Evaluating Models of Human Behavior in an Adversarial Multi-Armed Bandit Problem

Marcus Gutierrez[1], Jakub Černý[2], Noam Ben-Asher[3], Efrat Aharonov[4], Anjon Basak[1], Branislav Bošanský[2], Christopher Kiekintveld[5], and Cleotilde Gonzalez[6]

[1] University of Texas at El Paso, El Paso, TX, 79968, USA
{mgutierrez22,abasak}@miners.utep.edu
[2] Czech Technical University in Prague, Czech Republic
{jakub.cerny,branislav.bosansky}@agents.fel.cvut.cz
[3] Army Research Laboratory, Adelphi, MD 20783, USA
noam.ben.asher@gmail
[4] Carnegie Mellon University, Pittsburgh, PA, 15213, USA
eaharono@andrew.cmu.edu
[5] University of Texas at El Paso, El Paso, TX, 79968, USA
{cdkiekintveld}@utep.edu
[6] Carnegie Mellon University, Pittsburgh, PA, 15213, USA
coty@cmu.edu

Abstract. Developing good computational models of human behavior is an important aspect of designing autonomous agents that interact with humans. We consider the problem of predicting how humans learn interactively in an adversarial Multi-Armed Bandit (MAB) setting. Our domain is motivated by the use of deception in cybersecurity and the need to design effective decoys to lure attackers. We ran a behavioral study in which humans act as cyber attackers, and try to learn the (possibly randomized) defense strategy for assigning nodes in the network to be decoy targets over many interactions. We tested humans against three types of defensive strategies: a fixed strategy, a randomized mixed strategy based on a game-theoretic solution, and an adaptive strategy based on a bandit learning algorithm. Our results show that humans have the most difficulty learning against the adaptive defense, followed by the randomized equilibrium strategy. We also evaluated five different models for predicting how the human players learn to play against these defender strategies. We compare the predictive quality of these models using our experimental data, 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 and decision making in this adversarial domain.

Keywords: Cognitive Models · Behavioral Game Theory · Learning Agent Capabilities · Multiagent Learning

1 Introduction

With the increased popularity of autonomous systems, the question of how humans interact with these systems becomes increasingly important to the design of these systems, including the problem of how to secure them against cyber attackers. Humans are imperfect agents, but they are capable of fast learning in some settings and 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. In this work, we focus on comparing different computational models that attempt to capture and emulate human decision-making when interacting in an adversarial Multi-Armed Bandit (MAB) setting. The MAB is a decision making paradigm studied both within the machine learning community and the cognitive modeling community, where it is used to study how humans learn in probabilistic settings with feedback and uncertainty.

Here, we are motivated by a class of deceptive interactions that arise in cybersecurity when network defenders deploy decoys in the network (i.e., honeypots) to detect and thwart attackers. Honeypots are designed to be attacked to waste the attacker's resources and provide information to the defender [15]. Attackers wish to avoid detection by these honeypots by interacting only with real systems in the network. Simply deploying a collection of honeypots and never changing their configurations (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 learning dynamic should be able to deploy defensive strategies that are harder to learn and defeat over the long term.

We model this scenario as a repeated adversarial interaction between a network defender and attacker. The defender may incur a cost to protect some network resources using honeypots, and the attacker tries to maximize the value of the attacked resource while avoiding honeypots. We are interested in particular in how the attacker may be able to learn the defender's deception strategy and avoid honeypots based on previous experience. Formally, this scenario has many similarities to an adversarial version of a multi-armed bandit problem, though there are some differences from standard MAB formulations.

We conducted an experiment where each human participant plays the role of the attacker and tries to maximize her profits over numerous rounds against one of 3 algorithmic defenders of varying complexity. The *static pure* defender selects the same honeypot configuration every round. This defender acts as our baseline for how quickly humans can learn in a static environment. The *static equilibrium* defender plays a fixed distribution over all honeypot configurations, which maps most closely to the stochastic MAB problem. Lastly, the *adaptive Learning with Linear Rewards* (LLR) defender plays a dynamic strategy that reacts to the attacker's moves. Each defender provides a unique learning challenge for the human attackers and we wish to observe the participants' behavioral differences that arise due to each defender.

We then analyze the predictions of 4 behavioral models and 1 cognitive model that attempt to emulate human behavior. The behavioral models we investigate originate from the MAB literature and attempt to address the explore-exploit dilemma directly, ideally as a human would. Meanwhile, the cognitive models take an indirect approach of solving the dilemma by modeling the same cognitive processes that are stimulated in human decision making, such as memory activation, recency biases, and frequency biases. We investigate the effectiveness of these models when confronted with varying levels of defense complexities.

We analyze the effectiveness of each predictive model in capturing the human data using a variety of measures, including switching among different nodes (i.e., a measure of exploratory behavior), conditional switching (i.e., switching when winning or losing a round), and the proportion of optimal play (i.e., how often the player selects the best action). Using these measures, we show that a modified version of Thompson Sampling [2] and a cognitive model based on Instance-Based Learning (IBL) Theory [10] best emulate the dynamic behavior of human participants in this adversarial environment. By understanding the processes that drive human decision-making in the presence of uncertainty, and developing predictive human-like models, we move closer to developing automated defensive agents. Such models could be used to defeat sophisticated, learning adversaries, including other humans.

2 Honeypot Cyber Deception Domain

We designed a model that focuses on the learning aspects of an adversarial cybersecurity interaction, motivated by honeypot deception. In this scenario, an attacker and defender compete over multiple resources (nodes) in the network belonging to the defender with the following parameters: 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. At the beginning of the interaction, each node is initialized with the non-negative parameters. At the beginning of each round, the defender spends some budget D to turn some subset of the nodes into honeypots, such that the total cost of defended nodes 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 undefended, the attacker receives the reward $v_i - c_i^a$, and the defender receives a reward of 0. On the other hand, 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[7]. At the end of the round, the interaction resets, and the process repeats each round.

The only feedback the attacker receives is the reward for her action. Therefore, the attacker can only partially and indirectly observe the defender's behavior. The defender observes the individual honeypot placements. So, if the defender captures the attacker with a honeypot, the defender knows which honeypot node was responsible for the capture. If the defender does not capture the

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

attacker and there are more than 1 undefended nodes, the defender can never be certain about which node the attacker chose. This style of feedback is known as semi-bandit feedback in the MAB literature. Given our focus on the study of high-level decision-making, only general cognitive skills are needed from those humans playing the role of the attacker. Cybersecuriy knowledge does not play a role in making decisions regarding "honeypot" configuration or realism [3].

3 Learning in Multi-Armed Bandits

This cyber deception scenario maps approximately to a decision making paradigm known as the Multi-Armed Bandit (MAB), which have been investigated extensively in both the literature on human learning and machine learning agents. In a MAB, individuals (agents) learn by repeatedly choosing among multiple options (arms), each of which is associated to a probability of reward, observed through immediate feedback after a choice. In theories of decisions from experience, two-arm bandit problems are a classical example of research paradigms used for modeling human decisions and learning from experience (e.g., [8]). More broadly, MAB problems are conceived of as reinforcement learning strategies where a decision maker wishes to optimize her profit over repeated interactions by selecting different arms. MAB tasks have been very useful in the study of human decision making, characterizing the common exploration-exploitation tradeoff (e.g., [17]).

One such MAB solution, Thompson Sampling (TS), a strategy that maximizes the expected reward from a randomly drawn belief, is commonly used to model the exploration-exploitation dilemma in MAB problems. However, it is unclear how TS and other models actually capture the way humans learn in MAB tasks, and whether they capture human's sensitivity to exploration and exploitation tradeoffs (e.g., [14]). To answer this question, we analyze human behavior against the predictions of these models.

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 [9, 12]. Sripa et al. notably ran an experiment with 451 human participants playing the MAB [16], and applied a Bayesian learning model to explain the human data. Zhang et al. extend this work by improving the participant behavioral prediction with a Knowledge Gradient model [20]. 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.

The MAB has been well-studied in recent years for its generalizability. In context of our cyber deception model, the MAB has multiple applications to model robust attackers and resilient defenders. In the Experimental Design section, we discuss an algorithmic defender, adaptive LLR, that was developed in the context of a combinatorial and stochastic MAB domain. Meanwhile, Burtini et al. provides an in-depth survey on the MAB literature [5].

In contrast to these and other models often used in MAB tasks (e.g., TS), cognitive models of human behavior are less common. Cognitive models represent the cognitive mechanisms (e.g, memory, learning, forgetting) which are essential elements for human learning [10]; however, these models have not been yet tested in adversarial MAB settings. In this research we offer a unique context to test cognitive models of human learning and decision making, and pair them against models often used in MAB tasks (e.g., TS). The adversarial situation in which learning takes place provide insights on how humans learn against algorithms of variable dynamics and adaptability.

4 Experimental Design

We designed a repeated adversarial interaction with 6 arms (which we will refer to as nodes) to be played over 50 rounds as seen in Figure 1. We recruited 304 human participants on Amazon's Mechanical Turk [4]. Of the 304 participants, 130 reported female and 172 reported male with 2 participants reporting as other.

All participants were above the age of 18 and had a median age of 32. The experiment averaged roughly 10 minutes from start to finish and the participants were paid US $1.00 for completing the experiment. The participants 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 bonus payment was intended to incentivize participants to play as best they could.

In a realistic cybersecurity environment, the domain knowledge of the attacker plays an important role as to which vulnerabilities to exploit and how to to gain access to a system. When recruiting the participants in our study, we held no requirements or assumptions about the cybersecurity knowledge of the participants. To address this, we take the pessimistic assumption that if the participant (as the attacker) tries to attack a non-honeypot node, they perform a guaranteed successful attack. Taking this approach, we elevate all of our recruited participants to the level of expert hackers. All that remains for the attacker is deliberate which node to target for an attack, which boils down to basic human cognition. Real world expert hackers will share the same level of cognition with our participants, allowing the recruited participants to accurately represent real cyberattackers at the described level of abstraction [3].

4.1 Scenario

The defender receives 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. For ease of the participant, we simplified the visible rewards on the nodes, where the reward $v_i - c_i^a$ for attacking a non-honeypot appears as the positive top number in the

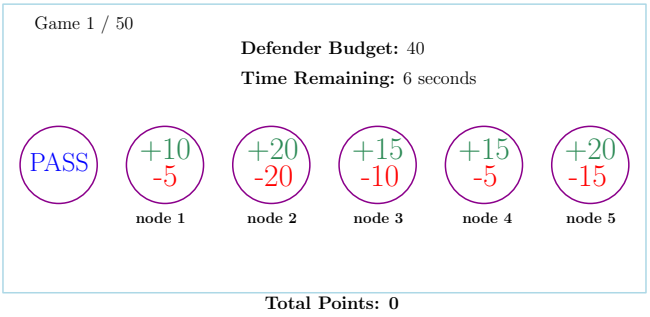

Fig. 1: User interface for the Honeypot Intrusion Game.

node. Meanwhile, the loss for attacking a honeypot $-c_i^a$ appears as the lower, negative number inside the node. Table 1 shows the actual parameters for each node.

We designed the nodes such to fit a variety of risk-reward archetypes (e.g., low-risk/low-reward, high-risk/high-reward, low-risk/high-reward). The intuition is to allow for differences in strategies and learning. These differences in known parameters provide a noticeable difference from the traditional MAB. For instance, in the first round, the attacker is making an informed decision based on the attack costs and rewards.

	pass	node 1	node 2	node 3	node 4	node 5
v_i	0	15	40	25	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.

4.2 Behavioral Measures

For the analysis for how humans learn and adapt to their environment and opponent, we look at 4 measures associated with their behavioral performance. Later on we compare predictive algorithms using the same measures. We primarily consider switching actions as a behavioral heuristic, which have been used by humans to make decisions in dynamic and complex environments [18].

Switching: This common measure has been well studied with regards to human decision-making and learning [9]. High switching indicates high exploration and low switching indicates exploitation in the case of a static defender and static environment.

Switching with Honeypot: Here, we look at the case where the decision maker switches nodes to attack after triggering a honeypot (i.e., receiving a negative reward). This corresponds with the "Lose-Shift" aspect of Win-Stay-Lose-Shift [13], a common strategy used in economics.

Switching without Honeypot: In this case, we look at where the decison maker switches nodes after attacking a real node (i.e., receiving a positive reward). This opposes the "Win-Stay" aspect of the Win-Stay-Lose-Shift strategy (i.e., "Win-Shift").
Proportion of Optimal Play: Here, we examine the actual per-round performance of the decision maker. We define optimal play for the attacker as attacking a node that provided the highest expected reward. When facing a static defender in a static environment, the optimal node(s) will remain the same. Versus a dynamic defender, the node(s) that provide the highest expected value may change from round to round.

4.3 Defenders

For the experiment, we deployed 3 different defenders to analyze the impact a dynamic defense has on human learning. We utilize the *Static Pure Defender*, the *Static Equilibrium Defender*, and the *Adaptive LLR Defender* that learns from its own action observations. Each defender creates a different level of learning complexity for the human participants. In this study, we are not investigating the best defense strategy versus humans. Instead, we are interested in analyzing the impact that varying levels of defense complexity have on human learning and decision making.

Static Pure Defender: This defender employs a "set and forget," purely static defense that attempts to maximize its total value by assuming it must commit to a single, pure, non-stochastic strategy for a single round. This defender tries to spend its budget to protect the highest valued nodes. For the scenario seen in Figure 1, the defender always defends nodes 2 and 5, leading to nodes 3 and 4 as optimal nodes for attacking. This defender acts as a baseline for human learning. With this defender, we investigate the upper bound on how quickly humans can learn a defense. 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. 101 total human participants played versus this defender.

Static Equilibrium Defender: This defender plays over a fixed distribution of defenses (combinations of nodes to be honeypots). The defense is a Mixed Strategy Nash Equilibrium. It optimizes the defender's expected utility assuming only a single, non-repeated interaction against a fully rational attacker. An optimal strategy of the attacker in this equilibrium is to attack node 4. Playing optimally leads to an expected total value of ≈ 447 points for the attacker. 103 total human participants played versus this defender.

Adaptive Learning with Linear Rewards Defender: The last defender, known as Learning with Linear Rewards (LLR) [7], provides a deterministic, yet adaptive defense as it tries to maximize its reward by balancing exploration and exploitation.

Algorithm 1 describes the LLR algorithm in effect where $\mathcal{A}_a$ defines the set of all individual basic actions (nodes to defend). In the described scenario from Figure 1, $\mathcal{A}_a$ is the set containing all 5 nodes. LLR uses a learning constant L, which we set to $L = 3$ for our scenario since this is the maximum number of

Algorithm 1 Learning with Linear Rewards (LLR)

1: //INITIALIZATION
2: If $\underset{a}{max}|\mathcal{A}_a|$ is known, let $L = \underset{a}{max}|\mathcal{A}_a|$; else, $L = N$
3: **for** $t = 1$ to N **do**
4: Play any action a such that $t \in \mathcal{A}_a$
5: Update $\left(\hat{\theta}_i\right)_{1\times N}$, $\left(m_i\right)_{1\times N}$ accordingly
6: **end for**
7: //MAIN LOOP
8: **for** $t = N + 1$ to ∞ **do**
9: Play an action a which solves the maximization problem

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

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

nodes we can play in a defense. LLR then 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 arm i has been played. After each round these vectors are updated accordingly.

After the initialization phase, LLR solves the maximization problem seen in equation 1 and deterministically selects the subset of nodes that maximizes the equation each round until the end of the game. This deterministic nature of LLR indirectly adapts to the attacker's moves. It has no concept of an opponent, but instead is trying to balance between nodes with high observed means (i.e., have captured the attacker often in the past) and less frequently played nodes (which the attacker may move to in order to avoid capture). We say that LLR indirectly adapts to the attacker's actions.

In this scenario, the attacker can never fully exploit the deterministic strategy of the defender because of the partial observability aspect of the interaction. This defense leads to the optimal node(s) changing in each round as adaptive LLR learns. 100 total human participants played versus this defender.

5 Behavioral Results

The analyses of human data results in clear performance patterns among the 3 defenders. The pure defender predictably performed the worst, yielding an average score of 611.93 points to the human attackers, just over 100 points short of the maximum possible points achievable against the pure defender. Next, the equilibrium defender performed significantly better, yielding only an average of 247.81 points to the human attackers, a full 200 points short of the maximum expected points achievable by a human attacker. Finally, the LLR stood as the

most resilient defender versus the human attackers with an average of 172.6 points yielded to the participants. Table 2 shows the aggregate statistics of the human attacker performance.

	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 2: Aggregate data of participants' end-game attacker rewards.

Figure 2 shows the cumulative utility frequencies among the participants. Visually, we can see the pure defender yielded many points to many participants with only a couple of outliers. The participant who received -375 points vs the static pure defender likely tried to speed through the game without trying to optimize total reward.

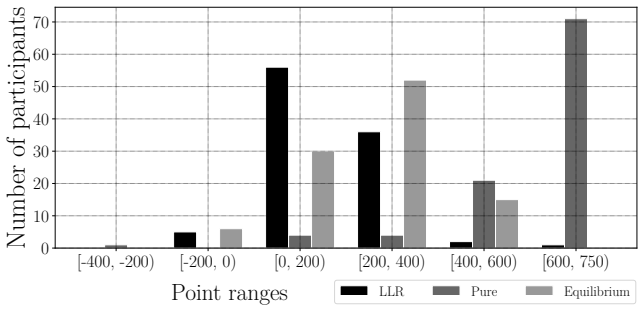

Fig. 2: Frequencies of total cumulative utility ranges.

LLR is designed to solve a combinatorial MAB problem where it assumes a static, stochastic environment. Here, LLR does not take into account the attacker's adaptive nature. We do not make any claims that LLR stands as a perfect adversary to human decision-making. However, we do note that LLR, a deterministic yet adaptive strategy, outperforms the static equilibrium defense.

When analyzing the proportion of the human population that played optimally per round, we can see the differences in learning curves among the various defenders. The rightmost graph in Figure 3 shows the frequency of optimal decisions over the course of the 50 rounds.

In Figure 3, we note that participants playing against the static pure defender learn very early on 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, and a significant advantage of the LLR only emerges after at least 20 practice rounds.

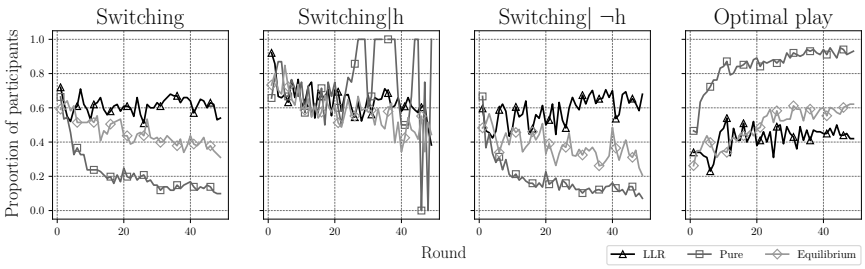

Fig. 3: The proportions of participants switching nodes to attack or playing optimally over time. The high switching after triggering a honeypot seen in round 26 from participants facing the static pure defender is a small portion of the population as less than 10% of the population were triggering honeypots past round 25.

We also investigate another essential human decision-making measure: switching [11]. This measure will help explain the human participants understanding of the defenders' strategies and provide insight into exploratory processes. How the human attackers change their switching patterns over time can help explain how confident they are in their understanding of the defender's strategy. We observe in the leftmost graph in Figure 3 that the overall proportion of switching decreases over time, particularly when participants face the static pure defender; meanwhile, when the participants face the adaptive LLR defender, they seem to stay in a high proportion of switching over the course of the 50 rounds.

The middle left graph in Figure 3 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. Because adaptive LLR improves its beliefs about a node's expected payoff after playing it, after capturing the attacker, the capturing node will more likely be selected in future rounds. Because of 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 experimental scenario, the attacker should always attack node 4, regardless of triggering a honeypot or not.

In Figure 3 (middle right), the effects of the defender on the participants' switching behavior show more differences when the attackers did not trigger a honeypot (i.e., received a positive reward). Concerning the static pure defender and static equilibrium defender, decrease in switching demonstrate a move towards a more exploitative strategy and understanding of the static defense. On the other hand, participants who competed against LLR, maintained high

switching throughout the entirety of the game. In general, adpative 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.

6 Predictors

In this section, we examine the ability of 4 behavioral predictor models to emulate participants' performance in the experiment, and compare them to an IBL cognitive model [10]. Predictive models that explain human decision-making and learning processes in adversarial environments can inform us about the underlying mechanisms that influence decision making as well as provide predictions. Such predictor models can support the development of better defenders that hinder human attacker learning in cybersecurity settings. We utilize these behavioral predictors as they have been known to capture human performance in numerous MAB settings [16, 20, 2].

The 5 predictor models we investigate in this paper are not an exhaustive representation of the behavioral models or cognitive models in the literature. We selected these models as they have all shown promise in predicting human performance in classic MAB settings. However, not all of our defenders accurately map to the traditional MAB problem and so other, more complex models with more parameters may struggle when MAB assumptions are broken (e.g., non-stochasticity). In future work, we will investigate more predictive models for our current and future defenders.

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 $\epsilon \in \{0, 1\}$. With probability ϵ, ϵ-Greedy attacks uniform randomly (exploration) and with probability $(1 - \epsilon)$, 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 towards the end of the interaction. The predictor starts with $\epsilon = 1$ and decreases it linearly towards $\epsilon = 0$ 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 [2]. 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. For this particular scenario, we use $W_i(\theta_i) = v_i \cdot \theta_i - c_i^a$ where $\theta_i \sim Beta(S_i+1, F_i+1)$ samples from a Beta distribution where α is the number of positive rewards S_i

and β is the number of negative rewards F_i. By utilizing this Beta distribution, this will favor successes more than failures.

	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 3: The distances of the predictions of individual predictors or IBL models from human data, calculated using RMSE metric. The used measures 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.

Instance-Based Learning: An IBL model [1, 8, 19] describes a learning attacker with a specific memory-recalling and similarity-identifying ability of "instances" in memory. An instance in IBL is a representation of declarative knowledge, including the following components:
Situation capturing the contextual attributes of node i relevant to the attacker.
Decision representing the choice to attack node i_t^* which is expected to yield the maximal outcome
Feedback the consequences of choice execution is captured by the outcome received from the environment.

An IBL instance provides a unified representation of a decision made in a specific situation, and its outcomes. In this study the feedback is the net payoff calculated as a difference between a successful attack 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 past experiences on current decision based on the similarity between the situations.

In the Honeypot 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 $(-c_i^a)$ in case the target of the attack 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:

$$i_t^* \leftarrow_{i \in N} BV_t(i) \tag{2}$$

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

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

where $\gamma \in (0, 1]$ is a uniformly randomly sampled real number 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 $sim(i, i') = 1 - |(v_i - 2c_i^a) - (v_{i'} - 2c_{i'}^a)|/20$.

A separate IBL attacker model was fitted to human attacker data when playing against one of the algorithmic defenders. Calibration of parameters values used exhaustive search over a wide range of values for each parameter with 350 repetitions for each combination. Calibration used a multiobjective optimization minimizing average RMSE (see Equation 5) of all measures. The resulting three sets of parameters are: $(\sigma = 0.2, d = 0.1, S = 0.6)$ 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.

6.1 Simulation Results

To analyze the predictors' effectiveness in adequately emulating human behavior, we developed a simulation with identical settings to the Mechanical Turk experiment. Each predictor played against each of the 3 defenders in the same scenario 100 times. In this section, we look at the same performance measures found in the results section, specifically the proportion of optimal play and switching. 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 a RMSE metric as

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

where m is a performance measure and T is a number of rounds (in our case, 50).

In Table 3, IBL was able to account for human behavior the best on most of the measures when playing against Pure and LLR algorithmic defenders. In contrast, TS plays most closely to human performance when playing against an Equilibrium defender. This makes sense as the static equilibrium defender most closely reflects the standard stochastic MAB setting that TS was designed for. ϵ-Greedy, ϵ-Greedy Decreasing and WSLS seem to perform poorly in general.

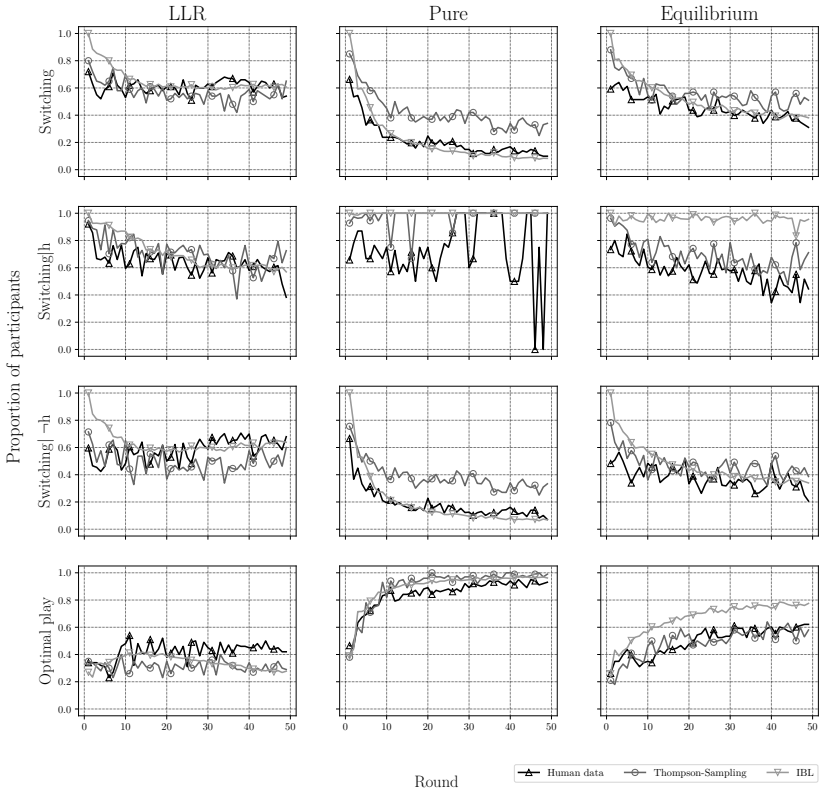

Fig. 4: The proportions of participants switching nodes to attack or playing optimally over time against different defenders according to human data, predicted by TS and predicted by fitted IBL models.

Examining parameters of the IBL attacker models playing against the static pure and adaptive LLR defenders indicates that the values differ mainly in the decay of instances in memory. This suggests that when facing a static pure defender, the IBL Fmodel tends to incorporate more experiences from the past while the IBL model that confronts an adaptive LLR defender pays more attention to recent experiences rather than relying on distant experiences. This observation is in agreement with recent findings from human experiments, suggesting that humans that best adapt to changing conditions of an environment generally have "poor memory" (e.g., high decay) compared to participants that show more "sticky" or less adaptable behavior [6]. This is because in dynamic situations the most representative conditions of the near future is the near past, and humans adapt their memory according to the dynamics of the environment [10].

However, these predictor observations may only paint part of the picture. For example, when confronting the adaptive LLR defender, the actual overall point performance of human participants is much lower than the 4 behavioral

	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 4: The aggregate points statistics of individual predictors, IBL model or humans against LLR defender.

predictors as shown in Table 4. Nearly all 4 of the behavioral predictors double the median score of the human participants when facing the LLR defender. The IBL model, however, play the most closely to human performance versus the adaptive LLR defender. The IBL model comes rather close to the human data in relation to the 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: overall, over-time behavior or individualized behavior. The IBL model is able to produce predictions at all the three levels.

7 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.

References

1. Abbasi, Y., Kar, D., Sintov, N., Tambe, M., Ben-Asher, N., Morrison, D., Gonzalez, C.: Know your adversary: Insights for a better adversarial behavioral model. In: Conference of Cognitive Science Society (CogSci) (2016)
2. Agrawal, S., Goyal, N.: Analysis of thompson sampling for the multi-armed bandit problem. In: Conference on Learning Theory. pp. 39–1 (2012)
3. Ben-Asher, N., Gonzalez, C.: Effects of cyber security knowledge on attack detection. Computers in Human Behavior **48**, 51–61 (2015)
4. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on psychological science **6**(1), 3–5 (2011)
5. Burtini, G., Loeppky, J., Lawrence, R.: A survey of online experiment design with the stochastic multi-armed bandit. arXiv preprint arXiv:1510.00757 (2015)
6. Cheyette, S., Konstantinidis, E., Harman, J.L., Gonzalez, C.: Choice adaptation to increasing and decreasing event probabilities. In: 38th Annual Meeting of the Cognitive Science Society (CogSci 2016), Philadelphia, PA (2016)
7. Gai, Y., Krishnamachari, B., Jain, R.: 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 (2012)
8. Gonzalez, C., Dutt, V.: Instance-based learning: Integrating sampling and repeated decisions from experience. Psychological review **118**(4), 523 (2011)
9. Gonzalez, C., Dutt, V.: Exploration and exploitation during information search and experimential choice. Journal of Dynamic Decision Making **2**(1) (2016)
10. Gonzalez, C., Lerch, J.F., Lebiere, C.: Instance-based learning in dynamic decision making. Cognitive Science **27**(4), 591–635 (2003)
11. Inman, J.J., Zeelenberg, M.: Regret in repeat purchase versus switching decisions: The attenuating role of decision justifiability. Journal of consumer research **29**(1), 116–128 (2002)
12. Mehlhorn, K., Newell, B.R., Todd, P.M., Lee, M.D., Morgan, K., Braithwaite, V.A., Hausmann, D., Fiedler, K., Gonzalez, C.: Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision **2**(3), 191 (2015)
13. Robbins, H.: Some aspects of the sequential design of experiments. In: Herbert Robbins Selected Papers, pp. 169–177. Springer (1985)
14. Speekenbrink, M., Konstantinidis, E.: Uncertainty and exploration in a restless bandit problem. Topics in cognitive science **7**(2), 351–367 (2015)
15. Spitzner, L.: Honeypots: tracking hackers, vol. 1. Addison-Wesley Reading (2003)
16. Sripa, B., Mairiang, E., Thinkhamrop, B., Laha, T., Kaewkes, S., Sithithaworn, P., Tessana, S., Loukas, A., Brindley, P.J., Bethony, J.M.: 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 (2009)
17. Steyvers, M., Lee, M.D., Wagenmakers, E.J.: A bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology **53**(3), 168–179 (2009)
18. Todd, P.M., Gigerenzer, G.: Précis of simple heuristics that make us smart. Behavioral and brain sciences **23**(5), 727–741 (2000)
19. Zhang, C., Jiang, A.X., Short, M.B., Brantingham, P.J., Tambe, M.: Defending against opportunistic criminals: New game-theoretic frameworks and algorithms. In: International Conference on Decision and Game Theory for Security. pp. 3–22. Springer (2014)

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