Modeling Cognitive Biases in Decision-Theoretic Planning for Active Cyber Deception

Aditya Shinde
THINC Lab, School of Computing
The University of Georgia
Athens, USA
adityas@uga.edu

Prashant Doshi
THINC Lab, School of Computing
The University of Georgia
Athens, USA
pdoshi@cs.uga.edu

ABSTRACT

This paper presents an approach to modeling and exploiting cognitive biases of cyber attackers in planning for active deception. Sophisticated cyber attacks are primarily orchestrated by human actors. Hence, we focus on the human aspect of the attacker's decision-making process. Humans deviate from rational decision-making due to various cognitive biases. Here, we focus on *fundamental attribution error* (FAE) and *confirmation bias* and their role in cyber deception because these biases contribute to humans being deceived. We use the decision-theoretic planning framework of finitely-nested factored I-POMDP (I-POMDP$_\chi$), which allows us to explicitly model FAE in multi-agent settings and build cognitive models of the attackers. We show how these biases impact their beliefs as they act and obtain more information about the environment and the adversary. The tractability of the I-POMDP$_\chi$ also allows for modeling agents at a higher strategy level where the optimal policy relies on induction and exploitation of these biases. Hence, we also present an I-POMDP$_\chi$-based rational defender agent that can model the attacker's beliefs under the influence of FAE and confirmation bias from a higher strategic level, and exploit them. Our experiments in simulated interactions show that the I-POMDP$_\chi$-based defender agent can induce FAE in an attacker to distort the attacker's beliefs. Consequently, the defender agent can exploit the attacker's cognitive biases to extend the duration of the attack to facilitate the attacker's intent recognition in a controlled environment. Our work provides a general decision-theoretic formulation of FAE and confirmation bias, and demonstrates its role in planning for agent-based active cyber deception.

KEYWORDS

Multi-agent decision-making; Cybersecurity; Computational modeling; Cognitive biases

ACM Reference Format:
Aditya Shinde and Prashant Doshi. 2024. Modeling Cognitive Biases in Decision-Theoretic Planning for Active Cyber Deception. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Realistic cyber attacks are rarely a single-step process. The initial intrusion, which remains a focus of many efforts in cyber defense research, only serves to establish a foothold in the target. Thereafter, attackers often follow a sequence of steps to achieve their objectives. These may include information gathering, escalating privileges, locating the target, and finally, causing business impact or gaining a strategic advantage. The diverse sets of tools and techniques that attackers use to perform these are well documented in the widely adopted MITRE ATT&CK matrix [26]. As such, sophisticated attacks are a multi-step process and tend to occur over an extended duration. We may model this extended attack interaction as a sequential decision-making problem between the attacker and defender. Defenders can utilize various techniques to disrupt, misdirect and isolate attackers to achieve defense goals such as minimizing the impact and learning about the attackers to preclude future attacks. Cybersecurity has begun to use deception (e.g., honey pots) as a central tactic to detect intruders and, more recently, to learn more about the attackers' preferences, capabilities, and motives [25]. Much research in AI-based cyber defense has adopted a game-theoretic perspective toward designating hosts as honey pots [3, 11, 14, 23].

AI-based cyberdeception commonly ascribes *rational* behavior to attackers. However, many cyberattacks are primarily orchestrated by human actors. It is well-known that human decision-making deviates from rational behavior due to the influence of cognitive biases [16]. Hence, in this paper, we focus on the human elements of the attacker's decision-making process, which may lead to suboptimal behavior due to cognitive biases. Specifically, we model the effects of *fundamental attribution error* (FAE) and *confirmation bias* on the attacker's beliefs. Previous work has shown that these biases play a role in humans being deceived: recent work on a game-theoretic integration of FAE studies the role it plays in misattributing agent behavior due to lack of context awareness [9]. Similar efforts were recently made to understand the psychological and cognitive impacts of cyber deception on human attackers [12]. We take a model-based approach and present a decision-theoretic formulation of select cognitive biases such as the FAE and confirmation bias that abet human deception. We then use this model to simulate cyber attack scenarios and study how the biases skew the attacker's beliefs. Importantly, we further show how this effect can be exploited to enhance deception.

A recent multi-agent based approach to interactive cyber deception applies the factored I-POMDP framework, labeled as I-POMDP$_\chi$, to model the interaction between an attacker and a

defender [25]. The defender agent modeled as an I-POMDP$_\mathcal{X}$ adaptively utilizes decoys and aims to deceive the attacker in order to extend the interaction, which facilitates inferring the attacker's intent. I-POMDP$_\mathcal{X}$ allows explicit modeling of the opponent and enables recursive reasoning about the opponent's actions. Hence, we adopt this framework and extend it with *cognitive models* of the opponent to study the effects of FAE and confirmation bias. The I-POMDP$_\mathcal{X}$ formulation gives us a subjective view of the interaction and enables reasoning about the biased attacker's beliefs. Also, as I-POMDP$_\mathcal{X}$ is scalable, the interaction is modeled at higher strategic levels where the defender reasons about the attacker's FAE and strategically exploits it during the interaction to deceive the attacker.

2 BACKGROUND

We consider a host-based perspective of the interaction between the attacker and the defender. In this section, we briefly review host-based cyber deception modeled using I-POMDP$_\mathcal{X}$.

2.1 I-POMDP$_\mathcal{X}$ for Cyberdeception

Interactive POMDPs (I-POMDPs) are a generalization of POMDPs to sequential decision-making in a multi-agent environment [6, 15]. I-POMDP$_\mathcal{X}$, introduced previously [25], is a factored variant of the I-POMDP framework and utilizes algebraic decision diagrams (ADDs) [2] to represent the factors for the agents' transition, observation, and reward functions compactly. This representation allows it to scale to problems with large state and observation spaces. Formally, an I-POMDP$_\mathcal{X}$ for agent i in an environment with one other agent j is defined as,

$$\text{I-POMDP}_\mathcal{X} = \langle \mathcal{IS}_i, A, T_i, \mathcal{Y}_i, O_i, \mathcal{R}_i, OC_i \rangle$$

where $\mathcal{IS}_i$ is the factored interactive state space consisting of physical state factors $\mathcal{X}$ and agent j's models M_j, which may be intentional or subintentional [5]. In a finitely-nested I-POMDP$_\mathcal{X}$, the set M_j is bounded and constructed similarly to finitely-nested I-POMDPs. The action set A is the set of joint actions of both agents. T_i defines the transition function represented using ADDs as $P^{a_i}(\mathcal{X}'|\mathcal{X}, A_j)$ for all $a_i \in A_i$. As a consequence of the *model non-manipulability* assumption in I-POMDPs, the transition function is defined over the physical states and does not include the other agent's models. Set $\mathcal{Y}_i$ contains the variables that constitute an observation. O_i is the observation function represented using ADDs as, $P^{a_i}(\mathcal{Y}_i'|\mathcal{X}', A_j)$. As a consequence of the *model non-observability* assumption, the observation function directly informs only about the physical state space. $\mathcal{R}_i$ defines the reward function for agent i, which is also represented as an ADD, $\mathcal{R}^{a_i}(\mathcal{X}, A_j)$. The reward function in I-POMDPs assigns rewards to physical states and actions. OC_i is the agent's optimality criterion, which may be a finite horizon H or a discounted infinite horizon where the discount factor $\gamma \in (0, 1)$.

The I-POMDP$_\mathcal{X}$ based agent i recursively updates the beliefs of agent j. This recursive belief update is similar to that defined for I-POMDPs but with ADDs compactly representing the factors. The I-POMDP$_\mathcal{X}$ belief update is computed as:

$$b_i^{a_i,o_i}(\mathcal{X}', M_j') = \sum_{\mathcal{X}, M_j} b_i(\mathcal{X}, M_j) \times P^{a_i,o_i}(\mathcal{X}', M_j'|\mathcal{X}, M_j) \quad (1)$$

where the ADD $P^{a_i,o_i}(\mathcal{X}', M_j'|\mathcal{X}, M_j)$ represents the transition probabilities for all interactive state variables given action a_i and observation o_i. The ADD is computed as:

$$\begin{aligned} P^{a_i,o_i}(\mathcal{X}', M_j'|\mathcal{X}, M_j) &= \sum_{A_j, \mathcal{Y}_j'} P^{a_i,o_i}(\mathcal{Y}_j', M_j', \mathcal{X}', A_j|M_j, \mathcal{X}) \\ &= \sum_{A_j} P(A_j|M_j)\ P^{a_i}(\mathcal{X}'|\mathcal{X}, A_j) \times \sum_{\mathcal{Y}_j'} P^{a_i}(\mathcal{Y}_i'|\mathcal{X}', A_j) \\ &\quad \times P^{a_i}(M_j'|M_j, \mathcal{Y}_j', \mathcal{X}', A_j). \end{aligned} \quad (2)$$

Here, the ADD $P^{a_i}(M_j'|M_j, \mathcal{Y}_j', \mathcal{X}', A_j)$ represents the recursive belief update transition $\tau_{\theta_j'}(b_j, a_j, o_j', b_j') \times O_j(s, a_i, a_j, o_j')$ of the original I-POMDP. ADD $P(A_j|M_j)$ is obtained by recursively solving the opponent's I-POMDP$_\mathcal{X}$.

Recent work on I-POMDP solution techniques utilizes Monte Carlo Tree Search for scalable approximate solutions [17, 24]. These methods, while being scalable, enumerate states and do not operate on factored representations that we utilize to model the domain. Interactive point-based value iteration offers a scalable alternative that utilizes a point-based approximation technique for solving finitely-nested I-POMDPs [7]. The I-POMDP$_\mathcal{X}$ leverages the compact ADD representations for factors along with this scalable point-based approximation to obtain the subject agent's policy. Toward this, the I-POMDP$_\mathcal{X}$ utilizes a factored representation of α-vectors. Subsequently, the backup equation is generalized using factored representation and ADD operations as follows:

$$\begin{aligned} \Gamma^{a_i,*} &\leftarrow \alpha^{a_i,*}(\mathcal{X}, M_j) = \sum_{A_j} R^{a_i}(\mathcal{X}, A_j) P(A_j|M_j) \\ \Gamma^{a_i,o_i} &\overset{\cup}{\leftarrow} \alpha^{a_i,o_i}(\mathcal{X}, M_j) = \gamma \sum_{\mathcal{X}', M_j'} P^{a_i,o_i}(\mathcal{X}', M_j'|\mathcal{X}, M_j) \\ &\quad \times \alpha^{t+1}(\mathcal{X}', M_j'), \forall \alpha^{t+1} \in \mathcal{V}^{t+1} \\ \Gamma^{a_i} &\leftarrow \Gamma^{a_i,*} \oplus_{o_i} \arg\max_{\Gamma^{a_i,o_i}} (\alpha^{a_i,o_i} \cdot b_i), \\ \mathcal{V}^t &\leftarrow \arg\max_{\alpha^t \in \bigcup_{a_i} \Gamma^{a_i}} (\alpha^t \cdot b_i), \quad \forall b_i \in B_i. \end{aligned} \quad (3)$$

Here, $\mathcal{V}^{t+1}$ is the set of α-vectors from the next time step and b_i is a belief point from the set of considered beliefs B_i. A popular way of building B_i is to project an initial set of belief points forwards for H time steps using the belief update of Eq. 1.

2.2 FAE and Confirmation Bias

In cyber-attack scenarios, attackers spend significant time gathering information about their target and making inferences based on their observations. Hence, in this paper, we focus specifically on select biases that affect the inference process of humans. *Fundamental attribution error* is known to affect the attribution process in human reasoning. FAE is the tendency of an observer to overestimate interpersonal factors compared to environmental factors. Studies show that humans underestimate the effects of the environment and the situation while characterizing the behavior of their peers [22]. Recent work on a game-theoretic formulation of FAE illustrates the role it plays in negotiations, specifically, buyer-seller interactions [9]. The formulation utilizes an agent's *coarse knowledge* of the state to simulate the skewed inference. While a rational agent models the opponent's behavior conditioned on the state, an agent affected by FAE models the opponent's behavior independent of the state. Such coarse modeling produces a biased reasoning process that may erroneously attribute observations to

the opponent instead of the state. A rational agent can exploit this biased inference caused by coarse reasoning to deceive the opponent about her *type*. Importantly, this game-theoretic formulation shows that *coarse thinking* [20] can lead to FAE.

Another bias afflicting the acquisition of information in humans is the *confirmation bias*. While confirmation bias manifests in human reasoning in many forms [21], we specifically focus on the phenomenon of *overweighting positive confirmatory evidence*. This effect leads to the agent overweighting observations that conform to her predicted belief state. Further, the agent underweights observations that are contrary to her belief state. Recent work on extended goal recognition shows that confirmation bias acts like inertia – keeping the agent slightly biased towards the hypothesis formed from the initial few observations [19].

We model and consider the effects of attacker FAE and confirmation bias in decision-theoretic settings. In the next section, we define and explain our formulation of these biases using a simple illustration.

3 COGNITIVE MODELING IN I-POMDP$_\mathcal{X}$

We model the cognitive biases introduced in Section 2.2 within the I-POMDP$_\mathcal{X}$ framework introduced in Section 2.1. We start by introducing FAE from a decision-theoretic perspective. Analogously to Ettinger and Jehiel [9], we demonstrate FAE as a consequence of coarse thinking, although our modeling pertains to the state and differs from theirs. We also introduce a model for weighting observations in a manner consistent with confirmation bias.

Consider a simple cyber defense scenario posed as a two-agent decision-making problem. The physical state consists of a host system which can either be a *critical system*, `HostType = c`, or a *honeypot* system, `HostType = hp`. Thus, `HostType = {c, hp}` forms the physical state space. There are two adversarial agents in the interaction: the *attacker* and the *defender*. Attacker agent a at strategy level $l = 1$ models a defender agent d at strategy level $l = 0$. For the sake of simplicity, we assume that both agents get perfect observations about the state and the attacker gets perfect observations about the defender's actions. The defender agent may either be *passive* with less capability, or *active* that can defend against the attacker. The passive defender passively observes and logs the attacker's actions. On the contrary, the active defender is capable of swift countermeasures to defend a critical system using the `Defend` action.

Definition 1 (Level 0 defender agent)**.** The level 1 attacker models the level 0 defender as a factored POMDP agent, which is defined as the tuple $\langle \mathcal{X}, A, T_d, \mathcal{Y}_d, O_d, \mathcal{R}_d, OC_d \rangle$. Here, $\mathcal{X}$ is the set of physical state variables. The other elements of the tuple are analogous to those defined for the I-POMDP$_\mathcal{X}$ in Section 2.1, but for a single agent.

In this illustration, $\mathcal{X} = \{\texttt{HostType}\}$. We assume that the physical state `HostType` does not change. Hence, T is defined as $P(\mathbf{x}'|\mathbf{x}, a_d) = 1$ when $\mathbf{x} = \mathbf{x}'$ and 0 otherwise $\forall \mathbf{x} \in \mathcal{X}, \mathbf{x}' \in \mathcal{X}, a_d \in A_d$. The defender's set of actions is defined as $A_d = \{\texttt{Defend}, \texttt{No-Op}\}$. Additionally, the attacker models the defender agent as having different levels of capability through frames. The set of defender frames is then defined as $\hat{\Theta}_d = \{\hat{\theta}_{act}, \hat{\theta}_{pass}\}$. For the passive defender, $\hat{\theta}_{pass}$, the optimal policy is to simply perform `No-Op` no matter the type of the host. On the other hand, the active defender's policy will be,

$$\pi_{\hat{\theta}_{act}}(\texttt{HostType}) = \begin{cases} \texttt{Defend}, & \texttt{HostType = c} \\ \texttt{No-Op}, & \texttt{HostType = hp} \end{cases}$$

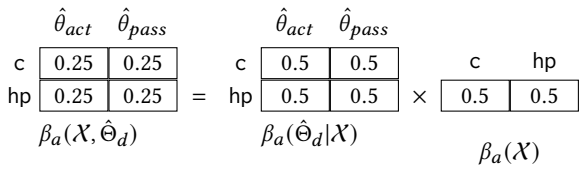


Figure 1: A rational attacker's beliefs over the defender's models are conditioned on the host type due to her ability to distinguish the host types. We also show the attacker's belief over the joint space obtained as $\beta_a(\mathcal{IS}_a) = \beta_a(\mathcal{X}) \times \beta_a(\hat{\Theta}_d|\mathcal{X})$.

Definition 2 (Level 1 attacker I-POMDP$_\mathcal{X}$)**.** The level 1 attacker is an I-POMDP$_\mathcal{X}$ agent defined as the tuple $\langle \mathcal{IS}_a, A, T, \mathcal{Y}_a, O_a, \mathcal{R}_a, OC_a \rangle$. $\mathcal{IS}_a$ is the attacker's interactive state space, $\mathcal{X} \times \hat{\Theta}_d$. The transition function $P(\mathbf{x}'|\mathbf{x}, a_d) = 1$ when $\mathbf{x} = \mathbf{x}'$ and 0 otherwise $\forall \mathbf{x} \in \mathcal{X}, \mathbf{x}' \in \mathcal{X}, a_a \in A_a, a_d \in A_d$. We assume the attacker gets perfect observations about the defender's action and the state, but not of the defender's frame which must be inferred.

The attacker's set of actions is defined as $A_a = \{\texttt{Attack}, \texttt{No-Op}\}$. Here, `Attack` is an abstract action that yields some reward when performed on a critical system. Also, attacking a system when a defender defends it incurs some cost for the attacker. Thus, ideally, the attacker aims to attack a critical system, `HostType = c`, with a passive defender $\hat{\Theta}_d = \hat{\theta}_{pass}$. We assume perfect observability of the state for the defender.

Example 1 (Rational attacker's beliefs at timestep t)**.** Consider the attacker agent's probability space $\langle \mathcal{IS}_a, \Sigma_a, \beta_a \rangle$, where $\mathcal{IS}_a = \mathcal{X} \times \hat{\Theta}_d$, and Σ_a is the sigma algebra of measurable subsets of $\mathcal{IS}_a$. For finite sets $\mathcal{X}$ and $\hat{\Theta}_d$, Σ_a is generally the powerset of $\mathcal{X} \times \hat{\Theta}_d$: $\Sigma_a = 2^{\mathcal{IS}_a}$. β_a is the attacker's probability measure function which assigns a *belief* to the elements in Σ_a such that $\beta_a(\mathcal{IS}_a) = 1$. Fig. 1 shows a rational attacker's beliefs over $\mathcal{IS}_a$.

We assume the attacker gathers information in the first step. Consequently, the attacker's observation function reveals the state of the system, $\mathcal{X} = \mathbf{x}'$, and the defender's action, $A_d = a_d$. Using these observations, the attacker agent computes her posterior belief distribution over $\mathcal{IS}_a$ using the Bayes rule, $\beta_a(\hat{\Theta}'_d, \mathcal{X}'|A_d = a_d) = \alpha P(A_d = a_d|\mathcal{X}', \hat{\Theta}'_d) \sum_{\mathcal{X}, \hat{\Theta}_d} P(\mathcal{X}', \hat{\Theta}'_d|\mathcal{X}, \hat{\Theta}_d) \beta_a(\mathcal{X}, \hat{\Theta}_d)$ where α is the normalizing constant and $P(A_d|\mathcal{X}, \hat{\Theta}_d)$ is essentially given by the defender's optimal policy for each frame. Note that this update is analogous to the I-POMDP$_\mathcal{X}$ belief update described in Eq. 1 with the assumption of perfect observability and no transitions.

Example 2 (Rational attacker's beliefs at timestep $t+1$)**.** Let $\mathcal{Y}'_i = \{\texttt{hp}, \texttt{No-Op}\}$ be the observed physical state and defender's action, respectively. The attacker's posterior belief, $\beta_a(\mathcal{X}', \Theta'_d)$ is shown in Fig. 2.

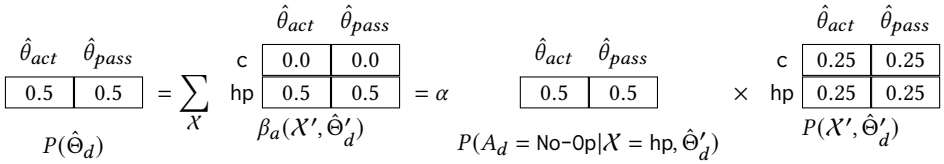


Figure 2: A rational attacker's posterior belief over the interactive state space $\mathcal{IS}_a$ is shown above as $\beta_a(\mathcal{X}', \hat{\Theta}'_d)$. $P(\hat{\Theta}_d)$ the belief over the opponent's frame and can be computed as $\sum_{\mathcal{X}} \beta_a(\mathcal{X}, \hat{\Theta}_d)$

We now illustrate how the attacker's posterior belief changes when she is unable to distinguish between the host system's types, `HostType = c` or `HostType = hp`, which prevents her from modeling the relationship between the defender's behavior and the physical state.

3.1 Coarse Thinking and FAE

In our treatment of coarse thinking, we consider scenarios where an agent cannot disambiguate between multiple states and aggregates them into a single partition. Such partitioning is analogous to the concept of *information partitions* first introduced by Halpern [10] and Aumann [1] for representing knowledge and reasoning with it. The partitioning of states has also been used before to define finite-level type spaces in Bayesian Markov games [4]. Note that our state partitioning, which is based on an agent's lack of knowledge, is fundamentally distinct from state-aliasing [27], which is an artifact of learned representations.

Recall that rational agents have the ability to perfectly model the physical state space because they can assign belief measures to individual elements of $\mathcal{IS}$. Now, consider a scenario in which the attacker is unable to disambiguate between a *critical system* `HostType = c` and a *honeypot* `HostType = hp` due to being unaware of the existence of honeypots in the network. Such an attacker reasons over a coarser state space using state *partitions*. We model these state partitions by ensuring that the set of physical events $\Sigma_X = \sigma(\{\texttt{c}, \texttt{hp}\})$ instead of the previous $\Sigma_X = 2^X$; $\sigma(\cdot)$ defines the smallest sigma algebra[1] containing the grouping $\{\texttt{c}, \texttt{hp}\}$. Consequently, the belief function β can no longer assign a probability measure to individual states `c` and `hp` because they are not measurable in the defined probability space. Instead, β now assigns a probability to the element $\{\texttt{c}, \texttt{hp}\}$ of the partition.

In the following examples, we utilize such partitioning to model an agent's coarse thinking.

Example 3 (Biased attacker's beliefs at time step t)**.** A human attacker, who is likely to exhibit common biases, may model the defender's behavior independently of the type of the system. We simulate this by assigning an element of the partition `HostType = {c,hp}` as the attacker's only state. Consequently, the attacker can no longer distinguish between individual states `HostType = c` and `HostType = hp`. The resulting belief over a coarse $\mathcal{IS}_a$ is shown in Fig. 3.

[1] $\Sigma \subseteq P(X)$ for a set X and its powerset $P(X)$ is a σ-algebra if the following conditions are true; $X \in \Sigma$, Σ is closed under complementation, and Σ is closed under countable unions

Analogously to the rational attacker, we simulate the biased attacker's belief update on receiving the observation $\{\texttt{hp}, \texttt{No-Op}\}$. The belief update gives us a posterior distribution over $\mathcal{IS}_a$ with a coarse representation of $\mathcal{X}$. Due to the partitioning of the state space, the observation function that informs the attacker about `HostType` is no longer valid because the attacker cannot assign probabilities inside the partition. We simulate the belief update on the partitioned state space to show that FAE results from the inference process on a coarse representation of the state. The following example shows the attacker's posterior belief over the partitioned $\mathcal{IS}_a$.

Example 4 (Biased attacker's beliefs at time step $t+1$)**.** The biased attacker's posterior belief is obtained by simulating one step of the I-POMDP$_{\mathcal{X}}$ belief update mentioned in Eq. 1 with the prior belief as given in Example 3. Fig. 3 shows the posterior belief after observing that the defender agent performs action `No-Op`.

The attacker agent, on observing that the defender performs the `No-Op` action, updates her posterior distribution over the defender's models by weighting them according to their likelihood of performing the `No-Op` action independently of `HostType`. Recall that the active defender's optimal policy, $\pi_{\hat{\theta}_{act}}$, suggests action `Defend` for `HostType = c`. Because of her coarse representation, the attacker ignores this conditioning of $\pi_{\hat{\theta}_{act}}$ on `HostType` causing her to estimate $P(A_d = \texttt{No-Op} | \hat{\theta}_{pass}) > P(A_d = \texttt{No-Op} | \hat{\theta}_{act})$ as $\hat{\theta}_{act}$ performs `Defend` with non-zero probability while $\hat{\theta}_{pass}$ performs `No-Op` only. Such an (erroneous) update results in the attacker believing that the defender is more likely to be passive (i.e., of the frame $\hat{\theta}_{pass}$). If instead, the attacker was rational and able to model the dependence between the defender's policy and `HostType`, she would infer after observing `HostType = hp` that $P(A_d = \texttt{No-Op} | \texttt{HostType = hp}, \hat{\theta}_{pass}) = P(A_d = \texttt{No-Op} | \texttt{HostType = hp}, \hat{\theta}_{act})$ as shown in Fig. 2. A rational attacker then stays neutral in her belief about whether the defender is active or not. The grouping of the state space makes such conditioning impossible, inducing the attacker to draw inferences about the defender types while being ambiguous about the physical state.

This faulty inference is consistent with the phenomenon of FAE. Thus, I-POMDP$_{\mathcal{X}}$ can explicitly model FAE resulting from an agent's coarse thinking. This decision-theoretic formulation of FAE is novel and enables its explicit modeling and manifestation.

3.2 Confirmation Bias

After establishing initial access, attackers are uninformed about the system's state. Hence, their initial observations dominate their

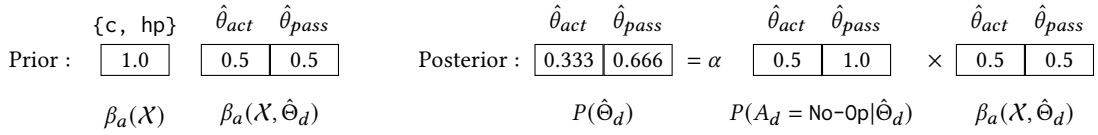


Figure 3: Coarse state representation causes the attacker to model the defender's behavior independently of the state. $\beta_a(\mathcal{X})$ shows the attacker's belief over the coarse state space. Due to the coarse representation, $\beta_a(\mathcal{X}, M_d)$ becomes a uniform distribution indicating that the attacker believes all defender behaviors to be equally likely.

hypotheses about the environment. Confirmation bias plays a critical role in such situations. Specifically, once the attacker forms a belief about the defender's frame, subsequent observations that contradict her belief are weighted less by the attacker.

Consider a subsequent interaction between the attacker and the defender agents after the attacker has observed the defender's initial actions and formed a belief about her frame. In our illustration, we assume that the active defender, $\hat{\theta}_{act}$, can now deploy decoy files to deceive the attacker when `HostType = hp`. Thus, the optimal policy for the active defender is,

$$\pi_{\hat{\theta}_{act}}(\texttt{HostType}) = \begin{cases} \texttt{Defend}, & \texttt{HostType = c} \\ \texttt{Deceive}, & \texttt{HostType = hp} \end{cases}$$

The optimal policy for the passive defender stays the same as previously defined.

Example 5 (Attacker's prior belief during a subsequent interaction)**.** Assume a scenario with the attacker already biased in her belief about the defender's frame due to FAE. Left Fig. 4 shows the attacker's prior belief during the second interaction.

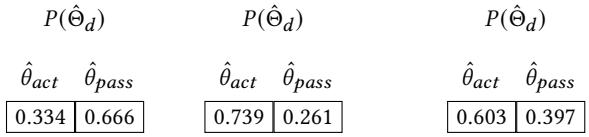


Figure 4: Confirmation bias causes the attacker to underweight evidence that does not conform to the predicted belief. The left figure shows the attacker's prior belief. The middle figure shows the rational updated belief. The right figure shows the updated belief due to confirmation bias.

For the sake of illustration, let the attacker observe the defender performing the `Deceive` action. Eqn. 1 computes the attacker's posterior belief over the defender's frame.

Example 6 (Attacker's posterior belief after a normative belief update)**.** On observing the `Deceive` action by the defender, the attacker updates her belief over the defender's frame. Fig 4 shows the attacker's posterior belief following the observation. The attacker models the active defender's ability to employ deception. Consequently, after observing deception, the attacker correctly updates her belief that the defender is active.

We now show how confirmation bias causes the attacker to discount the defender's observed behavior. Let $b(\mathcal{X}, \Theta_d)$ be the attacker's prior belief, $A_a = a_a$ be the attacker's action, and $\mathcal{Y}'_a = \mathbf{y}$ the observation at time step t. The posterior belief $b(\mathcal{X}', \Theta'_d)$ at time step $t+1$ due to confirmation bias is given by,

$$P(\mathcal{X}', \Theta'_d | a_a, \mathbf{y}) = \frac{\sum_{A_j} \left(\prod_{i=1}^{n} P(Y_i = y_i | \mathcal{X}', Aj)^{\gamma_i}\right) \beta_{\text{pred}}}{\sum_{\mathcal{X}', A_j} \left(\prod_{i=1}^{n} P(Y_i = y_i | \mathcal{X}', Aj)^{\gamma_i}\right) \beta_{\text{pred}}} \quad (4)$$

where, $\beta_{\text{pred}} = \sum_{\mathcal{Y}'_j} P^{a_i}(\mathcal{Y}'_j, M'_j, \mathcal{X}', A_j | M_j, \mathcal{X})$.
The ADD $P^{a_i}(\mathcal{Y}'_j, M'_j, \mathcal{X}', A_j | M_j, \mathcal{X})$ is exactly the same as the ADD $P^{a_i, o_i}(\mathcal{Y}'_j, M'_j, \mathcal{X}', A_j | M_j, \mathcal{X})$ defined in equation 2 without restriction of $\mathcal{Y}_i = o_i$. γ_i is the *weighting factor* for observation variable Y_i given by, $\gamma_i = (1 + ||P(Y_i = y_i | \mathcal{X}', Aj) - P_{a_a}||_1^1)^{-1}$.

Example 7 (Attacker's posterior belief due to confirmation bias)**.** The *weighting factor* now influences the attacker's belief update. The defender's observed action, `Decieve`, contradicts the attacker's prior belief that the defender is passive, yielding a small weighting factor. Consequently, the attacker underweights this observation. Fig. 4 shows the attacker's posterior belief due to confirmation bias.

This scenario shows the inertial property of confirmation bias which causes agents to underweight evidence that does not conform to their beliefs.

4 EXPERIMENTS

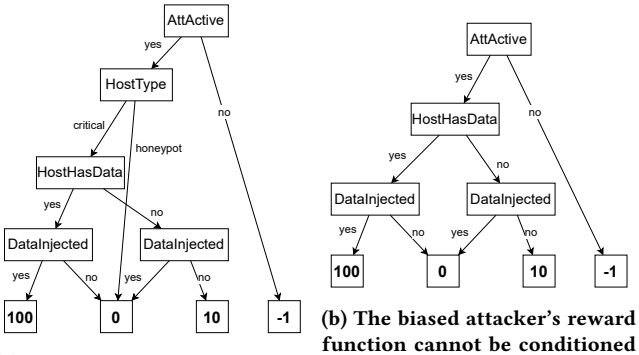
In the previous section, we illustrated the effect of cognitive biases on the attacker's beliefs about the defender's capabilities. We simulate interactions between our model of a biased human attacker and an I-POMDP$_{\mathcal{X}}$-based defender agent. Below, we briefly summarize the domain and discuss the results of our experiments. The code for the I-POMDP$_{\mathcal{X}}$ solver, the domain, and the supplementary material are available at https://github.com/dityas/CaffeineBravery

4.1 The Active Cyberdefense Domain

Both agents reason over a common physical state space representing the host system and recursively model each other's beliefs to compute an optimal adversarial policy. Shinde and Doshi [25] recently introduced the *cyber deception domain* enabling I-POMDP$_{\mathcal{X}}$-based active defenders to recognize an attacker's intent using deception. We adopt this domain with a few modifications to the state space. We include a detailed description of the domain with the supplementary material.

In the cyber deception domain, a set of discrete random variables describes the state of the host system on which the interaction occurs. We add the `HostType` state variable to this domain, along with

its partitioning introduced previously in Section 3.1 to represent the type of host system. The HostType does not change because the interaction occurs on a honeypot. We include state variables that model the presence of valuable data and privilege escalation opportunities for the attacker– the HostHasData and EscAccounts state variables. Other state variables that describe the attacker's progress– DataFound, AccFound, DataInjected, and AttActive, are exactly as defined in the cyber deception domain. Due to the partial observability of the host system, the attacker agent gets noisy observations from her actions. We model the attacker's noisy observations using an observation function that probabilistically maps the state of the system to the attacker's observations. Similarly, the defender agent must rely on her log inference capabilities to infer the attacker's actions. We assign a noisy observation function to the defender to model the ambiguity in realistic log inference systems. A detailed description of the state and observation variables and action sets for both agents is included in the supplementary material.



(a) The rational attacker's reward function is conditioned on the HostType state.

(b) The biased attacker's reward function cannot be conditioned on HostType state variable due to her coarse representation.

Figure 5: The attacker's reward function $R^{\texttt{EXIT}}(\mathcal{X})$ assigns a high reward for successful data manipulation.

To achieve her goal, the attacker employs information-gathering actions and impact-causing actions. Information-gathering actions inform the attacker about the state of the host system. Impact-causing actions enable the attacker to manipulate the state. We include both types of actions in our domain. To ensure realistic modeling of attacker actions, we utilize the MITRE ATT&CK matrix [26] as a reference. The attacker's reward function assigns a high reward for the manipulation of valuable data when HostType = critical. However, in the case of a biased attacker, the reward is not conditioned on HostType due to her coarse representation of this state variable. We show the ADD representation of the reward function for both cases– a rational attacker, and a biased attacker in Figure 5.

We model the active and passive defender agents analogously to the ones defined in Section 3. The passive defender agent simulates existing deception strategies that simply try to log attacker activity on honeypots without actively engaging them. In contrast, the active defender agent aims to evict an attacker present on a critical host system and engage the attacker on a honeypot system. Consequently, we condition the active defender's reward function on the HostType state variable. We include deceptive and preventive capabilities in the active defender's action set. The defender can evict the attacker from a critical system. However, there is a high cost associated with performing this action in the attacker's absence because this would imply that a legitimate user was evicted. Consequently, the defender agent has to utilize her log inference capabilities to infer the attacker's presence. On a honeypot system, the defender can deploy data decoys to bait and engage the attacker. We use the MITRE D3FEND matrix [18] as a reference to model these actions and their effects on the state of the host system.

We model the interaction between an active defender agent at strategy level $l = 3$ and an attacker agent at level $l = 2$. In the following section, we present and discuss the results of simulated interactions between the two agents.

4.2 Simulated Interactions

Human attackers act and infer information differently from rational agents. To investigate this departure from rationality, we compare our model of a biased human attacker with a rational attacker. The rational attacker agent has perfect state representation and a normative belief update. We perform 100 simulations where each simulation continues until the attacker exits with each attacker type – a rational attacker, an attacker with FAE, and an attacker with FAE and confirmation bias. We address these attackers as *rational*, *fae*, and *fae_cbias* for the sake of this discussion.

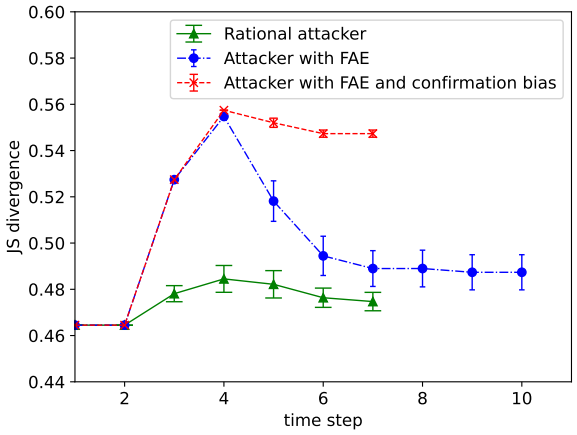
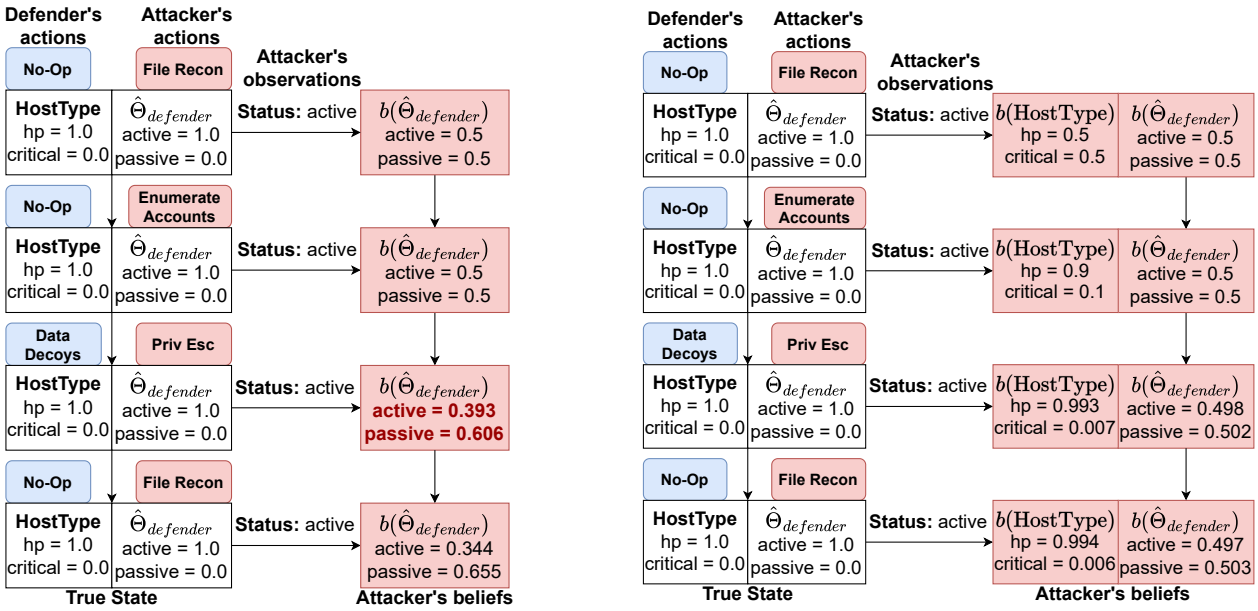


Figure 6: JS divergence between the attackers' beliefs and the defender's type indicates that a combination of FAE and confirmation bias causes a significant error in attribution.

Initially, the simulated attackers are uninformed about the defender's frame, as indicated by their identical Jensen-Shannon (JS) divergence values in the first time step in Figure 6. As the interaction proceeds, around time step 3, both *fae* and *fae_cbias* predict that an active defender will attempt to use defensive measures to prevent their progress. However, the defender agent, knowing that the HostType is a honeypot, instead deploys decoys to keep



(a) The attacker's beliefs shown in the right column deviate from the true states shown on the left due to fundamental attribution error

(b) A rational attacker can reason about the state because of her perfect state representation. Consequently, the attacker avoids FAE by correctly conditioning the defender's behavior on the state

Figure 7: A comparison of beliefs between *fae* and *rational* shows *rational*'s ability to utilize context and remain neutral about the defender.

the attacker engaged using the `DataDecoys` action. Both *fae* and *fae_cbias*, having a coarse state representation update their belief about the defender's frame while being unaware of the `HostType` state, causing them to attribute the lack of defensive actions entirely to the defender's ability. Figure 6 shows this attribution error. The blue and red lines trace the JS divergence of these attackers. The blue line denotes *fae*, and the red line denotes *fae_cbias*. At step 3, both attackers have significantly higher JS divergence values than *rational*. In contrast to this behavior, *rational* starts forming a belief about the `HostType` state based on her observations while gathering information using the `FileRecon` and `EnumAccounts` actions. Consequently, in step 3, when *rational* does not observe a reaction from the defender, she updates her belief about the defender's frame by conditioning it on the `HostType` state. Thus, *rational* avoids FAE by being aware of the context in attributing observations.

Figure 7 highlights this difference by simulating both belief updates on the same action-observation sequence as described previously. Both, *fae* and *rational* are initially uninformed about the defender's frame as shown in step 1. In Figure 7a, steps 3 and 4 show *fae*'s beliefs due to FAE. Figure 7b presents a comparative rational belief update with minimal error in attribution. *Rational*'s utilization of context to reason about her observations is evident from her successively higher belief that `HostType = honeypot`, and minimal error in attribution in steps 3 and 4.

Following the faulty inference due to FAE, the attacker continues her attempts at locating and manipulating critical data. While both *fae* and *fae_cbias* cannot differentiate between honeypots and critical systems, they do expect a defender, if present, to engage them. Consequently, *fae* attributes her discovery of valuable data equally to the `HostHasData` state, and the active defender's attempt to keep her engaged. However, for *fae_cbias* who also exhibits confirmation bias, the attribution of data discovery to an active defender contradicts her prior belief of the defender being passive. Consequently, confirmation bias causes *fae_cbias* to underweight this contradicting evidence and instead maintain her belief that the defender is passive. Fig 6 captures this difference in inference due to confirmation bias after step 4. *Fae*, using a normative belief update, shown in blue, starts correcting her belief about the defender type indicative from the steep drop in divergence. In contrast, *fae_cbias*, shown in red, is slow to apply this correction. Such delayed correction of beliefs by ignoring contradicting evidence is behavior consistent with confirmation bias. Figure 8 shows this delayed correction process caused by confirmation bias by comparing it with a rational belief update.

In our simulations, the rational attacker promptly exited the system after recognizing `HostType = honeypot` in the early stages of the interaction. Consequently, the defender agent, on also predicting a likely early exit, did not compute a policy to engage the attacker. The rational attacker stayed in the host system for a mean duration of **2.37** interaction steps. In contrast, the biased human attackers were more persistent due to their unawareness of the system being a honeypot and thinking that the defender was either

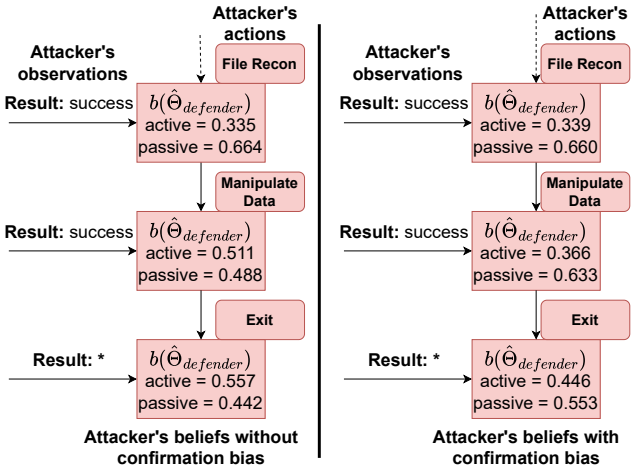


Figure 8: Confirmation bias induces the attacker to underweight contradicting evidence and delays belief correction

absent or not active. Consequently, the defender agent utilized effective deception strategies to engage them on the honeypot. The *fae_cbias* attacker continued the attack for a mean duration of **6.26** interaction steps. This difference in interaction durations was statistically significant ($p = 1.001 \times 10^{-32}, n_1 = n_2 = 100$ for a Wilcoxon rank-sum test). Engaging the attacker for a longer duration enabled the defender to observe a wider range of attacker capabilities. In our simulations, excluding the `No-Op` and `Exit` actions, the rational attacker only performed `FileRecon`, an information-gathering action, before recognizing `HostType = honeypot` and exiting the system. The rational attacker performed an average of **1.19** unique actions in every interaction. In contrast, the biased human attacker (*fae_cbias*) performed a mean of **4** unique actions excluding `No-Op` and `Exit`. The difference in the number of unique actions performed by the rational attacker and the biased attacker was statistically significant ($p = 2.523 \times 10^{-34}, n_1 = n_2 = 100$ for a Wilcoxon rank-sum test). The biased attacker started the interaction with the `EnumAccounts` action to discover escalation opportunities. The attacker then escalated her privileges using the `PrivEsc` action. Subsequently, the attacker located valuable data using `FileRecon`. Finally, the attacker performed the `ManipulateData` action to achieve her objective and conclude the attack. The biased attacker's behavior spanned her entire action set. The defender agent was thus able to deceive the attacker into behaving exactly as if she were attacking a real system. Computing this optimal defender strategy took around 16 mins 20 secs on a quad-core Intel i3 with 16GB RAM for 3 levels of nesting. As such, the factored representation of the framework scales well along multiple dimensions.

Our simulations demonstrate that the cognitive modeling of human attackers facilitates the development of viable deception strategies to engage attackers on sandboxed systems.

5 RELATED WORK

Recent work on AI-based techniques for cyberdefense include various game-theoretic approaches to modeling the interaction between attackers and defenders [3, 11, 14, 23]. While these works are similar to ours in their strategic use of honeypots, we model deception at the deeper level of beliefs of the involved agents. Particularly interesting is the work by Ettinger and Jehiel on modeling psychological biases relevant to deception from a game-theoretic perspective [9]. They proposed a framework to model deception in equilibrium strategies against players with coarse reasoning. In contrast, we adopt a decision-theoretic approach for explicitly modeling the agents instead of computing equilibrium policies. Additionally, we use the concept of state partitions to represent context unawareness. Our work shows that such representation plays a critical role in modeling FAE. Masters et al. recently proposed a model of confirmation bias to deceive observers engaging in goal recognition [19]. They use proximity from optimal paths to compute a weighting factor. In contrast, our work does not rely on path planning and instead weights observations according to their proximity to the predicted belief.

Doshi et al. [8] proposed the empirically informed I-POMDP to model human behavioral data in general-sum and fixed-sum games. They augmented the I-POMDP framework to model human decision-making based on known cognitive literature. In contrast, we focus on biases that play a significant role during cyberattacks. Another study on decision-making in cyberattacks was the Tularosa experiment [12] which recorded participants during a red-team exercise. Their studies showed that attackers exhibited biases such as confirmation bias, sunk cost fallacy, ambiguity effect, and self-serving bias [13]. Our work provides an analogous model-based framework to study some of these biases from the subjective perspective of the participating agents.

6 CONCLUSION

The augmented I-POMDP$_\chi$ framework presents a decision-theoretic formulation of coarse thinking, fundamental attribution error, and confirmation bias, and how it may benefit cybersecurity. This approach of modeling cognitive biases and utilizing them to engage attackers is a pioneering approach to cybersecurity. Conventional cybersecurity has been an arms race with attackers and defenders developing tools and techniques to overcome their opponent's capabilities. However, human behavior has been a common denominator across most targeted cyberattacks. We leverage a cognitive model of human decision-making to predict and manipulate attacker behavior. Our work is a step toward leveraging such cognitive models of human behavior to engage cyber attackers that have an asymmetric advantage over defenders.

More broadly, our framework has applications in studying the effects of fundamental attribution error and confirmation bias in generic interactions. Our model-based approach has several potential applications in psychology, economics, and cognitive science.

ACKNOWLEDGMENTS

This research was supported, in part, by a grant from the Army Research Office under grant number W911NF-18-1-0288. We thank former graduate student Muhammed Abuodeh for initiating this research direction and assistance from Dr. Adam Goodie in UGA's Department of Psychology in understanding the confirmation bias.

7 ETHICAL IMPLICATIONS

Our work follows a long history of research on modeling the effects of cognitive biases on human decision-making and behavior. We recognize that the broader use of these computational models may raise potential ethical concerns and address these below.

Leveraging an attacker's cognitive biases for deception is a purely defensive capability as it enables the defender to engage with the attacker on sandboxed systems. Also, our model of coarse thinking and FAE relies on the attacker's lack of knowledge about the state. Consequently, our framework may not be effectively misused in an offensive capacity by the attacker because the defender has complete knowledge about the state of the honeypot system. We recognize that a malicious actor may benefit from utilizing our models to explore generic deception strategies. However, the strategic depth of our framework will allow the identification, analysis, and remediation of biased behavior in such scenarios, thereby negating any unfair advantage to the malicious actor. On a broader scale, we believe that the possibility for positive impact from our work far outweighs any unexpected potential for its misuse.

REFERENCES

[1] Robert J Aumann. 1999. Interactive epistemology I: knowledge. *International Journal of Game Theory* 28 (1999), 263–300.

[2] R Iris Bahar, Erica A Frohm, Charles M Gaona, Gary D Hachtel, Enrico Macii, Abelardo Pardo, and Fabio Somenzi. 1997. Algebric decision diagrams and their applications. *Formal methods in system design* 10, 2-3 (1997), 171–206.

[3] Thomas E. Carroll and Daniel Grosu. 2011. A game theoretic investigation of deception in network security. *Security and Communication Networks* 4, 10 (2011), 1162–1172. https://doi.org/10.1002/sec.242

[4] Muthukumaran Chandrasekaran, Yingke Chen, and Prashant Doshi. 2016. Bayesian Markov games with explicit finite-level types. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

[5] Daniel Dennett. 1986. Intentional systems. Brainstorms.

[6] Prashant Doshi. 2012. Decision Making in Complex Multiagent Settings: A Tale of Two Frameworks. *AI Magazine* 33, 4 (2012), 82–95.

[7] Prashant Doshi and Dennis Perez. 2008. Generalized Point Based Value Iteration for Interactive POMDPs.. In *AAAI*. AAAI Press, 63–68.

[8] Prashant Doshi, Xia Qu, Adam Goodie, and Diana Young. 2010. Modeling recursive reasoning by humans using empirically informed interactive POMDPs. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. 1223–1230.

[9] David Ettinger and Philippe Jehiel. 2010. A theory of deception. *American Economic Journal: Microeconomics* 2, 1 (2010), 1–20.

[10] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. 2004. *Reasoning about knowledge*. MIT press.

[11] Kimberly Ferguson-Walter, Sunny Fugate, Justin Mauger, and Maxine Major. 2019. Game theory for adaptive defensive cyber deception. In *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security*. ACM, Association for Computing Machinery, New York, NY, USA, 4.

[12] Kimberly Ferguson-Walter, Temmie Shade, Andrew Rogers, Michael Christopher Stefan Trumbo, Kevin S Nauer, Kristin Marie Divis, Aaron Jones, Angela Combs, and Robert G Abbott. 2018. *The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception*. Technical Report. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

[13] Kimberly J Ferguson-Walter, Maxine M Major, Chelsea K Johnson, and Daniel H Muhleman. 2021. Examining the efficacy of decoy-based and psychological cyber deception. In *30th USENIX Security Symposium (USENIX Security 21)*. 1127–1144.

[14] Sunny Fugate and Kimberly Ferguson-Walter. 2019. Artificial Intelligence and Game Theory Models for Defending Critical Networks with Cyber Deception. *AI Magazine* 40, 1 (2019), 49–62.

[15] Piotr J Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24 (2005), 49–79.

[16] D. Kahneman, P. Slovic, and A. Tverksy (Eds.). 1982. *Judgment under Uncertainty: Heuristic and Biases*. Cambridge University Press.

[17] Anirudh Kakarlapudi, Gayathri Anil, Adam Eck, Prashant Doshi, and Leen-Kiat Soh. 2022. Decision-theoretic planning with communication in open multiagent systems. In *Uncertainty in Artificial Intelligence*. PMLR, 938–948.

[18] Peter E Kaloroumakis and Michael J Smith. 2021. Toward a knowledge graph of cybersecurity countermeasures. *The MITRE Corporation* 11 (2021).

[19] Peta Masters, Michael Kirley, and Wally Smith. 2021. Extended goal recognition: a planning-based model for strategic deception. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 871–879.

[20] Sendhil Mullainathan, Joshua Schwartzstein, and Andrei Shleifer. 2008. Coarse thinking and persuasion. *The Quarterly journal of economics* 123, 2 (2008), 577–619.

[21] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[22] Lee Ross. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*. Vol. 10. Elsevier, 173–220.

[23] Aaron Schlenker, Omkar Thakoor, Haifeng Xu, Long Tran-Thanh, Fei Fang, Phebe Vayanos, Milind Tambe, and Yevgeniy Vorobeychik. 2018. Deceiving cyber adversaries: A game theoretic approach. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 2 (2018), 892–900.

[24] Jonathon Schwartz, Ruijia Zhou, and Hanna Kurniawati. 2022. Online Planning for Interactive-POMDPs Using Nested Monte Carlo Tree Search. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8770–8777.

[25] Aditya Shinde, Prashant Doshi, and Omid Setayeshfar. 2021. Cyber attack intent recognition and active deception using factored interactive pomdps. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 1200–1208.

[26] Blake E Strom, Andy Applebaum, Doug P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. 2018. *Mitre Att&ck: Design and Philosophy*. Technical Report. MITRE Corp.

[27] Steven D Whitehead and Long-Ji Lin. 1995. Reinforcement learning of non-Markov decision processes. *Artificial intelligence* 73, 1-2 (1995), 271–306.