What Attackers Know and What They Have to Lose: Framing Effects on Cyber-attacker Decision Making

Edward A. Cranford[1], Cleotilde Gonzalez[1], Palvi Aggarwal[1], Milind Tambe[2], and Christian Lebiere[1]
[1]Carnegie Mellon University, [2]Harvard University

Many cybersecurity algorithms assume adversaries make perfectly rational decisions. However, human decisions are only boundedly rational and, according to Instance-Based Learning Theory, are based on the similarity of the present contextual features to past experiences. More must be understood about what available features are represented in the decision and how outcomes are evaluated. To these ends, we examined human behavior in a cybersecurity game designed to simulate an insider attack scenario. In a human-subjects experiment, we manipulated the information made available to participants (concealed or revealed decision probabilities) and the framing of the outcome (as losses or not). An endowment was given to frame negative outcomes as losses, but these were not framed as losses when no endowment was given. The results reveal differences in behavior when some information is concealed, but the framing of outcomes only affects behavior when all information is available. A cognitive model was developed to help understand the cognitive representation of these features and the implications of the behavioral results.

INTRODUCTION

In many applications of cybersecurity, the underlying algorithms are often based on game-theoretical formulations that assume the adversary makes perfectly rational decisions. However, it is well known that humans behave far differently than predicted under these assumptions (Tversky & Kahneman, 1974). Human cognition operates on imperfect memories and is limited in capacity, which can lead to seemingly irrational decisions that are based on prior experiences (Gonzalez, Lerch, & Lebiere, 2003; Gonzalez, 2013). Memory retrieval processes that operate on these experiences can lead to biases which further influence decisions (e.g., Cranford et al., 2019, 2020; Lebiere et al., 2013). Because experiences are encoded along dimensions of the features of the situation, it is important to understand how humans represent information in memory in order to make more accurate predictions of human behavior and ultimately inform better defenses for cybersecurity. The present research explores how the representation of contextual features and outcomes influence human decisions.

Humans make decisions based on the information available to them and each piece of information has an influence on the eventual decision (Lebiere, 1999; Martin, Lebiere, Fields, & Lennon, 2018). However, humans also have limitations on the amount of information they can process at once and tend to rely on heuristics and other processing shortcuts to solve problems and make decisions (Tversky & Kahneman, 1974; Gigerenzer & Todd, 1999). In fact, one commonly recommended strategy for cyber defense is to overload adversaries with information, which can lead to decision errors (Rowe & Rushi, 2016). Defenses that assume the adversary makes perfectly rational decisions also assume they have access to all available information in order to make rational-best decisions. However, Cranford et al. (2018) showed that even when given all possible information about the decision probabilities, humans do not seem to systematically factor such information into their decisions. Even for defenses that consider bounded rationality, the impact of feature representations is unclear.

If a defense algorithm or cognitive model assumes some information is represented in the decision when it is not, or assumes an incorrect representation, then the predictions of human behavior may be incorrect and may lead to sub-optimal solutions. For example, humans commonly make decision errors by failing to represent important information (e.g., running a red light) and concealing information can influence decisions (e.g. car salesmen and magicians fool the human mind by manipulating the saliency of information). Similarly, it is well known that humans represent values differently depending on whether they are framed as losses or not (i.e., framing effect, Tversky & Kahneman, 1981).

Human decisions from experience take as input the contextual features of the situation and generate decisions based on expected outcomes (Gonzalez et al., 2003). Therefore, two important dimensions to focus on are (1) what features are represented in the situation context and (2) how outcomes are processed and evaluated. The present research aims to better understand how these two factors influence human decision making in a cybersecurity game that simulates an insider attack scenario, called the Insider Attack Game (IAG).

Insider Attack Game

Cranford et al. (2018) designed the IAG to investigate how deceptive signals influence attacker decision making in an abstracted cybersecurity scenario. A screenshot of the task interface is shown in Figure 1A. Humans play the role of the attacker, depicted in the center surrounded by 6 targets. The attacker must first decide what target to attack. Two analysts monitor the six targets, and each target shows the probability that a target is being monitored, the points the attacker would win if they attack the target and it is not monitored (yellow stars), and the points they would lose if they attack the target and it is monitored (red stars). After selecting a target, the attacker is presented a message and asked if they would like to continue the attack or withdraw (Figure 1B). Sometimes the message indicates that the target is not being monitored, which is always truthful. However, other times the message indicates that the target is being monitored, and only sometimes is there an analyst actually monitoring the target. If the attack is withdrawn the attacker receives 0 points, but if the attack is

continued the attacker is rewarded or penalized based on the underlying coverage of the target. All decisions are self-paced, and no domain expertise is required. Attackers try to earn as many points as possible across 4 rounds of 25 trials (targets change each round), which is translated into a cash payout.

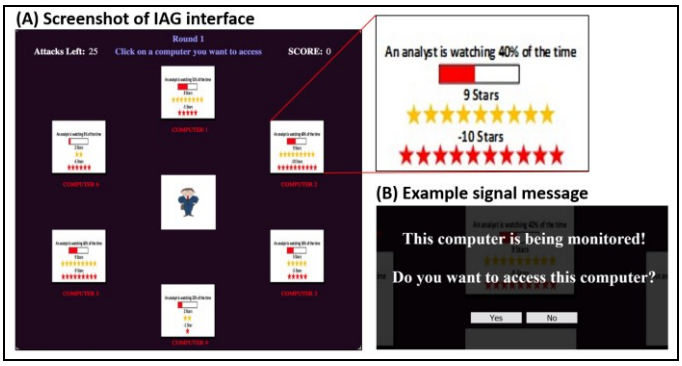

Figure 1. Screenshot of the IAG (A) and an example signal message indicating a target is being monitored (B). The first line is omitted when indicating a target is not being monitored.

The goal of adding deceptive messages is to improve defenses by deterring attacks when a target is not monitored and extending the perceived coverage of the network. The rate at which deceptive messages are sent is optimized through game-theoretic algorithms such as the Strong Stackelberg Equilibrium with Persuasion (*peSSE*; Xu, Rabinovich, Dughmi, & Tambe, 2015). However, the defense is contingent on sustaining belief in the signal while assuming adversaries make perfectly rational decisions. For this reason, in human studies, participants are provided all information regarding the payoff structure. Based on these values, the *peSSE* sends signals at a rate that makes the expected value of attacking given a signal equal to the expected value of withdrawing, and therefore a perfectly rational adversary will defer to the safe option and always withdraw. However, even with all possible information to aid decision-making, humans do not make rational-best decisions. Cranford et al. (2019) showed that humans attack far more often than predicted by the *peSSE* (~80% overall compared to the predicted 33%). This behavior was explained under the Instance-Based Learning Theory (IBLT; Gonzalez, Lerch, & Lebiere, 2003; Gonzalez, 2013) and an IBL cognitive model was developed that accurately predicts human performance in the IAG (with an impressive RMSE = 0.04 and $r = 0.80$; Cranford et al., 2018, 2019).

IBL Model of Human Behavior in the IAG

According to IBLT, human decisions from experience are based on the similarity of the current situation to past situations, modulated by the recency and frequency of those past experiences in memory (Gonzalez, 2013). An IBL model was created in the ACT-R cognitive architecture (Anderson & Lebiere, 1998; Anderson, Bothell, Byrne, Douglass, & Lebiere, 2004), which provides a theoretical framework that accurately simulates human-like cognition and processes such as memory retrieval, pattern matching, and decision making.

In the IBL model, described in more detail in Cranford et al. (2018), experiences (or instances) are represented by the contextual features of the decision. For example, in Figure 2, the contextual features include the information available in the environment, including the reward, penalty, and monitoring probabilities, as well as the action taken and the associated utility, or outcome of the decision. Each experience is saved in memory and when a new decision is to be made, an expected outcome is retrieved from memory that represents a weighted average across all memories based on similarity of the contextual elements and activation strength of the memory. In ACT-R the activation strength is determined by the recency of the memory and its frequency of occurrence. A Boltzmann softmax equation determines the probability of retrieving an instance based on its activation strength. The IBL model uses ACT-R's blending mechanism (Gonzalez et al., 2003; Lebiere, 1999) to retrieve an expected outcome of attacking a target based on a consensus of past instances. The expected outcome is the value that best satisfies the constraints of all matching instances weighted by their probability of retrieval.

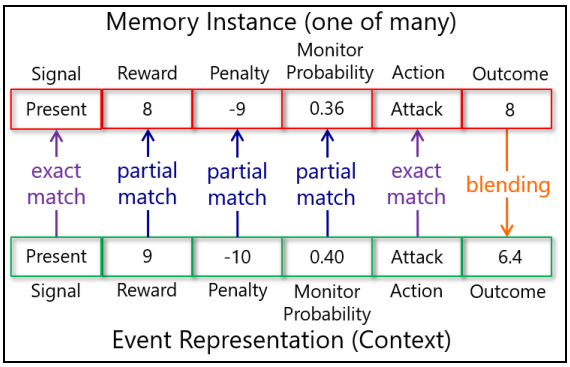

Figure 2. Example representation of instances in IBL.

First, the model selects a target by cycling through each of the targets and generating an expected outcome of attacking based on the reward, penalty, and monitoring probability features. The target with the highest expected outcome is selected. The context is then augmented with a feature representing the signal and the model generates a new expected outcome of attacking given the signal, but does not include the reward, penalty, and monitoring probability features in this decision because they are not present on the screen during this stage of the decision process. In fact, representing the target's features in this decision resulted in poorer fit to human data, suggesting that, when presented a signal, humans only factor the features of the signal and not the target. Finally, a decision to attack is made if the expectation is greater than zero, else the model withdraws, and ground truth feedback is given.

The model saves two instances to memory each trial. One represents the expectation generated during the decision to continue the attack or withdraw (includes the features: signal, action, and expected outcome), and the other represents the ground truth decision and feedback received (includes all features: signal, reward, penalty, monitoring probability, ground truth action, and ground truth outcome). While ground truth experience alone would predict a lower probability of attack, aligning with the statistics of the environment, storing the expectations drives a confirmation bias in which the availability of additional positive instances in memory perpetuates a behavior to attack even after suffering losses.

In summary, humans do not compute all information and make rational-best decisions, but instead make decisions based on experiences represented by the important features of the situation. The present study manipulations are aimed at better understanding how feature representation impacts decisions.

Experimental Manipulations and Hypotheses

Two observations were made from the human data reported in Cranford et al. (2019). First is that humans attack at a high rate from the very first trial and this pattern perpetuates throughout the game. It was hypothesized that, because participants' monetary payouts are not discounted for having negative total points, it is possible that early negative outcomes are not framed as losses. Instead they may be represented as higher than face value. According to IBLT, if the set of past experiences includes inflated outcomes, the generated expected outcome will also be inflated which could contribute to the high probability of attack.

To investigate how the framing of outcomes affect decisions, in the present study, we manipulated the number of points participants started with. In one condition, participants start with zero points as usual (NoLoss), but in the other condition participants begin with an endowment of 100 points (Loss). When given an endowment, early losses are predicted to be more meaningful and framed as losses (i.e., encoded as full penalty), but should not be framed as losses when the endowment is withheld (i.e., encoded as less than the full penalty). Therefore, according to IBLT, an endowment should increase the valuation of negative outcomes, resulting in lower expected outcomes, and thereby lowering the probability of attack. It is possible the effect will diminish in later rounds as participants accumulate points and have something to lose.

The second observation was that human decisions are not made through formal calculation of expected values based on the decision probabilities, but instead are based on experience, using features as representations of what happened in the past in order to generate expectations of the future. Therefore, to investigate how context representations affect decisions, we manipulated the information made available to participants. In one condition participants were provided all possible information (Info), but in the other condition the monitoring probabilities were withheld (NoInfo).

According to perfect rationality, providing more information is predicted to reduce the probability of attack because adversaries should be able to use the decision probabilities to calculate expected values and make the rationally-best decision to withdraw when given a signal. However, according to IBLT, concealing features such as the monitoring probability will affect how the current situation matches to past experiences, thus affecting what chunks are included in the retrieval set, and thus the expected outcomes. Because target features are only considered during the selection decision, it is predicted that concealing information will largely affect selection behavior and only indirectly influence the probability of attack insofar as shifting attacks toward different targets with different coverage probabilities will result in different experiences (e.g., more experiences of loss would result in lowered expected outcomes and lowered probabilities of attack).

IBL Model Modifications & Predictions

The cognitive model was modified to make predictions about how humans would behave in each of the conditions. To simulate the effects of not having a 100-point endowment, the model was modified so that any negative outcomes were changed to be no greater than the number of positive points available. To simulate the effects of concealing the monitoring probabilities, we simply removed the feature from instances in memory and so target selection decisions are based solely on the reward and penalty values. Figure 3 shows the model predictions of the probability of attack (i.e., the proportion of trials in which participants chose to continue the attack instead of withdraw). The model predicts a main effect of framing and also of information, and a slight reduction in probability of attack over time, but no interactions.

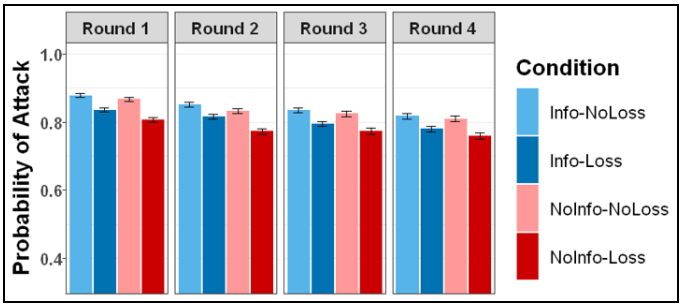

Figure 3. Mean probability of attack per round comparing the model in the Info-NoLoss, Info-Loss, NoInfo-NoLoss, and NoInfo-Loss conditions.

METHODS

Design

The design was a 2 (Framing) X 2 (Information) between subject design. For Framing, participants were either endowed 100 points to start the game (Loss) or given no endowment (NoLoss). For Information, participants were either provided information regarding the monitoring probabilities (Info) or this information was concealed (NoInfo).

Participants

100 participants were recruited via Mechanical Turk for each of the four conditions. However, some participants were removed from analysis due to incomplete data arising from technical issues, resulting in final sample sizes of 100, 98, 99, and 99 for the NoInfo-Loss, NoInfo-NoLoss, Info-NoLoss, and Info-Loss conditions, respectively. In the NoLoss conditions, participants were paid a base payment of $1.00, but in the Loss conditions, they were given 100 points to start the game. Participants could then earn $0.01 per point earned in the game up to a maximum total payment of $5.50.

Procedure

After providing informed consent, participants read instructions about their payout and the gameplay. In the Info condition, participants were informed about the monitoring

probabilities, but were not in the NoInfo condition. Participants answered a short quiz about the instructions and played 5 practice trials before beginning the main experiment. Participants continued to play the game for four rounds of 25 trials each as described in the Introduction. After completing the game, participants were given feedback about their results, completed a brief survey, and thanked for their participation. Payment was awarded within 24 hours of completion.

RESULTS & DISCUSSION

The data was analyzed for the probability of attack, as described above with the model predictions, and target selection preferences. The selection preferences examined the proportion of trials that participants selected each target.

The mean probability of attack for each round is shown in Figure 4, comparing Info (blue) to NoInfo (red) conditions and NoLoss (lighter) to Loss (darker) conditions. A mixed-effects ANOVA, with Round included as a within-subjects factor, revealed a main effect of Information, $F(1,392) = 4.08$, $p = .044$, and a main effect of Round $F(3,1176) = 15.27$, $p < .001$, but not of Framing, $p = 0.333$, and there were no interactions, all $p > .175$. Replicating prior research, participants attack gradually less across rounds. Of importance, and consistent with current hypotheses, participants attacked more often when information regarding the monitoring probability was available than when this information was concealed. However, closer inspection reveals that the effect is only present in rounds 1-3 in the NoLoss condition, all $p < .036$, diminishing in round 4, $p = .086$, and is not present in any rounds of the Loss condition, all $p > .248$.

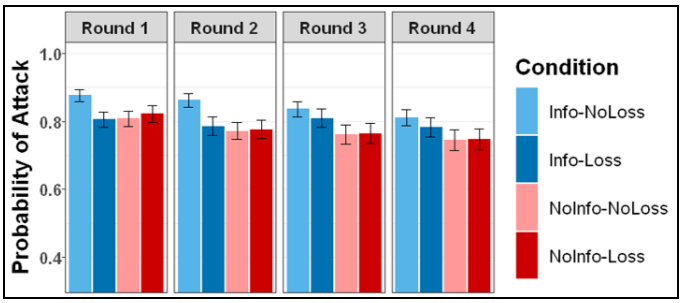

Figure 4. Mean probability of attack per round comparing humans across the Info-NoLoss, Info-Loss, NoInfo-NoLoss, and NoInfo-Loss conditions.

Although analysis revealed no main effect of Framing, nor a 3-way interaction, further inspection revealed a marginal interaction between Framing and Round within the Info condition, $F(3,588) = 2.31$, $p = .075$. Consistent with hypotheses, there is an early effect of providing an endowment whereby participants withdrew more often in Round 1, $F(1,196) = 6.34$, $p = .013$, and Round 2, $F(1,196) = 5.25$, $p = .023$, but not in Rounds 3 and 4, both $p > .425$. In the NoLoss condition, early losses do not deter future attacks as much as in the Loss condition, likely because they are represented as lower than face value (i.e., losses do not impact their monetary payout and memories reflect this experience), but this effect diminishes across rounds after points have been acquired and losses become more meaningful.

The model predictions match well to human data in the Info-NoLoss (RMSE = 0.04, $r = 0.73$), Info-Loss (RMSE = 0.05, $r = 0.71$), and NoInfo-Loss (RMSE = 0.05, $r = 0.73$) conditions, but not as well in the NoInfo-NoLoss (RMSE = 0.07, $r = 0.78$) condition in which the model attacks more often, predicting an effect of Framing that was not observed in the human data. To better understand why this effect of Framing was not observed in the NoInfo condition for humans, but was for the model, we must look further at how the concealment of monitoring probabilities influenced the selection behavior.

Figure 5 shows the mean probability of selecting each target for humans, plotted along the three features: reward, penalty, and monitor probability. As can be seen, there is not much difference between Framing conditions. However, there is a large observable difference between Information conditions. Concealing the monitoring probability shifts selection towards targets with higher rewards and lower penalties. Unbeknownst to participants in the NoInfo condition, these targets are also monitored more often, and therefore signaled more often, which explains why these participants attack less often than participants who are given the monitoring probability. Participants in the Info condition also tend to select targets with higher rewards, but also factor monitoring probabilities and select targets with the more moderate values.

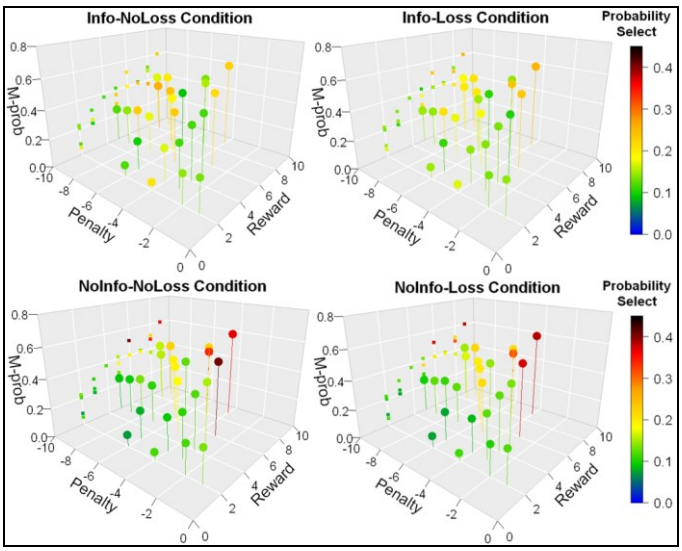

Figure 5. Mean probability of selecting a target, plotted by the reward, penalty, and monitoring probability (M-prob) target features, comparing humans across the Info-NoLoss, Info-Loss, NoInfo-NoLoss, and NoInfo-Loss conditions.

Meanwhile, the model did not exhibit this same shift in selection behavior as humans, but instead was stable across conditions and resembled humans in the Info condition (see Figure 6). The discrepancies of probability of attack between humans and the model can be explained by these differences in selection. For humans, attacking highly covered targets in the NoInfo condition could have led to the lower probability of attack observed when no endowment was given because, as participants experience more signals and decide to withdraw, they also experience fewer gains, thereby lowering the future expected outcomes and thus the overall probability of attack.

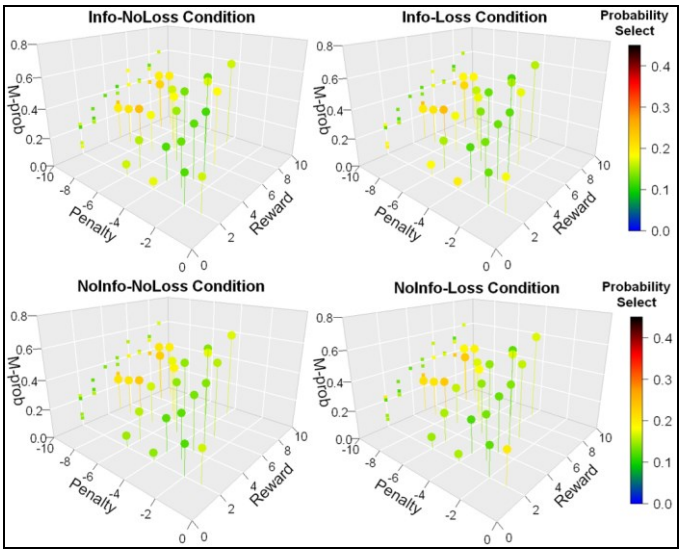

Figure 6. Mean probability of selecting a target, plotted by the reward, penalty, and monitoring probability (M-prob) target features, comparing the model across the Info-NoLoss, Info-Loss, NoInfo-NoLoss, and NoInfo-Loss conditions.

In summary, the manipulations had a large effect on target selection preferences. Concealing the monitoring probability shifted target selection preferences toward targets with higher reward/penalty ratios, which influenced the probability of attack. The effect of framing was only observed when participants could rely on monitoring probabilities to select more moderately covered targets. When selecting more frequently covered targets that are signaled more often in the NoInfo conditions, participants withdraw more often, leading to fewer experiences of losses for which the endowment has less of an impact on expected outcomes. Future research is aimed at modifying the model to capture these selection preferences observed when the monitoring probabilities are concealed. These modifications will help us gain a better understanding of how feature representation affects human decision making.

CONCLUSIONS

In conclusion, the present research demonstrated the importance of understanding how humans represent and use the information available to them in making decisions. Firstly, it is important to understand how humans represent outcomes and values because, if a model assumes a penalty is greater than the human perceives, then it will likely make predictions that overestimate the effectiveness of the defense. It is also important to consider how outcomes are framed because manipulating this feature can impact how attackers perceive outcomes which can influence their behavior.

Secondly, model predictions are limited to the extent that they can capture the important features of the decision. For example, Martin et al. (2018) showed how humans learn to use and weight the available features in the environment for making classification decisions. The features that are important for achieving the agent's goals are those that weigh more heavily in the decisions, as was evident in Somers, Mitsopoulos, Lebiere, and Thomson's (2019) research on saliency. In their research, decision errors were observed when the saliency, or relative importance, of the critical decision feature was low. As demonstrated here, it is important to understand how humans represent the information available and to be mindful of what information is made available. While too much information could overload an adversary and cause decision errors, presenting certain information could have adverse effects on defenses. Future research is therefore aimed at modifying the model to better capture selection preferences in order to gain a better understand of how concealing monitoring probabilities shifts selection preferences and affects attacking behavior.

ACKNOWLEDGEMENTS

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

REFERENCES

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

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.

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. In *Proceedings of the 40th annual conference of the Cognitive Science Society* (pp.258-263). Madison, WI: Cognitive Science Society.

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., & Lebiere, C. (2019). Towards personalized deceptive signaling for cyber defense using cognitive models. In *Proceedings of the 17th Annual Meeting of the International Conference on Cognitive Modeling*. Montreal, CA.

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

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591-635.

Gonzalez, C. (2013). The boundaries of instance-based learning theory for explaining decisions from experience. *Progress in Brain Research*, 202, 73-98.

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

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*.

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*, Special Issue BRIMS 2017.

Rowe, N. C., & Rrushi, J. (2016). *Introduction to Cyberdeception*. Switzerland: Springer.

Somers, S., Mitsopoulos, K., Lebiere, C., & Thomson, R. (2019). Cognitive-Level Salience for Explainable Artificial Intelligence. In *Proceedings of the 17th Annual Meeting of the International Conference on Cognitive Modeling*. Montreal, CA.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-548.

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