Cognitive Salience of Features in Cyber-attacker Decision Making

Edward A. Cranford (cranford@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Sterling Somers (sterling@sterlingsomers.com)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Konstantinos Mitsopoulos (cmitsopoulos@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Christian Lebiere (cl@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Abstract

While much is known about how humans make decisions based on the recency, frequency, and similarity of past experiences, much less is known about how humans weigh the contextual features and the impact it has on decisions. The present study uses a novel method of introspecting into a cognitive model of human decision making in an abstract cyber security game to gain insight about the cognitive salience of the features. The results show that cognitive salience can provide valuable evidence about how and why individuals make their decisions. The implications of these results are discussed with regard to theory and application.

Keywords: cognitive salience; cyber deception; cognitive models; instance-based learning, ACT-R

Introduction

Feature representation plays an important role in human decisions, and while much is known about how experiences shape decisions through instance-based learning (e.g., frequency, recency, and similarity effects; Gonzalez, 2013), much more needs to be understood about what features are represented in human decisions and how those features weigh in the decision. Instance-based learning (IBL) models have accurately modeled human behavior across a number of tasks including supply chain management (Gonzalez & Lebiere, 2005), social dilemmas (Lebiere, Wallach, & West, 2000), two-person games (Sanner et al., 2000, West & Lebiere, 2001), repeated binary-choice decisions (Gonzalez & Dutt, 2011), and multi-stage Stackelberg Security games (Cranford et al., 2020a). According to Instance-Based Learning Theory (IBLT; Gonzalez, 2013; Gonzalez, Lerch, & Lebiere, 2003) human decisions from experience are generated through the aggregate retrieval across past experiences, based on the feature similarity of the current situation to past situations.

While IBL models provide evidence for the underlying mechanisms involved in decision making, as well as evidence that the representations used in a specific model sufficiently describe its respective task, they do not provide any insight into the degree to which a decision maker weighs particular features of the decision. We believe that thwarting a would-be attacker could be more successful if we had insight into the features they find salient.

We consider the salient feature of a decision to be the feature that most influenced that decision and the greater degree of salience a feature has, the more influential it was in the decision. While the term, salience, might imply attention, in our use the salient feature may not be the most attended feature by some measure of attention (eye gaze, etc.). It may be the case that a feature is attended more than others but ultimately does not contribute to a decision.

Our salience mechanism is somewhat analogous to gradient-based saliency used in image classification (Grün et al., 2016). Gradient-based salience techniques calculate the gradient of a prediction with respect to the input image to estimate the importance of pixels. The result of this process is often a heatmap of pixels that, when overlaid on the original image, provide some insight into what parts of the image were most important for the classification.

We term our approach *cognitive salience* in contrast to the gradient-based approach. We use the term *cognitive* for two reasons. First, our approach calculates salience by taking the derivative of a theorized memory retrieval mechanism, blending, the mechanism underlying decision making in IBL models (Lebiere, 1999). Second, the features of a cognitive model are typically of a higher-level of abstraction than pixels, usually conceptual terms, which is typical of a cognitive-level description.

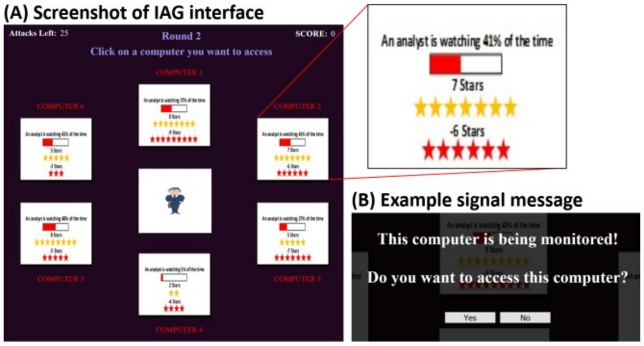
In the present work, we examine the saliency of features in a model of human decisions in a cybersecurity game called the Insider Attack Game (IAG). IBL models of human decisions in the IAG revealed cognitive biases, such as confirmation bias, that emerge naturally through memory retrieval processes, and lead participants to attack far more often than predicted by perfect rationality (Cranford et al., 2020a). While much has been learned by comparing model performance to humans and making inferences about human behavior based on the model mechanisms and processes, examining feature salience can provide further useful information regarding the relative importance of features when making decisions. These insights could prove useful in further informing how human decisions are shaped through their unique experiences, how representation of features impact decisions, and also for designing more effect cybersecurity defenses.

In what follows, we first describe the IAG and an IBL model that accurately captures human behavior in the game. Next, we describe the method for deriving cognitive salience from IBL model decisions. Assuming the model accurately reflects human decision-making processes, we apply our salience technique to the model to gain insight into how humans might weigh features when making decisions in the IAG. Finally, the results are discussed regarding their implications for theories of human decision-making and applications to cybersecurity.

IBL model of Attackers in the Insider Attack Game

The Insider Attack Game (IAG) was designed as a two-stage Stackelberg Security Game (SSG) to investigate the influence of deceptive signals on cyber-attacker decision making (Cranford et al., 2018). Players take the role of an insider attacker and make repeated decisions of "hacking" computers. In the first stage, attackers must decide which of six targets to attack, as depicted in Figure 1A. However, they must avoid the two analysts (defenders) that monitor one target each. An example target is shown in the zoomed inset of Figure 1A. Attackers are presented all information about the reward received if they successfully attack a target that is not monitored, the penalty received if they attack a target that is monitored, and the probability that the target is being monitored. After selecting a target, in the second stage, the attacker is presented with a message signaling whether the computer is being monitored (e.g., see Figure 1B). The message is always truthful when claiming a target is not being monitored. However, the attacker is informed that the message is sometimes deceptive when claiming the target is being monitored. The attacker must decide to either continue the attack and earn the reward or penalty, depending on the true underlying coverage, or withdraw and earn zero points. Attackers are incentivized to earn as many points as possible across four rounds of 25 trials each; a new set of targets are presented each round.

The defense algorithm in the IAG, the Strong Stackelberg Equilibrium with Persuasion (*peSSE*; Tambe, 2011; Xu et al., 2015), was designed to optimize the rate at which deceptive messages are sent so that belief in the signal is maintained, but does so under assumptions that adversaries make perfectly rational decisions. In the first stage, the algorithm optimizes the allocation of the two defenders based on the reward and penalty values of the targets. The algorithm effectively equalizes the expected value of all targets so that no target is more preferred than another and assigns defenders to targets across the trials according to the derived probabilities (these are the monitoring probabilities attackers see). In the second stage, the algorithm optimizes the rate of deceptive signals. If signals were only truthful, the expected value would be negative when indicating that a target is being monitored. The *peSSE* determines the probability with which to send deceptive signals so that the expected value given a signal increases to zero, and under assumptions of perfect rationality, an adversary will still defer to the safe option and withdraw. In the IAG, with a 1:3 defender-target ratio, the signal is present and truthful on 1/3rd of trials, on average. Therefore, the *peSSE* can send deceptive signals on another 1/3rd of trials when the target is not monitored. Thus, the peSSE increases the perceived coverage of the system by finding the optimal combination of bluffing (sending a deceptive message that the target is monitored when it is not) and truth-telling (sending a truthful message that the target is covered) so that a rational attacker would always withdraw in the presence of a signal.


(A) Screenshot of IAG interface
Attacks Left: 25
Round 2
SCORE: 0
Click on a computer you want to access
COMPUTER 1
An analyst is watching 37% of the time
8 Stars
-9 Stars
COMPUTER 6
An analyst is watching 41% of the time
5 Stars
-3 Stars
COMPUTER 2
An analyst is watching 41% of the time
7 Stars
-6 Stars
An analyst is watching 48% of the time
8 Stars
-5 Stars
COMPUTER 5
An analyst is watching 27% of the time
5 Stars
-7 Stars
COMPUTER 3
An analyst is watching 5% of the time
2 Stars
4 Stars
COMPUTER 4
An analyst is watching 41% of the time
7 Stars
-6 Stars
(B) Example signal message
Yes
No

This computer is being monitored!

Do you want to access this computer?

Figure 1: Screenshot of the IAG (A) and an example signal message (B). The first line of the message is omitted if the signal is absent. The zoomed inset shows a target, including the value of the reward if the attack is successful (yellow stars), the value of the penalty if the attack is unsuccessful (red stars), and the monitoring probability (given as percentage in text and graphically as a fillable gauge).

Humans, however, do not always make rational-best decisions. Instead, human decisions in the IAG can be explained under Instance-Based Learning Theory (IBLT; Gonzalez, 2013; Gonzalez et al., 2003), and an IBL model was created that captures this behavior (Cranford et al., 2018; 2020a). According to IBLT, decisions are made by generalizing across past experiences, or instances, that are similar to the current situation. In the IBL model, instances are represented by the contextual features of the decision. For example, in Figure 2, the features include the information available in the environment: the *reward*, *penalty*, *monitoring probability*, and *signal*, the *action* taken, and its associated utility, or *outcome*. 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 their probability of retrieval.

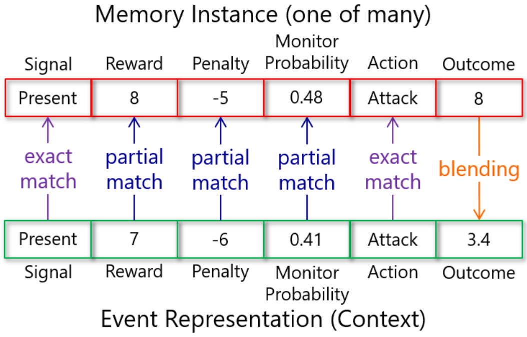


Figure 2: Example representation of instances in IBL.

The IBL model was created in the ACT-R cognitive architecture (Anderson & Lebiere, 1998; Anderson et al., 2004), which provides a theoretical framework that accurately simulates human-like cognition and processes such as memory retrieval, pattern matching, and decision making. In ACT-R, the probability of retrieving an instance is based on its activation strength which is determined by its recency and frequency of occurrence, and its similarity to the current context. The IBL model uses ACT-R's blending mechanism (described in more detail in the next section; 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.

The IBL model played the same game as humans. In the first stage of the IAG, the features of the decision include the monitoring probability [0.0, 1.0], the reward [1, 10], and the penalty [-1, -10]. The model generates an expected outcome for each target, via blending across previous outcomes, and selects the target with the highest expected outcome. In the second stage, the only feature in the decision is the signal [present, absent] and the model generates a new expected outcome of attacking. A straightforward decision rule is then applied: if the value is greater than zero the model attacks, else it 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). Storing the expectations as well as the ground truth drives a confirmation bias in which the availability of additional positive instances in memory (i.e., from the positive expectations generated prior to deciding to continue an attack) perpetuates a behavior to attack when faced with a signal, even after suffering losses.

Cranford et al. (2020a) showed that humans attacked about 80% of trials on average, far more than the predicted 33% of perfectly rational attackers (i.e., on average, signals are absent on only 1/3rd of trials). The IBL model very accurately captures this behavior across trials in the game (overall total RMSE = 0.04 and r = 0.73), as shown in Figure 3 (adapted from Cranford et al., 2020a). The pattern of spikes across trials can be attributed to the schedule of coverage and signaling, which was the same for each player and reflects experiences of success/failure given the probability of seeing a signal. In fact, the correlation between the probability of seeing a warning and the probability of attacking is -0.84 for humans and -0.89 for the model.

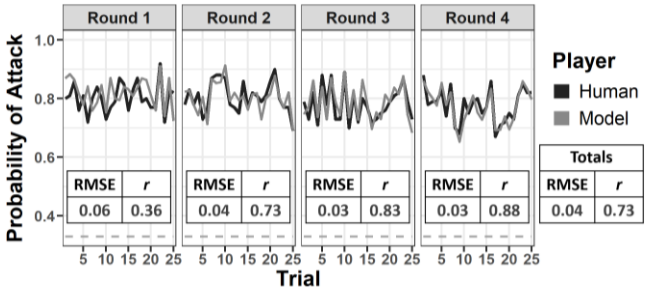


Figure 3. Mean probability of attack across trials for human participants compared to the IBL model runs.

The data presented in Figure 3 averages across substantial individual differences in behavior. In addition to capturing mean human performance across trials, the model also captures the full distribution of attack probabilities as seen in Figure 4 (adapted from Cranford et al, 2020a). Like humans, approximately 40% of participants (e.g., model runs) attack greater than or equal to 95% of trials. In another study that examined human behavior in the IAG, Cranford et al. (2020b) reported that approximately 23% of participants that attacked greater than or equal to 95% of trials also reported that they ignored the signal in their decisions. The study reported in Cranford et al. (2020a) did not collect such data, but we can assume similar responses would have been made.

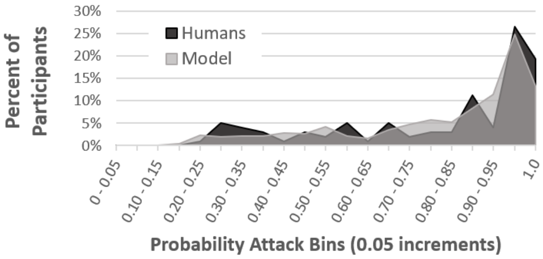


Figure 4: Probability of attack distribution for human participants compared to the IBL model runs.

In summary, the IBL model very accurately captures human behavior in the IAG and has proven useful in making inferences about human decision making in the task. Humans do not compute all information and make rational-best decisions, but instead make decisions based on past experiences, which are represented by the important features of the situation. These decisions are heavily influenced by the dynamics of memory retrieval processes which result in emergent cognitive biases (e.g., recency, frequency, and confirmation bias). These biases lead to overweighting of certain outcomes that, often, results in inflated expectations.

Humans fail to fully comply with the signal because they are more likely to expect a positive outcome than a negative one as belief in the signal deteriorates. While much has been learned about how experience influences decisions in the IAG regarding recency and frequency of instances, it is less clear how humans weigh the features in their decisions. Therefore, in the present study, we examine the salience of the features during the selection and attack decisions of the model to inform why certain decisions are made and if there are differences between types of participants in how information is processed that leads to the observed individual differences in attack behavior, as described in Figure 4.

Blending and Cognitive Salience

The cognitive salience of a feature can be derived from the blended retrieval mechanism. The blending mechanism in ACT-R retrieves an estimated outcome by interpolating across previously experienced outcomes (Lebiere, 1999). That interpolation process is weighted by the contextual similarity of the present instance to previous instances and is computed with the following equation:

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

The value, V, therefore is an interpolated value based on matching chunks i, weighted by their retrieval probability P_i. The complete blending process is outlined in Figure 5. The retrieval probability, equation 2, is derived from a Boltzmann softmax function that is based on the activation strength of chunks, which is influenced by power laws of frequency and recency, according to ACT-R theory of memory retrieval (Anderson & Lebiere, 1998; Anderson et al., 2004), and also the similarity or match between the current instance in memory and past instances. The match score in equation 1 is equivalent to the similarity function, $Sim(V_t, v_{it})^2$, and is used to compare memory chunks v_{it} and candidate consensus values V_t. In the simplest case, where the values are numerical (i.e. the return R_T) and the similarity function is linear, the process simplifies to a weighted average by the probability of retrieval, as shown in equation 3 of Figure 5.

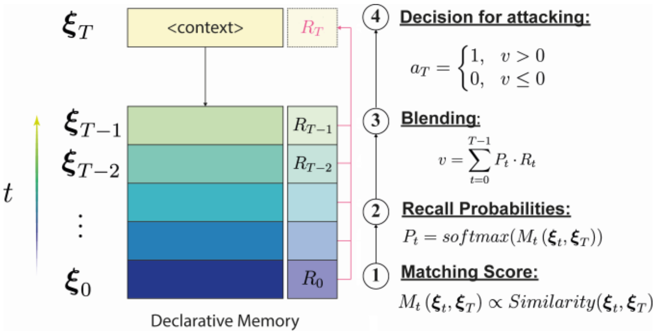

ξ_T
<context>
R_T
ξ_{T-1}
R_{T-1}
ξ_{T-2}
R_{T-2}
t
$\vdots$
ξ_0
R_0
Declarative Memory
4 Decision for attacking:
$a_T = \begin{cases} 1, & v > 0 \\ 0, & v \leq 0 \end{cases}$
3 Blending:
$v = \sum_{t=0}^{T-1} P_t \cdot R_t$
2 Recall Probabilities:
$P_t = softmax(M_t(\xi_t, \xi_T))$
1 Matching Score:
$M_t(\xi_t, \xi_T) \propto Similarity(\xi_t, \xi_T)$

Figure 5: Description of blending mechanism.

We calculate salience by taking the derivative of the blending equation with respect to each feature:

$$S(V_t, f_k) = c \sum_{i=1}^{n} P_i \cdot \left(\frac{\partial Sim(f_k, v_{ik})}{\partial f_k} - \sum_{j=1}^{n} P_j \frac{\partial Sim(f_k, v_{jk})}{\partial f_k} \right) \cdot v_{it}$$

This derivative gives us the degree of influence a particular feature (f_k) had in a decision (V_t). The value S can be infinitely positive or negative. While the direction of the value provides information about how the feature is used, to compare between features, the magnitude tells us which feature has the greater impact on the decision. Therefore, in all analyses below, we examine the absolute values of S.

Cognitive salience was first applied in an explainable artificial intelligence context (Somers et al., 2019), where ACT-R was used to model a reinforcement learner (RL). In that context, the baseline equations in ACT-R, which are responsible for effects of recency and frequency, were not used because in an RL context, there is no reason to expect decay in memory. This is the first time that cognitive salience has been applied in a human experimental context.

Salience Analysis

We examined the salience of features during the first-stage, selection decision and during the second-stage, attack decision of the IBL model. During target selection, a perfectly rational attacker would display no preference for features because all targets have the same expected value. No one feature is more informative than another and do not differentiate targets. However, it is possible that, for boundedly rational humans that derive expected outcomes from very limited experiences, selection preferences could emerge if one feature becomes more salient than another. It is hypothesized that saliencies will be higher when few instances are available in memory, thus skewing the mental representation of expected values. With more experiences, the attacker should gradually and implicitly learn that all targets are of about equal values. Saliencies can inform us whether decisions reflect the statistics of the environment, that the features are by all accounts meaningless.

During the second-stage, attack decision, there are clear individual differences in the probability of attack. It is therefore hypothesized that, when the signal is present, salience for the signal is lower for those participants with high probabilities of attack. If the salience of the signal is low then, when a signal is present, the expected outcome generated should be higher because the decision will discriminate less between signal types, giving more weight to instances when a signal was not present and that have positive outcomes, thus driving up the blended outcome.

First-Stage Decision: Selection

Figure 6 shows the average magnitude (i.e., absolute value) of saliencies for the reward, penalty, and monitoring probability (Mprob) features of the selected target across the four rounds of the experiment. The saliencies presented in Figure 6 are calculated during the outcome generation process. When generating expected outcomes, the model interpolates from previous experiences, weighing those experiences by their similarity to the features of the current target. The saliencies indicate that the model initially displays differences between features, but quickly merge within a few trials. After merging, all saliencies start out relatively high and decrease over time. The dashed gray line in Figure 6 shows the mean expected outcome across trials on the secondary y-axis. The expected outcomes are initially inflated and gradually decrease over time along with saliencies. These results suggest that the model is learning that the targets have equal expected outcomes and no feature is more salient than another. The decreasing trend in saliencies implies that the features are less influential in the decision over time. The model does not have any explicit awareness or explicit modeling of this decrease.

While the average magnitude of saliencies indicates no preferences for any particular feature, Figure 7 examines the individual differences between model runs (i.e., players) regarding the relative magnitude of the saliencies. The ternary plot takes the overall mean magnitude saliencies for each feature for each player and plots a point for each player to show the relative importance of each feature (i.e., the ratio between the three saliencies). The results show that players mostly display no preference for features as they learn over time that no feature is more meaningful and the saliencies trend downwards (as shown in Figure 6). Figure 7 shows that players mostly center around the middle point of the plot, indicating saliencies for features are all close to the same magnitude. However, all players display a higher saliency for one of the features, even if miniscule, and some do extend toward the corners of the ternary plot if only by a small percentage. Players were therefore split into 3 groups depending on the feature that is overall most salient, the reward, penalty, or Mprob, to examine if these groups display any target selection preferences.

Figure 8 shows a scatterplot of target selections by the reward and penalty values of the selected target. The size of the dots indicates the percent of selections for that target within a round. Thus, the dominant color for a point indicates a greater percent of selections for that group on that target. The distributions of penalties and rewards for each "Max Saliency" group are shown as marginal plots on the right and top, respectively. Figure 8 shows clear differences in target selection behavior between groups. The players that have higher saliency for the penalty tend to select the targets with higher penalties, which incidentally also have higher rewards. Meanwhile, players that have higher saliency for reward or Mprob tend to select the targets with lower penalties, and the ones with more moderate reward/penalty tradeoffs, more often.

Second-Stage Decision: Attack

In the second-stage decision, it was hypothesized that *aggressive* attackers (i.e., those that attacked ≥ 95% of the time) may weigh the signal differently than the other *cautious* attackers. Figure 9 shows the mean magnitude of saliencies

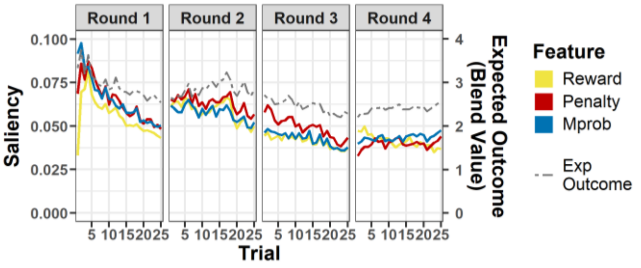


Figure 6: Mean magnitude of saliencies across trials for each feature of the selected targeted. The dashed gray line shows the mean expected outcome of the selected target.

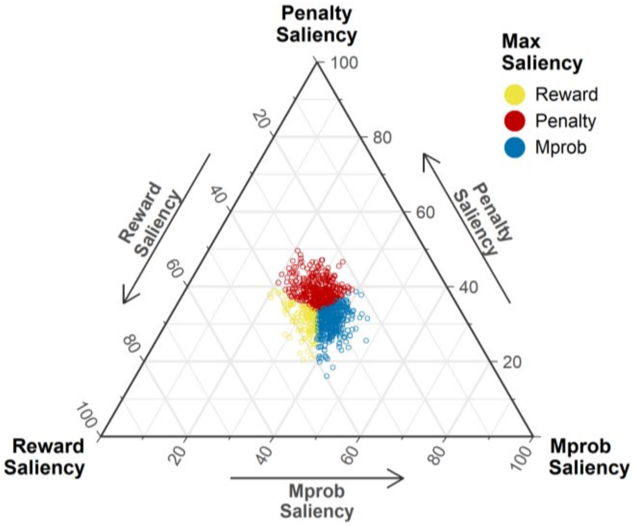


Figure 7: Mean relative saliency per player. Colors indicate the most salient feature for that player.

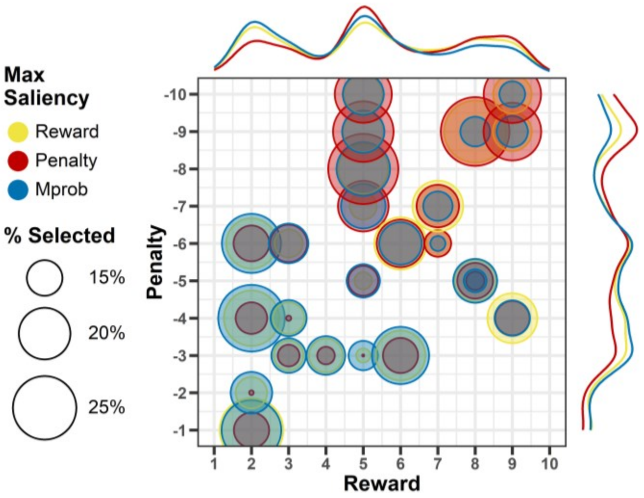


Figure 8: Scatterplot of reward and penalty of selected targets for each Max Saliency group. The size of the bubbles represent the percent of selections for the target within each Max Saliency group and round.

for the signal feature across the four rounds. The breaks in the solid line for the signal-absent condition are trials in which every target was scheduled to present a signal. When the signal is absent, the saliency for the signal is about equal in Round 1, but is higher for the aggressive attackers through the remainder of the game. For both player types, the saliency is overall higher when the signal is absent than when the signal is present. When the signal is present, however, the salience of the signal is lower for aggressive attackers than for the cautious ones. As predicted, when present, the signal is less influential in aggressive attackers' decisions than cautious attackers' decisions.

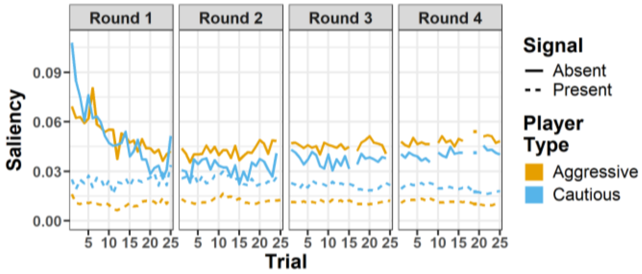


Figure 9: Mean magnitude of salience for the signal during the attack decision, for each signal and player type.

Figure 10 shows the mean expected outcome for the attack decision across the four rounds. An interesting interaction presents when compared to the pattern for saliencies. For expected outcomes, when the signal is absent, the aggressive attackers generate higher values than cautious attackers, echoing the pattern for saliencies. However, in contrast to saliencies, when the signal is present, cautious attackers generate lower expected outcomes than aggressive attackers.

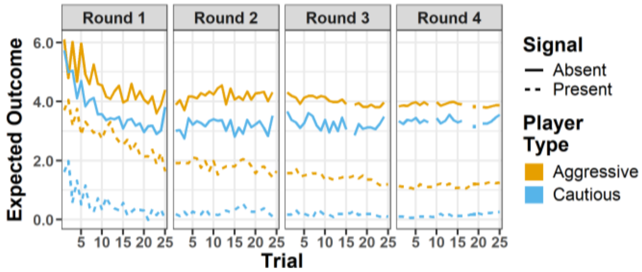


Figure 10: Mean expected outcome of the attack decision, for each signal and player type

These results suggest that salience influences decisions through its influence on memory retrieval. A higher saliency for a signal means that, for blending, more weight is given to past instances that have the same value as the current signal, and when saliency is low for the signal more weight is given to past instances that have a different value as the current signal so that probabilities are more evenly distributed across past instances. When a signal is absent, the higher salience for aggressive attackers means more weight is given to past instances whose signal is absent. Because these instances have only positive values, the expected outcomes are inflated. When a signal is present, the lower salience for aggressive attackers means the probability of retrieving past instances whose signal is absent is higher, and more evenly distributed across all targets (hence the expected outcome by Round 4 is approximately equal to the true expected value, irrespective of the signal, of 1.43), and again the expected outcomes are inflated. Meanwhile, cautious attackers that have higher salience for the signal when present have expected outcomes near zero, which is the true expected value given a signal.

Conclusion

The present study is the first to use this method for calculating cognitive salience to introspect into the model how humans weigh the contextual features in their decisions. The results provide additional insight into how the representation of features can influence decisions. Specifically, we can infer that players learn quickly that all targets have equal expected values and no feature is more informative than another. But also, because human decisions are based on limited experiences, some individuals have a slight preference for some targets that can be predicted by the degree of salience for a particular feature. An interesting path for future research will be to examine whether target selection preferences emerge if the targets are not of equal expected value and/or certain features are more indicative of success than others, as was shown in the original work on cognitive salience in an explainable artificial intelligence context (Somers et al., 2019). In that research, feature preferences emerge as certain features are more indicative of successful decisions.

In contrast to the selection decision, during the attack decision the degree of salience for the signal has a direct impact on the probability of retrieving past instances which in turn impacts the expected outcomes generated. The results provide unique insights into how individual differences can emerge through unique experiences. Understanding how an individual weighs the feature in their decisions provides valuable evidence about how information is processed and how it impacts decisions, which is vitally important for improving security defenses, especially for defenses that rely on adaptive and personalized defense (Cranford et al., 2020b). Therefore, one avenue for future research will be to validate the observed model results with human experiments designed to investigate what features are most important in decisions. For example, Cranford et al. (2020b) showed that aggressive attackers tended to report that they ignored the signal feature, and a model that omitted that feature from the representation was a better predictor of these aggressive attackers. As was demonstrated, understanding what features are important in a decision can inform the design of models and about the underlying representation of the decision. More accurate models can provide more accurate predictions about human behavior which can be used to improve security algorithms. Examining cognitive salience with cognitive models provides valuable information about individual differences between players, and future research aims at exploring its utility for designing more effective personalized, adaptive signaling schemes for cyber defense.

Acknowledgments

This research was funded by DARPA XAI contract FA8650-17-C-7710 and by the Army Research Office 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. (2020a). Towards personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science* (under review).

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* (pp. 1885-1894). Maui, HI.

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

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

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*. Macmillan (Palgrave Macmillan).

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

Grün, F., Rupprecht, C., Navab, N., & Tombari, F. (2016). A taxonomy and library for visualizing learned features in convolutional neural networks. arXiv preprint arXiv:1606.07757.

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

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* (pp. 185-193). NL: Universal Press.

Sanner, S., Anderson, J. R., Lebiere, C., & Lovett, M. C. (2000). Achieving efficient and cognitively plausible learning in Backgammon. *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann.

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.

Tambe, M. (2011). Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned. *Cambridge University Press.*

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.

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.