Evaluating Models of Human Adversarial Behavior Against Defense Algorithms in a Contextual Multi-Armed Bandit Task

Marcus Gutierrez (mgutierrez22@miners.utep.edu)
Computer Science Department, University of Texas at El Paso

Jakub Černý (cerny@disroot.org)
Computer Science Department, Nanyang Technological University

Noam Ben-Asher (noam.ben.asher@gmail.com)
Army Research Laboratory

Efrat Aharonov (efrat.aharonov@gmail.com)
Department of Social & Decision Sciences, Carnegie Mellon University

Branislav Bošanský (branislav.bosansky@agents.fel.cvut.cz)
Agent Technology Center, Computer Science Department, Czech Technical University in Prague

Christopher Kiekintveld (cdkiekintveld@utep.edu)
Computer Science Department, University of Texas at El Paso

Cleotilde Gonzalez (coty@andrew.cmu.edu)
Department of Social & Decision Sciences, Carnegie Mellon University

Abstract

We consider the problem of predicting how humans learn interactively in an adversarial Multi-Armed Bandit (MAB) setting. In a cybersecurity scenario, we designed defense algorithms to assign decoys to lure attackers. Humans play the role of cyber attackers in an experiment to try to learn the defense strategy after repeated interactions. Participants played against one of three defense algorithms: a stationary strategy, a static game-theoretic solution, and an adaptive MAB strategy. Our results show that humans have the most difficulty learning against the adaptive defense. We also evaluated five different models of attack behavior and compared their predictions against human data. We show that a modified version of Thompson Sampling and a cognitive model based on Instance-Based Learning Theory are the best at replicating human learning against defense strategies. We discuss how these models of human attacker can inform future cyberdefense tools.

Keywords: Cognitive Modeling; Reinforcement Learning; Intelligent Agents; Decision Making; Cybersecurity

Introduction

With the popularity of autonomous systems, the question of how humans interact with these systems becomes increasingly important (Gershman, Horvitz, & Tenenbaum, 2015). Humans are imperfect agents, but they are capable of learning and in some settings able to adapt to novel situations. Our ability to anticipate human behavior, to represent human decision making computationally, and to use these predictions to improve autonomous agents is critical to making autonomous systems more capable and secure.

We study an adversarial decision making setting framed in the context of cybersecurity. Humans attackers try to compromise a network while automated defender algorithms deploy decoys in the network (i.e., honeypots) to detect and thwart attackers. Honeypots are designed to waste the attacker's resources and provide information to the defender (Spitzner, 2003). Attackers try to avoid detection by honeypots. Deploying a fixed configuration of honeypots (i.e., a static defense) may capture an attacker in a single interaction. However, an adaptive attacker may learn the static honeypot defenses and actively avoid them in future interactions. A defender who can predict this attack learning dynamic should be able to deploy defensive strategies that are harder to learn and defeat over the long term. Our goal is to determine how human attackers behave against defense algorithms of various complexities, and to test cognitive models of adversarial behavior against other common behavioral models.

We model a cybersecurity scenario as a repeated Multi-Armed Bandit task (MAB) where a human attacker plays against an automated defender. MAB tasks have been useful in the study of human decision making, characterizing the common exploration-exploitation tradeoff (e.g., (Steyvers, Lee, & Wagenmakers, 2009)). However, our goal is to determine how a human attacker is able to learn the defender's deception strategy and avoid honeypots based on previous experience.

In a standard MAB, a decision maker select arms on a "slot machine" in each round and observes the outcome, typically with the value of each arm in the range $[0,1]$. The adversarial MAB considers an adversary (i.e., the algorithmic defender) who has control over the rewards of each node. Here, we consider a variation of the MAB in which each node i has bounded support interval $\{-c_i^a, v_i - c_i^a\}$. This allows the MAB agent to make more informed decisions in earlier

rounds. This maps naturally to an attacker who has probed the network prior to making an attack, and it relates to recent approaches to study learning and decision making under contextual MAB, where information about rewards is provided.

Learning in Multi-Armed Bandits

In a MAB, individuals learn by repeatedly choosing among multiple options that have varying probabilities of different rewards that are observed through immediate feedback after a choice. In theories of decisions from experience, two-arm bandit problems are a classical research paradigm used for modeling human decisions and learning from experience (e.g., (Gonzalez & Dutt, 2011)).

Experiments of human behavior have demonstrated that humans are able to learn in MABs by gradually transitioning from exploration of the available alternatives to exploitation of the most rewarding options while learning from feedback and experience (Gonzalez & Dutt, 2016; Mehlhorn et al., 2015). Sripa et al. notably ran an experiment with 451 human participants playing the MAB (Sripa et al., 2009), and applied a Bayesian learning model to explain the human data. Zhang et al. extended this work by improving the participant behavioral prediction with a Knowledge Gradient model (Zhang & Angela, 2013). Our current work differs from these works in that we consider differences in reward distributions. Specifically, the previously mentioned authors address human performance in stochastic settings. In this work, we consider humans in static, stochastic, and adversarial MABs settings and analyze the effects of each environment. Furthermore, we provide context to the human decision makers by advertising the potential gains and losses of each arm of the MAB.

Recent research has shown that humans are able to learn well in contextual MABs, and various algorithms have been used to replicate this human behavior, including Thompson sampling (Agrawal & Goyal, 2012; Speekenbrink & Konstantinidis, 2015). In contrast to these models often used in MAB tasks, cognitive models of human behavior represent the cognitive mechanisms (e.g, memory, learning, forgetting) which are essential elements for human learning (Gonzalez, Lerch, & Lebiere, 2003). We offer a unique paradigm to test cognitive models of human learning and decision making and pair them against other representations of behavior in MAB tasks, playing against defense algorithms of various complexities.

Honeypot Cybersecurity Game

In the Honeypot Cybersecurity Game (HCG) a defender places decoys to protect network resources (nodes) and the attacker aims to capture those resources. A screenshot of the user interface shows a network with 5 nodes (Figure 1). Each node in the network has the following values: v_i is the value of node i, c_i^a is the cost to attack node i, and c_i^d is the cost to defend node i. The reward $v_i - c_i^a$ for attacking a non-honeypot appears as a positive number on top of each node. The cost for attacking a honeypot $-c_i^a$ appears as a negative number at the bottom of the node.

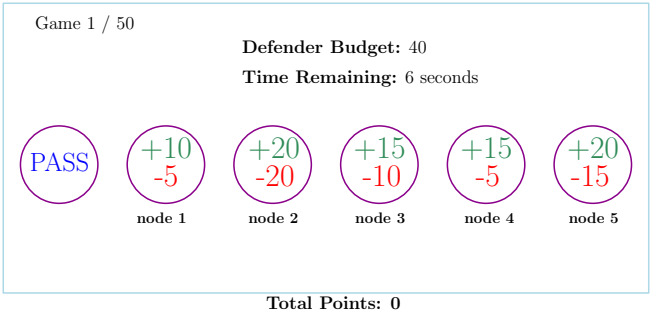

Figure 1: User interface for the HCG.

Table 1 shows the specific values used in the HCG for our experiments. We designed the node values to fit common risk-reward archetypes (e.g., low-risk/low-reward, high-risk/high-reward, low-risk/high-reward). The explicit values shown in each node give an attacker the possibility of making informed decisions that will be combined with experiential decisions as in (Lejarraga, Dutt, & Gonzalez, 2012).

	pass	node 1	node 2	node 3	node 4	node 5
v_i	0	15	40	35	20	35
c_i^a	0	5	20	10	5	15
c_i^d	0	10	20	15	15	20

Table 1: Node parameters for online human experiment.

At the beginning of each round, the defender spends her budget D to turn some subset of the nodes into honeypots, such that the total cost is $\leq D$. Once the defender deploys honeypots, the attacker selects a node to attack or passes. If the attacker's chosen node i is not a honeypot, the attacker receives the reward $v_i - c_i^a$, and the defender receives a reward of 0. If the attacker's chosen node i was a honeypot, the attacker receives the negative reward $-c_i^a$, and the defender receives the positive reward v_i[1]. At the end of a round with n trials, the game resets and a new round begins. The attacker and defender are only informed of the rewards they receive after each action, and do not directly observe the other player's choices (known as incomplete or semi-bandit feedback).

Defender Algorithms

We consider 3 different defender algorithms to investigate their impact on human adversarial decision making and learning. We expect these to create varying levels of difficulty for the human attackers to learn the defense policy.

The *Static Pure Defender* algorithm employs a "set and forget," defense that implements an unchanging, greedy strategy that spends the budget to protect the highest valued nodes. For the scenario in Figure 1, the defender always sets nodes 2 and 5 as honeypots, leading to nodes 3 and 4 being the optimal ones to attack. Against this defender, the attacker can gain a maximum of 750 total points in this specific scenario by always attacking node 3 or 4 for all 50 rounds.

[1]We assume $v_i \geq c_i^a$ and $\sum_{i \in N} c_i^d > D$.

The *Static Equilibrium Defender* plays according to a fixed probability distribution over the possible combinations of nodes to be honeypots. A new combination is selected randomly each round according to the distribution shown in Table 2. This is a game-theoretic Mixed Strategy Nash Equilibrium that optimizes the defender's expected utility assuming a single, non-repeated interaction against a fully rational attacker. The optimal strategy for the attacker against this strategy is to attack node 4, with an expected total value of ≈ 447 points for the attacker.

defended nodes	{1,3,4}	{2,3}	{2,5}	{3,5}
probability	≈ 0.303	≈ 0.095	≈ 0.557	≈ 0.0448

Table 2: Static Equilibrium Defender probabilistic strategy.

Algorithm 1 Learning with Linear Rewards (LLR)

If $\max_a |\mathcal{A}_a|$ is known, let $L = \max_a |\mathcal{A}_a|$; else, $L = N$
for $t = 1$ to N **do**
  Play any action a such that $t \in \mathcal{A}_a$
  Update $\left(\hat{\theta}_i\right)_{1\times N}$, $\left(m_i\right)_{1\times N}$ accordingly
end for
for $t = N+1$ to ∞ **do**
  Play an action a which solves the maximization:

$$a = \arg\max_{a\in\mathcal{F}} \sum_{i\in A_a} a_i\Big(\hat{\theta}_i + \sqrt{\frac{(L+1)\ln n}{m_i}}\Big), \qquad (1)$$

  Update $\left(\hat{\theta}_i\right)_{1\times N}$, $\left(m_i\right)_{1\times N}$ accordingly
end for

The *Adaptive Learning with Linear Rewards Defender (LLR)* (Gai, Krishnamachari, & Jain, 2012) plays an adaptive, learning defense strategy that tries to maximize reward by balancing exploration and exploitation using an approach designed for MAB learning. $\mathcal{A}_a$ in LLR is the set of all individual actions (nodes to defend). In the scenario from Figure 1, $\mathcal{A}_a$ is the set containing all 5 nodes. LLR uses a learning constant L, which we set to $L = 3$ since this is the maximum number of nodes we can play in a defense. LLR has an initialization phase for the first $N = 5$ rounds where it guarantees playing each node at least once. $\left(\hat{\theta}_i\right)_{1\times N}$ is that vector containing the mean observed reward $\hat{\theta}_i$ for all nodes i. $\left(m_i\right)_{1\times N}$ is the vector containing m_i, or number of times node i has been played. The vectors are updated after each round.

After the initialization phase, LLR solves the maximization problem in equation 1 and deterministically selects the subset of nodes that maximizes the equation each round until the end of the game. The algorithm tries to balance between nodes with high observed means (i.e., have captured the attacker often in the past) and exploring less frequently played nodes (which the attacker may move to in order to avoid capture). While LLR has no concept of an opponent, it indirectly adapts to the attacker based on the observations of previous rewards that depend on the attacker's strategy.

In this scenario, it is difficult for the attacker to fully exploit the strategy of the defender due to incomplete information. When facing a static defender in a static environment, the optimal node(s) will remain the same, but when facing LLR or another adaptive defender the node(s) providing the highest expected value may change from round to round.

Experimental Design

We recruited 304 human participants on Amazon's Mechanical Turk (AMT) where 130 reported female and 172 reported male with 2 participants reporting as other. All participants were above the age of 18, and the median age was 32. Participants interacted with one of the 3 defense algorithms for 50 rounds. 101 participants played against the Static Pure Defender; 100 played against the Static Equilibrium Defender; and 103 played against the LLR defender. Participants took roughly 10 minutes from start to finish. They were paid US $1.00 for completing the experiment and were given a bonus payment proportional to their performance in the 50 round game, ranging from US $0 to an extra US $3.25.

This task did not require cybersecurity knowledge and participants were given detailed instructions and definitions of the concepts needed to perform the task (e.g., honeypot). Participants were told that the defender has a budget $D = 40$ that limits the number of honeypot configurations (i.e., combinations of defended nodes). In each round, the participant attacks a node and receives either a positive reward $v_i - c_i^a$ or a negative reward $-c_i^a$ depending on the defender's action. The setup in Figure 1 was the same for every participant.

We analyzed 4 measures associated with participants' performance, and we compared predictive algorithms using the same measures. **Switching** is a common measure of exploration used in human decision-making and learning studies (Gonzalez & Dutt, 2016; Todd & Gigerenzer, 2000). High switching indicates high exploration and low switching indicates exploitation in the case of a static defender and static environment. **Switching with Honeypot** is a measure of switching after attacking a honeypot (i.e., receiving a negative reward). This corresponds with the "Lose-Shift" aspect of Win-Stay-Lose-Shift (WSLS) (Robbins, 1985), a common strategy studied in economics. **Switching without Honeypot** measures switching after attacking a real node (i.e., receiving a positive reward). This opposes the "Win-Stay" aspect of the WSLS (i.e., "Win-Shift"). Finally, **Optimal Play** is the fraction of decisions that have the actual highest expected value.

Behavioral Results

The results for the 4 dependent measures are shown in Figure 2. The rightmost graph in Figure 2 shows the frequency of optimal decisions over the 50 rounds. We note that participants playing against the static pure defender learn very early to play optimally and significantly improve over time, while the difference between the static equilibrium defender and adaptive LLR defenders is not clear early on. A significant advantage for LLR only emerges after at least 20 rounds.

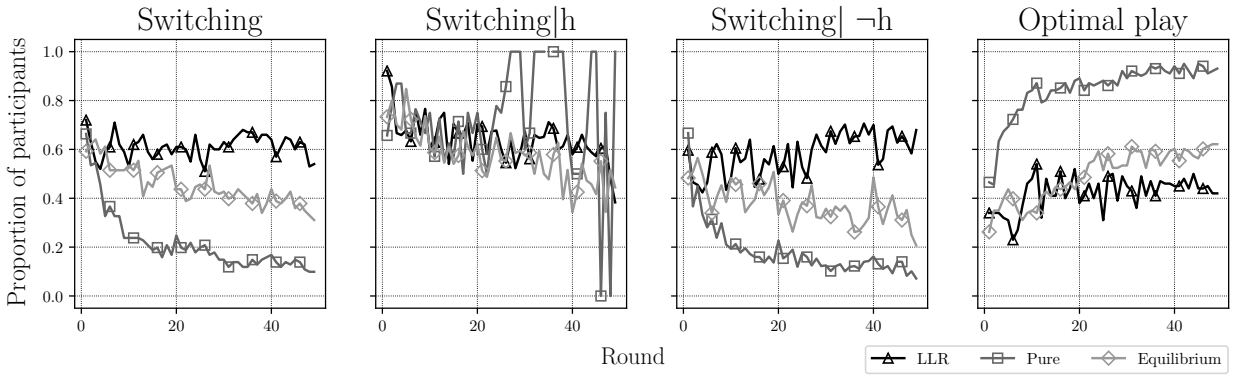

Figure 2: The proportions of participants switching nodes and playing optimally over time. The high switching after triggering a honeypot seen in round 26 from participants facing the static pure defender is a small portion of the population.

We also observe in the leftmost graph in Figure 2 that the overall proportion of switching decreases over time, particularly when participants face the static pure defender. When the participants face the adaptive LLR defender, they seem to have a high proportion of switching throughout the 50 rounds.

The middle left graph in Figure 2 describes the participants' switching behavior after triggering a honeypot. For the static pure defender, the attackers show noticeable spikes because only a few participants attacked the 20 point nodes (triggering the honeypots), upon which the players immediately switched. There are few differences between switching behavior when triggering honeypots of the participants who faced the equilibrium defender and those who faced adaptive LLR. We see a downward trend, hinting that the participants are moving from an early exploratory state to a more exploitative state. Since adaptive LLR updates it's beliefs about a node's expected payoff after playing it, if it captures an attacker that node will be more likely to be selected in the immediate future. Due to this adaptive behavior, switching when triggering a honeypot against adaptive LLR will be more beneficial than against the static equilibrium defender. When facing the static equilibrium defender in this scenario the attacker should always attack node 4, regardless of triggering a honeypot or not.

The middle right graph in Figure 2 shows distinct differences when the attackers did not trigger a honeypot (i.e., received a positive reward). Concerning the static pure defender and static equilibrium defender, decreases in switching demonstrate a move towards a more exploitative strategy and understanding of the static defense. Compare this with participants who faced the adaptive LLR defender where the switching remains high in comparison to the defenders. In general, adaptive LLR tries to react to the observed rewards and slowly moves from exploration to exploitation over time. High switching and remaining mobile is a good strategy against adaptive LLR. However, when we compare the participants' switching behavior with their performance versus adaptive LLR, it appears the participants were largely unable to learn the LLR strategy.

Overall, the pure defender predictably performed the worst (best for the human attackers), yielding an average score of 611.93 points. The equilibrium defender performed significantly better, yielding an average of 247.81 points. Finally, LLR was the most resilient defender against the human attackers with an average of 172.6 points yielded to the participants. Table 3 shows the aggregate statistics of the human attacker performance in terms of end-game attacker points.

	average	std. dev.	median	min	max
Pure	611.93	168.88	675	-375	750
Equ.	247.81	149.60	290	-185	570
LLR	172.6	123.02	160	-85	640

Table 3: Aggregate data of participants' end-game attacker points.

Adversarial Models

We evaluated 4 behavioral models and one cognitive model (IBL) (Gonzalez et al., 2003) to emulate participants' performance in the experiment. These models can give insights into the underlying mechanisms that influence decision making and support the development of better defense algorithms that hinder human attacker learning in cybersecurity settings. The models selected below are representatives of behavioral predictors that have been known to capture human performance in numerous MAB settings (Sripa et al., 2009; Zhang & Angela, 2013; Agrawal & Goyal, 2012).

Win-Stay-Lose-Shift: WSLS plays uniform randomly on the first round. If WSLS receives a positive reward, it attacks the same node again in the next round. Otherwise, it attacks another node uniform randomly. The "pass" action does not count as a positive reward.

ε-Greedy: This model addresses the exploration-exploitation dilemma directly with the parameter $\varepsilon \in \{0,1\}$. With probability ε, ε-Greedy attacks uniform randomly (exploration) and with probability $(1-\varepsilon)$, attacks the node with the highest observed average reward (exploitation).

ε-Greedy Decreasing: ε-Greedy Decreasing dynamically changes the parameter ε in order to prefer exploitation to-

	LLR				Pure				Equilibrium			
	Sw	Sw| h	Sw|¬h	OP	Sw	Sw| h	Sw|¬h	OP	Sw	Sw| h	Sw|¬h	OP
ε-G 0.2	0.317	0.258	0.353	0.153	0.146	0.325	0.121	0.163	0.189	0.245	0.164	0.138
ε-GD	0.236	0.173	0.309	0.205	0.39	0.259	0.392	0.239	0.211	0.179	0.25	0.159
WSLS	0.221	0.364	0.486	0.190	0.211	**0.079**	0.191	0.254	**0.104**	0.434	0.26	0.285
TS	**0.091**	0.121	0.140	0.137	0.210	0.318	0.21	0.076	0.124	**0.156**	**0.123**	**0.070**
IBL	0.109	**0.118**	**0.139**	**0.127**	**0.084**	0.347	**0.094**	**0.057**	0.136	0.163	0.164	0.152

Table 4: The distances of the predictions of individual predictors or IBL models from human data, calculated using RMSE metric. The measures we use are switching (Sw), switching after triggering a honeypot (Sw|h), switching after not triggering a honeypot (Sw|¬h) and optimal play (OP). Bold font indicates the lowest value in each column.

wards the end of the interaction. The predictor starts with $\varepsilon = 1$ and decreases it linearly towards $\varepsilon = 0$ at the end of the interaction, given a known finite horizon.

Thompson Sampling (TS): We follow the description of the TS algorithm as detailed by Agrawal and Goyal for Bernoulli Bandits (2012). We extend this version of the TS algorithm for the Bernoulli MAB by incorporating a support function $W_i(\theta_i)$ instead of selecting the action i with the maximum sample θ_i as described by Agrawal and Goyal. For this setting, we use a support function $W_i(\theta_i) = v_i \cdot \theta_i - c_i^a$ where $\theta_i \sim Beta(S_i + 1, F_i + 1)$ samples from a Beta distribution, thus the algorithm favors successes (S_i) over failures (F_i).

Instance-Based Learning: An IBL model (Gonzalez & Dutt, 2011) describes a learning attacker with an ability to recall and identify similar "instances" of past decisions using memory. An IBL instance represents a decision made in a specific situation, and the outcome feedback. The feedback here is the net payoff calculated as a difference between a successful and a failed attack, i.e., $v_i - 2c_i^a$. The IBL decision process has three main parameters: (1) decay, d, which specifies how past experiences are considered in current decisions based on time; (2) noise parameter σ, capturing random variability between experiences; and (3) the similarity, S, capturing the influence of the past on the present based on the similarity of the situations.

In the HCG game, an attacker can observe two possible outcomes of an attack on node i: a positive reward ($v_i - c_i^a$) when she attacks a real resource (success s_i) or a negative reward ($-c_i^a$) if the target is a honeypot (failure f_i). We denote an instance in memory representing a combination of situation, decision and outcome that was experienced in the past as $o(t') \in \bigcup_{i \in N} \{s_i, f_i\}$. In round t, an attacker targets a node i_t^* which maximizes a blended value (BV) as follows:

$$i_t^* \leftarrow \arg\max_{i \in N} BV_t(i) \tag{2}$$

$$BV_t(i) = (v_i - c_i^a) \frac{e^{A_t(s_i)}}{e^{A_t(s_i)} + e^{A_t(f_i)}} - c_i^a \frac{e^{A_t(f_i)}}{e^{A_t(s_i)} + e^{A_t(s_i)}} \tag{3}$$

$$A_t(o_i) = \ln \sum_{t' \in \{1,\ldots,t-1\}: o(t') = o_i} (t - t')^{-d} - \\ - S \sum_{i' \in N} (sim(i, i')) - \sigma \ln \frac{1-\gamma}{\gamma}, \tag{4}$$

where $\gamma \in (0, 1]$ is uniformly randomly sampled and *sim* is a similarity function. We used a linear similarity function that normalizes the net payoff from a decision based on the maximal payoff of 20 and is calculated as $sim(i, i') = 1 - |(v_i - 2c_i^a) - (v_{i'} - 2c_{i'}^a)|/20$.

We fit a separate IBL attacker model to human data when playing against each of the algorithmic defenders. We calibrated parameters values using exhaustive search over a wide range of values for each parameter with 350 repetitions for each combination. We used a multiobjective optimization minimizing average RMSE (see Equation 5) of all measures. The resulting three sets of parameters were: $(\sigma = 0.2, d = 0.1, S = 0.6)$ for the LLR defender, $(\sigma = 0.35, d = 1.2, S = 0.4)$ for the Pure defender and $(\sigma = 1.4, d = 0.5, S = 0.5)$ for the Equilibrium defender.

Simulation Results

To analyze the predictors' effectiveness in emulating human behavior we did a simulation with identical settings to the human experiment. Each predictor played against the 3 defenders in the same scenario 100 times. We consider the same performance measures as before. How well a predictor approximates human behavior is determined by a distance of a prediction $\{p\}_{t=1}^T$ from human data $\{hd\}_{t=1}^T$, calculated using the RMSE metric below where where m is a performance measure and T is a number of rounds.

$$RMSE_m(p, hd) = \sqrt{\frac{\sum_{t=1}^{50} (m(p_t) - m(hd_t))^2}{T}} \tag{5}$$

In Table 4, IBL accounted for the "most human" behavior on most of the measures when playing against the Pure and LLR defenders. In contrast, TS plays most closely to human performance when playing against the Equilibrium defender. This may be because the static equilibrium defender most closely reflects the standard stochastic MAB setting that TS was designed for. ε-Greedy, ε-Greedy Decreasing and WSLS perform poorly in general as predictors of human behavior.

However, these observations may only paint part of the picture. The actual overall point performance of human participants versus LLR is much lower than the 4 behavioral predictors as shown in Table 5. Nearly all 4 of the behavioral predictors double the median score of the human participants when facing the LLR defender. In contrast, the IBL model plays the most closely to human performance versus LLR. The IBL model comes rather close to the human data in relation to the

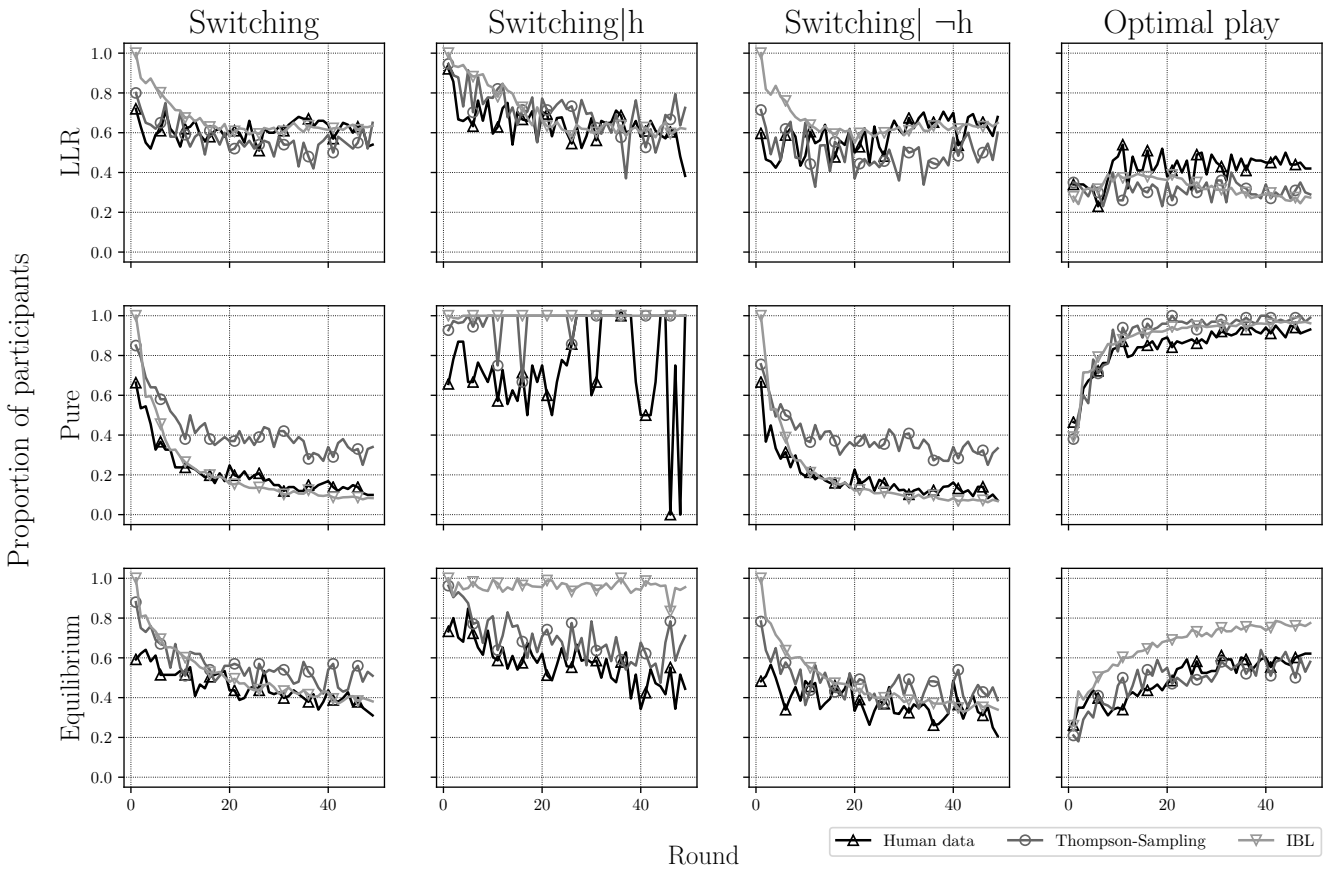

Figure 3: Comparison of the strategy predictions of TS and IBL with human data.

average and median scores. When considering all this information, it appears that the adaptive LLR defender exploited the human participants' learning mechanisms as well as IBL predicts. We can also see that humans may adopt different strategies depending on an opponent's strategy. Thus, when choosing a modeling approach there is a need to carefully select the granularity level at which predictions are needed: aggregate or individualized behavior. The IBL model can produce predictions at both levels.

	μ	σ	median	min	max
Human	172.6	123.02	160	-85	640
ε-G 0.2	303.9	140.3559	320	-75	640
ε-GD	265.1	99.55705	275	-115	480
TS	332	109.6275	330	90	585
WSLS	292.4	114.2686	287.5	35	590
IBL	198.9	193.44	220	-335	685

Table 5: Performance of predictors against the LLR defender in attacker points.

Conclusion

We study how humans learn in a novel version of an adversarial, contextual multi-armed bandit scenario motivated by a real-world cybersecurity scenario where defenders use deceptive decoys and attackers must learn to avoid them. We evaluated three different types of defensive strategies and showed that an adaptive defensive strategy was clearly the strongest against human players, and the hardest for them to learn. We also made novel comparisons between predictive models for emulating how humans learn in this type of adversarial setting, comparing leading models from both the MAB literature and cognitive science. We find that the best models (Thompson Sampling and IBL) are able to predict human behavior quite effectively, but that human attackers use different strategies depending on the adversary they are up against, and the best predictor may depend on this context. There are many interesting opportunities to improve both types of models especially in making personalized predictions for individuals and specialized context. However, the results so far have immediate practical implications for how we can design better strategies for deploying decoy systems to enhance cybersecurity. In particular, these systems must be adaptive to prevent attackers from easily learning the defensive strategy. The predictive models of attacker learning we have developed will also allow us to develop defenses that actively mitigate the ability of attackers to learn the defensive strategy.

Acknowledgements

This research was sponsored by the Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation here on. The authors also acknowledge the support of the OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/000 0765 Research Center for Informatics. The authors thank Orsolya Kovacs from the Dynamic Decision Making Laboratory at Carnegie Mellon University for her help with data collection.

References

Agrawal, S., & Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory* (pp. 39–1).

Gai, Y., Krishnamachari, B., & Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, *20*(5), 1466–1478.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

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

Gonzalez, C., & Dutt, V. (2016). Exploration and exploitation during information search and experimential choice. *Journal of Dynamic Decision Making*, *2*(1).

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

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Robbins, H. (1985). Some aspects of the sequential design of experiments. In *Herbert robbins selected papers* (pp. 169–177). Springer.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367.

Spitzner, L. (2003). *Honeypots: tracking hackers* (Vol. 1). Addison-Wesley Reading.

Sripa, B., Mairiang, E., Thinkhamrop, B., Laha, T., Kaewkes, S., Sithithaworn, P., ... Bethony, J. M. (2009). Advanced periductal fibrosis from infection with the carcinogenic human liver fluke opisthorchis viverrini correlates with elevated levels of interleukin-6. *Hepatology*, *50*(4), 1273–1281.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, *23*(5), 727–741.

Zhang, S., & Angela, J. Y. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (pp. 2607–2615).