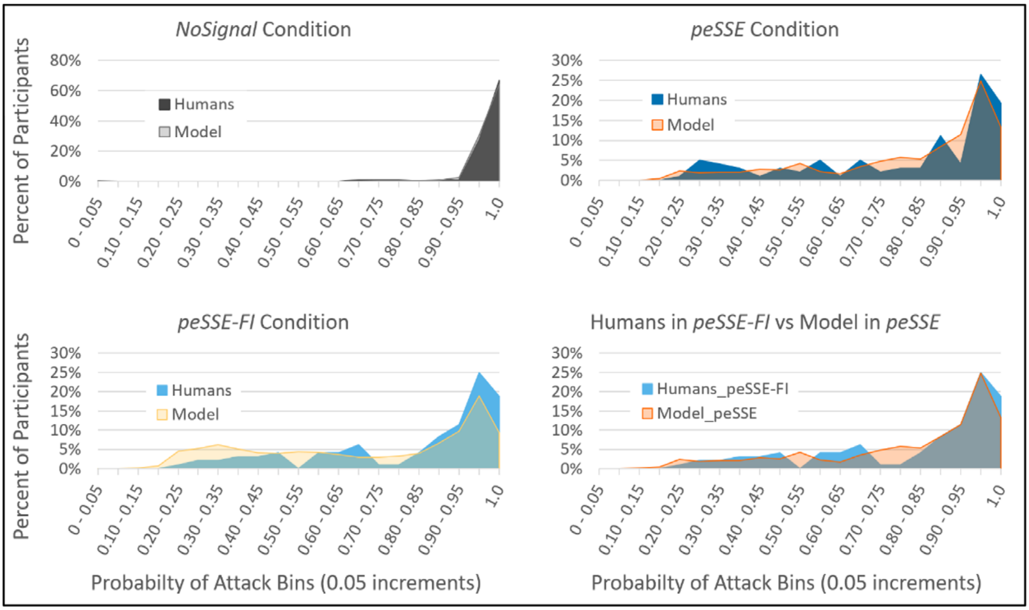


Fig. 5. Histogram of the mean probabilities of attack for humans compared to the IBL model in the *NoSignal*, *peSSE*, and *peSSE-FI* conditions. The bottom right panel shows that the model in the *peSSE* condition matches better with the humans in the *peSSE-FI* condition than the model in the *peSSE-FI* condition.

If we look more closely at the effects of signaling, the results show that when presented with a signal, humans (and, likewise, the IBL model in the *peSSE* condition) tend to continue the attack more often than withdraw. Fig. 6, panel A, shows the probability of attack when a signal is present. As before, the model in the *peSSE* condition matches well to humans in the *peSSE* and *peSSE-FI* conditions, while the model in the *peSSE-FI* condition underpredicts the probability of attack. While a p.r.p is predicted to never attack when a signal is present, humans attack much more often, near 75% probability. This is also more than the average probability of a loss following an attack given a signal (or 50%), indicating a bias toward attacking given a signal. In fact, we examined sensitivity (d') for detecting whether a signal is truthful or deceptive, and results show that humans in the *peSSE* and *peSSE-FI* conditions have an almost zero sensitivity (d' $= -0.36, -0.24$, respectively), which makes sense given the probability of attacking given a signal when a target is not covered ($M = 0.71, 0.72$, respectively) is almost equal to the probability of attack given a signal when a target is covered ($M = 0.72, 0.73$, respectively). Additionally, response bias scores (C) indicate a high bias toward attacking given a signal ($C = -2.12, -1.82$, respectively). As expected, these results indicate that humans are insensitive to detecting whether a signal is deceptive in the present task but also reveal biases to attack given a signal. Compared to humans, the cognitive model in the *peSSE* condition displayed similar sensitivity (d' $= -0.12$) and bias scores ($C = -1.56$) as humans in either condition. However, in the *peSSE-FI* condition, the model was similarly insensitive (d' $= -0.27$) but exhibited less bias toward attacking ($C = -0.41$), as reflected in the comparatively lower probability of attack seen in Fig. 4.

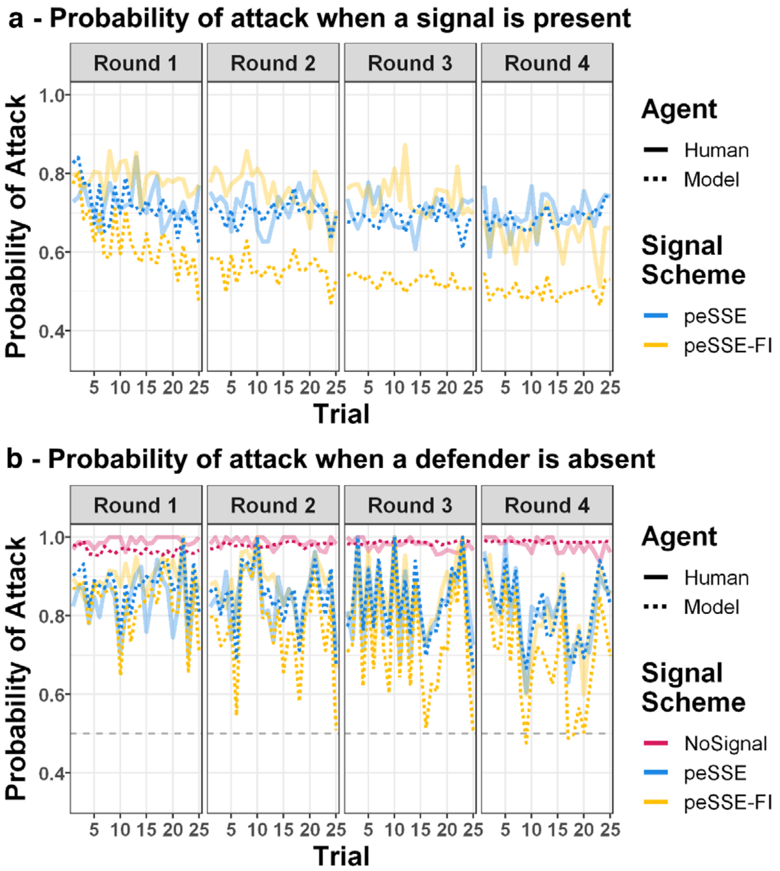


Fig. 6. Probability of attack for humans, compared to the IBL model when a signal is present (panel A) and when a defender is not monitoring the target (panel B). For panel A, the prediction of a perfectly rational player is not shown because it is at 0.0.

Ultimately, a primary goal of the peSSE is to reduce attacks on uncovered targets. Fig. 6, panel B, shows the mean probability of attack per round when a defender is not monitoring the selected computer. In the *peSSE*, because deceptive signals are presented, an average of 50% of trials when a defender is absent, a p.r.p. is expected to attack 50% of the time (i.e., every time a signal is *not* presented). Humans attack uncovered targets less often when deceptive signals are presented (i.e., in both the *peSSE* and *peSSE-FI* conditions), compared to never signaling but at a much higher rate (~80%) than was predicted by a p.r.p. Once again, we see that the model in the *peSSE* condition accurately predicts human behavior in the *peSSE* and *peSSE-FI* conditions better than the model in the *peSSE-FI* condition.

5.2.2. *Points earned*

There were no differences in the total points earned between signaling schemes (analyses can be found in the online Supplemental Material at https://osf.io/mfb4q); however, important

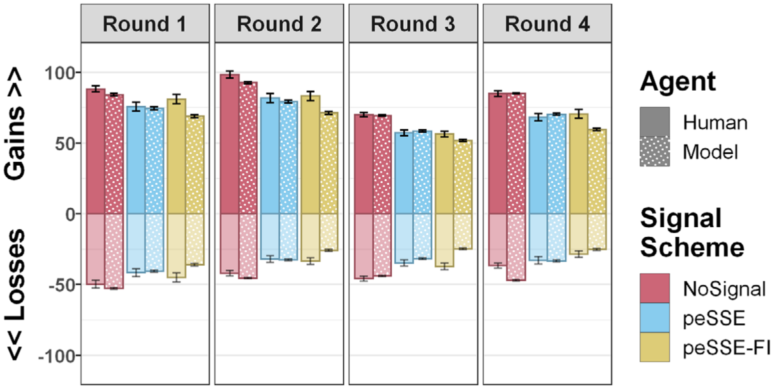


Fig. 7. Mean gains/losses per round for humans, compared to the IBL model in the *NoSignal, peSSE*, and *peSSE-FI* conditions.

differences are observed when points are separated into mean gains and losses. Gains are points earned from attacking a target when a defender is absent and is complementary to the analysis of attacks on uncovered targets, while losses are points lost from attacking when a defender is present. If deceptive signals are effective at influencing human behavior, then gains should be reduced, compared to not signaling. However, any losses indicate behavior that is inconsistent with a p.r.p. who would always withdraw attacks in the presence of a signal. Fig. 7 shows the mean points obtained per round for humans and the IBL model. For humans, a mixed-factors ANOVA revealed the mean gains were significantly different between signaling schemes, $F(2, 287) = 13.30$, $p < .001$, being lower in the *peSSE* and *peSSE-FI* conditions, compared to the *NoSignal* condition, both *t*s > 4.26, *p*s < .001. There were no differences between *peSSE* and *peSSE-FI* conditions. While deceptive signals reduce gains, compared to not signaling, participants still suffer losses in the *peSSE* and *peSSE-FI* conditions, although fewer than in the *NoSignaling* condition, indicating that, while effective, the peSSE signaling scheme could be improved.

5.2.3. Selection preferences

Finally, in addition to probabilities of attacking and points gained, the IBL model accounts very well for the selection behavior of humans. Fig. 8 shows the probability to attack a target, as a function of the probability of being selected, across rounds. The light bars represent the probability of selecting that target, and the dark bars represent the probability of attacking that target given it was selected. Targets are ordered by m-prob and then reward. In Round 1, both humans and the IBL model tend to select the higher reward targets more often, unless the m-prob is high. However, humans tend to select the moderate targets (i.e., reward = 5; m-prob = 0.36) more often than the model. In Round 2, both humans and the model tend to select the target with a higher reward and moderate m-prob more often. In Round 3, this trend continues and both humans and the model once again tend to select the option with a moderate m-prob more often. However, unlike the model, humans also tend to select the target that has a very large difference between the reward and penalty even though it also has a very high m-prob.

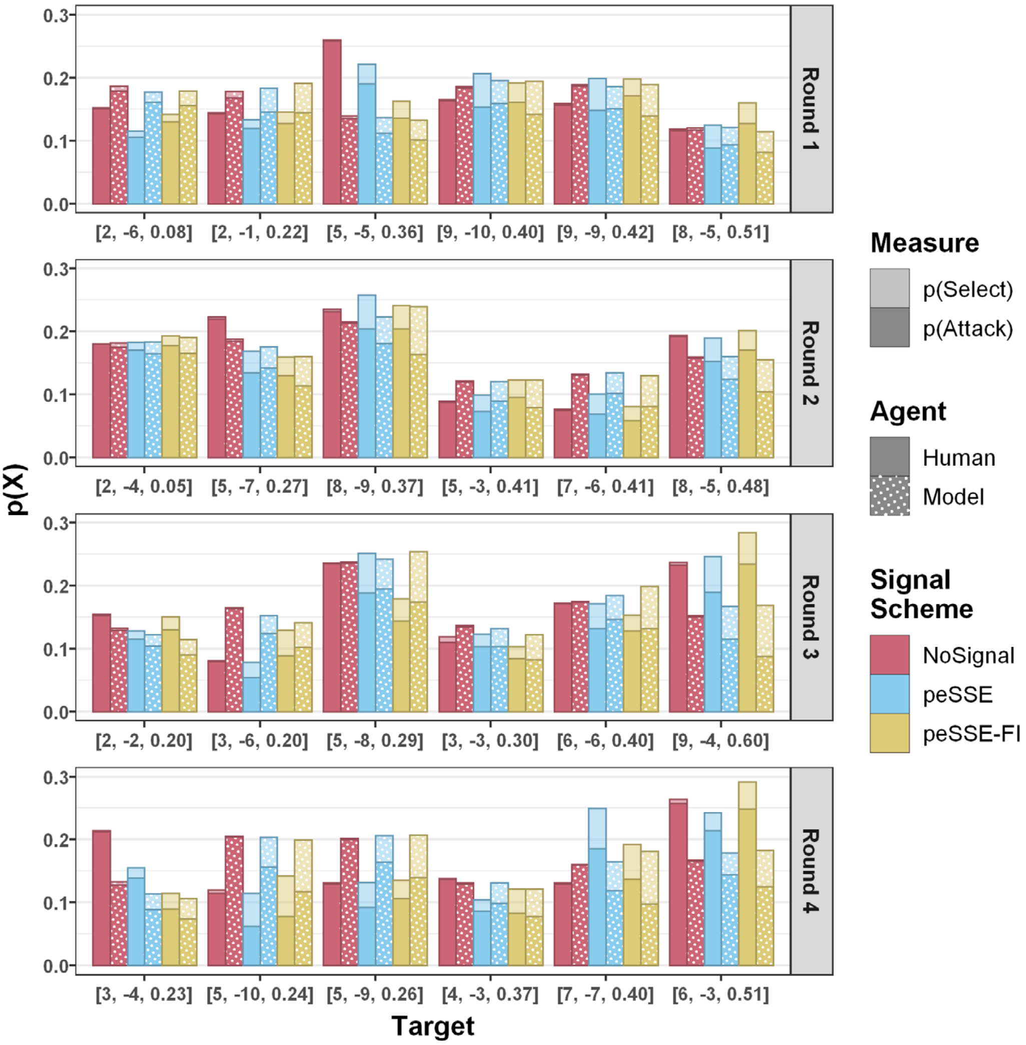


Fig. 8. Probability of attacking a target as a function of the probability of selecting a target, across rounds, for humans, compared the IBL model in the *NoSignal*, *peSSE*, and *peSSE-FI* conditions. Targets are labeled with their respective attribute values (reward, penalty, monitoring probability).

This trend is similar in Round 4, where humans prefer the target with the largest difference between reward and penalty, while the model tends to select targets with lower m-probs and moderate rewards.

6. Discussion

Compared to not signaling at all, the use of deceptive signals is effective in reducing the number of attacks on uncovered targets. However, the peSSE does not perform as well as predicted under the assumption of perfect rationality. Instead, when faced with a signal, rather than always withdrawing, humans tend to attack more often than not. Through instructions and experience, humans learn that the signal is sometimes deceptive and that attacking will sometimes result in a reward. These experiences, in turn, lead to continued attacks in the future, which sometimes result in a reward, perpetuating the probability of attacking. The IBL model sheds additional light on human behavior in the IAG.

The IBL model provides highly accurate predictions of human behavior, even capturing the trial-by-trial fluctuations that reflect the combination of coverage and signaling schedules and target selection preferences. Because the target selection distribution is similar to humans, the model experiences similar patterns of signaling and coverage that influence the probability of attack over trials. However, the model does not capture the trial-by-trial fluctuations in the *NoSignal* condition, indicating that the fluctuations are largely driven by the signals themselves.

Deceptive signals influence behavior, although the degree differs across individuals. Some individuals are more compliant with the signal and attack less, while almost half of the participants attacked greater than 95% of the time overall. As a general process model, the IBL model also produces the full range of human behavior without the need for knowledge engineering or parameter fitting. That is, the same model produces a range of behavior, from those that comply with the signal and make decisions that look perfectly rational to those that always attack and make satisficing decisions. The model accounts for these behaviors in an emergent way, highlighting the importance experience has on decision-making. Thus, while human decision-making in the IAG reflects the statistics of the environment, according to the IBL model, it is largely influenced by at least three factors: (1) memory retrieval dynamics across past experiences, (2) confirmation bias, and (3) representation of features of the decision.

The peSSE performs worse than expected because human biases (e.g., recency, frequency, and confirmation) lead to overweighting of positive outcomes and, in turn, expectations greater than zero. When past experiences of positive outcomes are more recent, or more frequent, then positive outcomes are more likely to be expected (likewise for past experiences of negative outcomes). In general, 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. Only for some did the expectation of a loss given a signal outweigh the expectations of a reward and persist throughout the game.

Human behavior is not solely driven by recency and frequency of past experiences, or else the probability of attack given a signal would more accurately reflect the statistics of the environment and be near 50% (i.e., the average probability of a win). However, humans attacked almost 75%. According to the IBL model, this tendency to attack can be explained by retrieving past instances that include memories for prior expectations in addition to the ground-truth outcome. This memory phenomenon manifests as a form of confirmation bias, where decisions are influenced by one's expectations (Gonzalez et al., 2003). The original IBL model (Cranford et al., 2018) only stored the ground-truth outcome and manually kept the expectations from being saved to memory and attacked nearly 50% of trials given a signal. The current model simply allows the expectations to be saved to memory as the architecture intended. The effect is that the additional instances alter the availability of information in memory. More specifically, the net effect on the next trial is the average of the expectation and the actual outcome instead of just the actual outcome. For example, when expectations are not stored, a negative experience would drive down future expectations and likely lead to a subsequent withdrawal action. However, when expectations are stored, the positive expectation will temper the impact of negative experiences on future expectations, and the model will be more likely to persist in attacking. The IBL model indicates that confirmation bias emerges from storing prior expectations in memory and explains why many humans continue to attack in the face of a signal (and also why some continue to never attack in the face of a signal).

Finally, the IBL model indicates that the representation of features of the decision is important. For example, in the original model in Cranford et al. (2018), the context of the selected target was represented in the decision to attack or withdraw, but this resulted in a much lower probability of attack on targets with higher m-probs because it served as an indication that the outcome would more likely be a loss. The current IBL model and humans show a much more equally distributed probability of attack across targets. Because the signal message occludes the targets, unless humans explicitly maintain the target information in working memory, it is likely not represented in the context of the decision.

Another example of the importance of accurately representing the features of the decision is that, in the *peSSE-FI* condition, providing humans with information regarding the deception probability does not further reduce attacks. The additional information does not change human behavior, and when the model does not consider that information (i.e., in the *peSSE* condition), it more accurately predicts human behavior in the *peSSE-FI*. This is an indication that participants do not consider, or otherwise know how to use, this information in making decisions.

7. General discussion

In the cyber domain, it is often very difficult to detect deception due to the lack of physical cues available (Riggio & Friedman, 1983). Using cognitive models and human experiments, the present study sought a better understanding of how humans make decisions when faced with deceptive signals in an insider attack scenario. The results showed that human behavior in the IAG is largely consistent with IBLT (Gonzalez et al., 2003). Human decisions in such dynamic and uncertain environments are made through the aggregated retrieval over past decisions based on the similarity of the current context to past instances, recency, and frequency. Human behavior in the IAG is a result of innate and learned biases, memory dynamics, and interaction with the environment. Decisions are influenced by individual experiences playing out over time. The IBL model accounts for these biases and shows that a major source of bias in the IAG is confirmation bias.

Without any overt cues to detect deception, humans must rely on exploration for detection in the cyber world. In the IAG, this means attacking when a signal is presented. However, once the possibility of deception is known, human decisions are made under uncertainty and born from experience. The presentation of a deceptive signal can weaken future belief in the signal if humans attack due to the experience of a positive outcome. However, attacks on truthful signals can rebuild that belief. For a majority of participants, those that attacked greater than 95% of the time, rebuilding that belief was difficult, if not impossible, because the effects of a positive outcome can persist long enough, through confirmation bias, until another positive outcome is experienced to reinforce their behavior. For cybersecurity, it is important to understand these human biases to construct effective defenses that can account for human bounded rationality and find the optimal rate of signaling (Cooney et al., 2019).

Adversaries will tend to trust signals that are very rarely deceptive. However, the defender gains very little from the use of deception in such cases. Meanwhile, using too much deception will erode the meaning of the signal and the adversary may ignore it. Deceptive signaling schemes have been optimized for an environment that assumes attackers are perfectly rational. However, the present results show that deceptive signals must be optimized for humans with imperfect memories that make boundedly rational decisions. Research is currently underway to find the optimal combination of bluffing and truth-telling that minimizes attacks on uncovered targets for boundedly rational humans. For example, we have developed game-theoretic models of signaling, and also of masking, that use machine learning methods to predict the likelihood of an attack in particular situations using real human data to inform the design of the defense strategy (Aggarwal et al., 2020; Cooney et al., 2019; Thakoor et al., 2020). Meanwhile, in other research, we have developed techniques that use the present IBL model to predict human behavior in real time to adapt the signaling scheme to the individual user (Cranford et al., 2020a, 2020b). These preliminary approaches that account for bounded rationality show slight improvements over game-theoretic models that assume perfect rationality.

One limitation of the peSSE is that it is static, while humans, like other animals, learn to adapt to signals through repeated experiences, and these experiences are unique to each individual (Eliason, 2018). While deception is an effective tool for preventing malicious behaviors, the experience of successfully calling a bluff can reduce compliance in the signal. Regaining trust in the signal is difficult if not impossible to do under static signaling schemes. Unlike statistical, or machine learning, models of attackers that explain the statistics of the environment and the probability of making decisions in particular situations and that rely on large amounts of data to make accurate predictions of human decisions, the IBL model is a behavior generative model that helps explain human behavior response to deceptive signals. It could therefore serve an applied role in predicting human behavior to aid the development of alternative signaling schemes. In fact, it has been used to accurately predict human attacker behavior against several other signaling schemes not reported here. In recent research, we have explored the possibility of using the cognitive model to adapt the signaling scheme online to be more effective against an individual attacker (Cranford et al., 2020a, 2020b). In Cranford et al. (2020a, 2020b), the IBL model is used to trace human behavior, make predictions about the probability of attack given a signal, and then adjust the signaling scheme to present signals at a rate that maintains a belief in the signal. Personalized signaling schemes that use a cognitive model to make predictions about human behavior can greatly improve security defenses by taking into account an individual attacker's history of experience and cognitive biases to better understand their preferences and tendencies. This research is in early development and future research is aimed at improving existing methods and exploring alternative methods that use the IBL model as a tool to adapt the signaling scheme to an individual to increase compliance with the signal.

The IBL model also highlights the importance of the representation of information in the context of the decision. Unlike a perfectly rational adversary, humans do not consider all available information, as was evident in the *peSSE-FI* condition. For example, in another experiment with the IAG, about 45% of participants did not seem to represent the signal in their decision, and a model that also did not represent the signal matched well to human data (Cranford et al., 2020b). In addition to predicting attack probabilities, an adaptive cognitive model could learn what features are important for an individual's decision-making (see Martin, Lebiere, Fields, & Lennon, 2018) to further improve predictions by adapting the model's representation of instances to match the individual. Concurrent research is computationally investigating what are the salient features of IBL decisions, and initial results prove the methodology useful and highlight the individual differences in what information humans use to make their decisions (Cranford, Somers, Mitsopoulos, & Lebiere, 2020; Somers, Mitsopoulos, Lebiere, & Thomson, 2019).

One concern of the present research is how well the cognitive model and our findings might generalize to real-world settings. For example, an assumption of the paradigm is that attackers make repeated decisions and learn from experience. However, in reality, it may be the case that individuals make few attacking attempts or that a series of attacks could come from multiple individuals. Fortunately, the IBL model could be used to predict the behavior of a group of attackers, or the average behavior across a window of time, and we still recognize the advantage of tools that protect against repeated attackers. Another concern is how well the model predictions would hold in real-world situations where the payoffs and costs are potentially much higher. However, we argue that IBL models are well-suited to account for any possible reward structure precisely because they base decisions by computing expected utilities of the potential options. To address these and other issues of ecological validity, we are currently investigating deceptive techniques in the more realistic Cyber Security Virtual Assured Network (CyberVAN) testbed using domain experts (Aggarwal et al., 2020).

In addition to making the enemy think we are near when we are far, Sun Tzu asserts we must also make the enemy think we are far when we are near. Another limitation of the peSSE is that it only uses deception when a target is unmonitored. When the signal is absent, players may attack with impunity. Therefore, research has begun to investigate the possible benefits of a defense signaling scheme that adds deception when a target is monitored by sometimes refraining from sending a signal (Cooney et al., 2019; Cranford et al., 2020a). If signals were always truthful, attackers would likely always comply because there is no evidence that attacking given a signal would result in a reward. However, using deception only when a target is unmonitored does not greatly improve defenses because humans are not perfectly rational and, once they know a signal is sometimes deceptive, will lose belief in the signal and (at least sometimes) attack. While attacks on unmonitored targets are reduced, they are not eliminated. By adding deception when a target is monitored, a signaling scheme could create further uncertainty in the mind of the attacker, and just a little uncertainty could add disproportional benefits in the rate of compliance with a signal.

Our research suggests that in the presence of limited security resources, defenders could use signaling to create uncertainty in the attacker's decisions. The benefits of deceptive signaling have been proven effective in physical security such as LA airport security and poaching (Tambe, 2011) and our research indicates the methods can be applied to the design of defense strategies for cybersecurity. Furthermore, our research provides additional insights about human decision-making processes under deception, emphasizing the importance of models that consider human bounded rationality. Although the IAG is a simple abstraction of naturalistic cyber scenarios, the theoretical developments of our architecture and insights from our cognitive model can be scaled up to complex systems that require greater expertise. Therefore, future research will further uncover how humans react to deception in the cyber domain and inform the design of more effective cyber defenses, scaling in both realism and complexity.

Note

1. In practice, this value is epsilon lower than zero so that the expected value of attacking a target given a signal remains slightly negative.

Acknowledgment

This research was supported by the Army Research Office and accomplished under grant number W911NF-17-1-0370.

References

Abbasi, Y. D., Ben-Asher, N., Gonzalez, C., Kar, D., Morrison, D., Sintov, N., & Tambe, M. (2016). Know your adversary: Insights for a better adversarial behavioral model. In A. Papafragou, Daniel J. Grodner, D. Mirman & J. Trueswell (Eds.), *Proceeding of the 38th annual conference of cognitive science society* (pp. 1391–1396). Austin, TX: Cognitive Science Society.

Aggarwal, P., Gonzalez, C., & Dutt, V. (2016). Cyber-security: Role of deception in cyber-attack detection. In D. Nicholson (Ed.), *Advances in human factors in cybersecurity* (Vol. *501*, pp. 85–96). Cham: Springer. https://doi.org/10.1007/978-3-319-41932-9_8

Aggarwal, P., Thakoor, O., Mate, A., Tambe, M., Cranford, E. A., Lebiere, C., & Gonzalez, C. (2020). An exploratory study of a masking strategy of cyberdeception using CyberVAN. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 446–450. https://doi.org/10.1177/1071181320641100

Almeshekah, M. H., & Spafford, E. H. (2016). Cyber security deception. In S. Jajodia, V. Subrahmanian, V. Swarup & C. Wang (Eds.), *Cyber deception* (pp. 25–52). Cham: Springer. https://doi.org/10.1007/978-3-319-32699-3_2

Al-Shaer, E., Wei, J., Hamlen, K. W., & Wang, C. (Eds.). (2019). *Autonomous cyber deception: Reasoning, adaptive planning, and evaluation of honey things*. Cham: Springer. https://doi.org/10.1007/978-3-030-02110-8

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–517. https://doi.org/10.1017/S0140525X00070801

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781315805696

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. https://doi.org/10.1037/0033-295X.111.4.1036

Battigalli, P. (2006). Rationalization in signaling games: Theory and applications. *International Game Theory Review*, *8*(01), 67–93. https://doi.org/10.2139/ssrn.635244

Bond, C. F. Jr. & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*(4), 477–492. https://doi.org/10.1037/0033-2909.134.4.477

Cho, I. -K., & Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, *102*(2), 179–221. https://doi.org/10.2307/1885060

Cooney, S., Wang, K., Bondi, E., Nguyen, T., Vayanos, P., Winetrobe, H., Cranford, E. A., Gonzalez, C., Lebiere, C. & Tambe, M. (2019). Learning to signal in the Goldilocks Zone: Improving adversary compliance in security games. In Brefeld U., Fromont E., Hotho A., Knobbe A., Maathuis M., Robardet C. (Eds.), *Machine learning and knowledge discovery in databases* (pp. 725–740). Cham: Springer. https://doi.org/10.1007/978-3-030-46150-8_42

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., & Lebiere, C. (2020a). Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science*, *12*(3), 992–1011. https://doi.org/10.1111/tops.12513

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., & Lebiere, C. (2020b). Adaptive cyber deception: Cognitively informed signaling for cyber defense. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Maui, HI, USA (pp. 1885–1894). https://doi.org/10.24251/HICSS.2020.232

Cranford, E. A., Lebiere, C., Gonzalez, C., Cooney, S., Vayanos, P., & Tambe, M. (2018). Learning about cyber deception through simulations: Predictions of human decision making with deceptive signals in Stackelberg Security Games. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, Madison, WI, USA (pp. 258–263).

Cranford, E. A., Somers, S., Mitsopoulos, K., & Lebiere, C. (2020). Cognitive salience of features in cyber-attacker decision making. In T. C. Stewart (Ed.), *Proceedings of the 18th annual meeting of the international conference on cognitive modeling*. University Park, PA: Applied Cognitive Science Lab, Penn State.

Ekroll, V., & Wagemans, J. (2016). Conjuring deceptions: Fooling the eye or fooling the mind? *Trends in Cognitive Sciences*, *20*(7), 486–489. https://doi.org/10.1016/j.tics.2016.04.006

Eliason, C. M. (2018). How do complex animal signals evolve? *PLoS Biology*, *16*(12), e3000093. https://doi.org/10.1371/journal.pbio.3000093

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

Gonzalez, C., Aggarwal, P., Cranford, E. A., & Lebiere, C. (2020). Design of dynamic and personalized deception: A research framework and new insights. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Maui, HI, USA (pp. 1825–1834). https://doi.org/10.24251/HICSS.2020.226

Gonzalez, C., Ben-Asher, N., Martin, J. M., & Dutt, V. (2015). A cognitive model of dynamic cooperation with varied inter-dependency information. *Cognitive Science*, *39*(3), 457–495. https://doi.org/10.1111/cogs.12170

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

Gonzalez, C., & Lebiere, C. (2005). Instance-based cognitive models of decision making. In D. Zizzo & A. Courakis (Eds.), *Transfer of knowledge in economic decision-making*. New York: Macmillan (Palgrave Macmillan.

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

Hertwig, R. (2015). Decisions from experience. In G. Keren & G. Wu (Eds.), Blackwell handbook of judgment and decision making *(pp. 239–267).* Chichester: Wiley-Blackwell. https://doi.org/10.1002/9781118468333.ch8

Hyman, R. (1989). The psychology of deception. *Annual Review of Psychology*, *40*, 133–154. https://doi.org/10.1146/annurev.ps.40.020189.001025

Jenkins, A., Zhu, L., & Hsu, M. (2016). Cognitive neuroscience of honesty and deception: A signaling framework. *Current Opinion in Behavioral Sciences*, *11*, 130–137. https://doi.org/10.1016/j.cobeha.2016.09.005

Juvina, I., Saleem, M., Martin, J. M., Gonzalez, C., & Lebiere, C. (2013). Reciprocal trust mediates deep transfer of learning between games of strategic interaction. *Organizational Behavior and Human Decision Processes*, *120*(2), 206–215. https://doi.org/10.1016/j.obhdp.2012.09.004

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

Lebiere, C., Gonzalez, C., & Martin, M. (2007). Instance-based decision making model of repeated binary choice. *Proceedings of the Eighth International Conference on Cognitive Modeling*, Ann Harbor, MI, USA (pp. 67–72). https://doi.org/10.1184/R1/6571190.v1

Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., & Anderson, J. R. (2013). A Functional model of sensemaking in a neurocognitive architecture. *Computational Intelligence and Neuroscience*, *2013*, 921695. http://doi.org/10.1155/2013/921695

Lebiere, C., Wallach, D., & West, R. L. (2000). A memory-based account of the prisoner's dilemma and other 2x2 games. *Proceedings of International Conference on Cognitive Modeling*, Groningen, the Netherlands (pp. 185–193).

Martin, M., Lebiere, C., Fields, M. A., & Lennon, C. (2018). Learning features while learning to classify: A cognitive model for autonomous systems. *Computational and Mathematical Organization Theory*, *26*, 23–54. https://doi.org/10.1007/s10588-018-9279-3

Moisan, F., & Gonzalez, C. (2017). Security under uncertainty: Adaptive attackers are more challenging to human defenders than random attackers. *Frontiers in Psychology*, *8*, 982. https://doi.org/10.3389/fpsyg.2017.00982

Mokkonen, M., & Lindstedt, C. (2016). The evolutionary ecology of deception. *Biological Reviews*, *91*(4), 1020–1035. https://doi.org/10.1111/brv.12208

Morgan, C. J., LeSage, J. B., & Kosslyn, S. M. (2009). Types of deception revealed by individual differences in cognitive abilities. *Social Neuroscience*, *4*(6), 554–569. https://doi.org/10.1080/17470910802299987

Morgulev, E., Azar, O. H., Lidor, R., Sabag, E., & Bar-Eli, M. (2014). Deception and decision making in professional basketball: Is it beneficial to flop? *Journal of Economic Behavior & Organization*, *102*, 108–118. https://doi.org/10.1016/j.jebo.2014.03.022

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Pawlick, J., Colbert, E., & Zhu, Q. (2019). A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Computing Surveys*, *52*(4), 1–28. https://doi.org/10.1145/3337772

Pita, J., Jain, M., Ordónez, F., Portway, C., Tambe, M., Western, C., & Kraus, S. (2008). ARMOR security for Los Angeles International Airport. *Proceeding of the Twenty-Third AAAI Conference on Artificial Intelligence*, Chicago, IL (pp. 1884–1885).

Riggio, R. E., & Friedman, H. S. (1983). Individual differences and cues to deception. *Journal of Personality and Social Psychology*, *45*(4), 899–915. https://doi.org/10.1037/0022-3514.45.4.899

Rowe, N. C., & Rrushi, J. (2016). *Introduction to cyberdeception*. Cham: Springer. https://doi.org/10.1007/978-3-319-41187-3

Sanner, S., Anderson, J. R., Lebiere, C., & Lovett, M. C. (2000). Achieving efficient and cognitively plausible learning in Backgammon. In P. Langley (Ed.), *Proceedings of the seventeenth international conference on machine learning*. San Francisco: Morgan Kaufmann. https://doi.org/10.1184/R1/6613298.v1

Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., & Meyer, G. (2012). PROTECT: A deployed game theoretic system to protect the ports of the United States. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Valencia, Spain (pp. 13–20).

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Sinha, A., Fang, F., An, B., Kiekintveld, C., & Tambe, M. (2018). Stackelberg Security Games: Looking beyond a decade of success. *Proceedings of the 27th international joint conference on artificial intelligence*, Stockholm, Sweden (pp. 5494–5501). https://doi.org/10.24963/ijcai.2018/775

Somers, S., Mitsopoulos, K., Lebiere, C., & Thomson, R. (2019). Cognitive-level salience for explainable artificial intelligence. *Proceedings of the 17th Annual Meeting of the International Conference on Cognitive Modeling*, Montreal, Quebec, Canada.

Stech, F. J., Heckman, K. E., & Strom, B. E. (2016). Integrating cyber-D&D into adversary modeling for active cyber defense. In S. Jajodia, V. Subrahmanian, V. Swarup & C. Wang (Eds.), *Cyber deception* (pp. 1–22). Cham: Springer. https://doi.org/10.1007/978-3-319-32699-3_1

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

Thakoor, O., Jabbari, S., Aggarwal, P., Gonzalez, C., Tambe, M., & Vayanos, P. (2020). Exploiting bounded rationality in risk-based cyber camouflage games. In Q. Zhu, J. S. Baras, R. Poovendran & J. Chen (Eds.), *Decision and game theory for security. GameSec 2020. Lecture Notes in Computer Science, 12513* (pp. 103–124). Cham: Springer. https://doi.org/10.1007/978-3-030-64793-3_6

West, R. L., & Lebiere, C. (2001). Simple games as dynamic, coupled systems: Randomness and other emergent properties. *Journal of Cognitive Systems Research*, *1*(4), 221–239. https://doi.org/10.1016/S1389-0417(00)00014-0

Xu, H., Rabinovich, Z., Dughmi, S., & Tambe, M. (2015). Exploring information asymmetry in two-stage security games. *Proceedings of the National Conference on Artificial Intelligence*, Austin, TX, USA (pp. 1057–1063).