Deception in Optimal Control

Melkior Ornik and Ufuk Topcu

***Abstract*— In this paper, we consider an adversarial scenario where one agent seeks to achieve an objective, and its adversary seeks to learn the agent's intentions and prevent the agent from achieving its objective. Thus, the agent has an incentive to try to deceive the adversary about its intentions, while at the same time working to achieve its objective. The primary contribution of this paper is to introduce a mathematically rigorous framework for the notion of deception within the context of optimal control. The central notion introduced in the paper is that of a belief-induced reward: a reward dependent not only on the agent's state and action, but also adversary's beliefs. While the paper focuses on the setting of Markov decision processes, the proposed framework allows for deception to be defined in an arbitrary control system endowed with a reward function. Furthermore, we discuss design of optimal deceptive strategies, under possible additional specifications on agent's behavior or uncertainty in agent's knowledge about the adversary's learning process, and show that such design reduces to known problems in control design on partially observable or uncertain Markov decision processes. The notion of deception is illustrated on a running example of "cops and deceptive robbers". We show that an optimal deceptive strategy of this example, designed using the theory introduced in the paper, follows the intuitive idea of how to deceive an adversary in such a scenario.**

I. Introduction

The concept of deception is naturally present in a variety of contexts that have an adversarial element. Examples include cybersecurity [1], [2], bio-inspired robotics [3], genetic algorithms [4], vehicle decision making [5], warfare strategy — a particularly voluminous study of the role of deception in war has been made in [6] — and interpersonal relationships [7]. A *deceptive strategy* employed by an agent in an adversarial setting rests on agent's two complementary goals:

1) reaching its objective,
2) modifying the adversary's beliefs about the nature of the objective — for instance, objective location, distance, or reward attained at an objective.

The desire to modify the adversary's beliefs is motivated by the assumption that the adversary would be able to block, or modify, access to the objective if it correctly identified its nature. Thus, the success of an agent at reaching its control objective often depends on the agent's ability to hide its true intentions from its adversary while still proceeding towards its objective, and satisfying any other constraints that may be placed on its behavior (e.g., safety specifications).

A simple example for a deceptive setting, which will serve as the running example throughout this paper, is given in Fig. 1 — we refer to it as *cops and deceptive robbers*. In the setting illustrated by Fig. 1, the agent ("robbers") seeks to move to a particular area of a state space ("bank"), which holds a reward. An adversary ("cops") knows that the agent is seeking to reach one of several possible objectives, but does not know which one. By observing the agent's movement, the adversary seeks to learn the agent's intentions, and change the nature of the objective ("set a trap"). Hence, it is in the agent's interest to make the adversary's beliefs about the agent's intentions as incorrect as possible, while still ultimately reaching its objective.

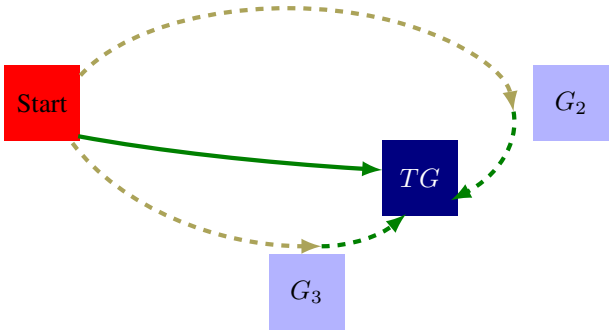

Fig. 1. An illustration of the running example of this paper. The agent seeks to take a path from the starting area, denoted in red, to its true goal (TG), denoted in dark blue. An adversary seeks to determine which of the three possible blue goals is the agent's true goal. If the agent takes the direct path (denoted in solid green) from the start to the true goal, the adversary can easily correctly infer its intentions. If the agent, however, first begins moving towards one of the false goals (dashed yellow paths) the adversary may assume that this goal is the agent's objective, and may only have very little time to change its opinion, after the agent does not end up going into a false goal, and instead turns towards the true goal (dashed green paths). Thus, the agent's top and bottom paths are deceptive.

In this paper, we describe the agent's *strategy* as controlled movements on a state space over a period of time, potentially constrained by a priori specifications on the agent's behavior. The agent's objective is encoded through a reward function that the agent seeks to maximize.

Existing work on formalizing deception is largely limited to application-specific settings (e.g., [2] considers deception in the context of cyber-attacks on networked systems, [8] deals with pursuit-evasion scenarios, and [9] discusses area denial), or two-player games (see, e.g., [10]–[14]). While the framework of the latter set of papers — and particularly [13] and [14], which deal with multi-stage games — bears most similarity to our work, such papers assume that the two

M. Ornik is with the Institute for Computational Engineering and Sciences, University of Texas at Austin. U. Topcu is with the Institute for Computational Engineering and Sciences and the Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin. Emails: {mornik@ices.utexas.edu, utopcu@utexas.edu}

This work was funded by grants W911NF-16-1-0001 from the Defense Advanced Research Projects Agency, FA8650-15-C-2546 from the Air Force Research Laboratory, and W911NF-15-1-0592 from the Army Research Office.

players are symmetric: each player has its own objective and can take different actions during the system run, potentially exploiting the other player's lack of knowledge about the state space (as in [13]) or irrationality (as in [14]). We are primarily interested in an attacker-defender setting where there is an asymmetry between the attacker (i.e., agent), who takes actions during the system run, and the adversary, who solely observes the attacker, learns from a predetermined model, and influences the agent's collected rewards based on its beliefs, but does not make any decisions itself. Additionally, we allow for the possibility that the adversary is able to perfectly observe the agent's state and actions at all times — the unknown element are the agent's intentions.

The primary contribution of this paper is to provide a formal definition of deception and deceptive strategies within the framework of optimal control, as well as discuss optimal design of deceptive strategies for a wide class of scenarios. Since a critical component of deception is modifying adversary's opinions, we begin our discussion by formalizing the notions of adversary's *belief space* and *belief-induced rewards*. We define these notions within the setting of Markov decision processes in Section II, while remarking that analogous definitions can be introduced for general control systems. Section III uses belief space and belief-induced rewards to define deception and optimal deceptive strategies, and provides the design of an optimal deceptive strategy for a memoryless adversary. As it is natural that the deceiving agent might not know everything about adversary's beliefs, Section IV discusses several classes of lack of agent's knowledge about the adversary, and considers design of optimal deceptive strategies for such classes. Section V is dedicated to a detailed analysis of the running example. Finally, in Section VI we briefly describe future work concerning deceptive scenarios, design of optimal deceptive strategies, and models of the agent's lack of knowledge.[1]

II. Beliefs

This section presents the first contribution of our paper: it defines the notion and role of beliefs in control systems with reward-based objectives. For the sake of simplicity of notation and conservation of space, the control systems discussed in this paper will be described by discrete-state, finite-time Markov decision processes. All definitions of this section can be directly generalized for general control systems evolving on any abstract state space, with any discrete or continuous set of times.

A *Markov decision process (MDP)* is defined by $\mathcal{M} = (S, A, P)$, where the state space S and action set A are finite, and the agent's dynamics on S are given by a stochastic model $\mathbb{P}(s_{t+1} = s') = P(s_t, a, s')$, where the transition probability function $P : S \times A \times S \to [0, 1]$ satisfies $\sum_{s' \in S} P(s, a, s') = 1$ for all $s \in S$, $a \in A$.

In order to motivate our definition of deception in a scenario where the agent has a reward-based objective, and an adversary is attempting to learn the agent's intentions and influence its achievement of the objective, we first consider a simplified scenario, where the agent has the same control objective, but the adversary does not exist. We refer to such a scenario as *nominal*.

Assume that MDP $\mathcal{M}$ comes equipped with a *nominal reward* function $R : S \times A \to \mathbb{R}$, assumed to be time-invariant for ease of notation. The control objective of an agent evolving in $\mathcal{M}$ is to maximize its accumulated reward over a period of time.

Definition 1: The *nominal optimal control policy* is given by

$$\underset{a_0,\ldots,a_T \in A}{\operatorname{argmax}} \ \mathbb{E}\left[\sum_{t=0}^{T} R(s_t, a_t)\right], \tag{1}$$

where s_t is the agent's state at time t and a_t is the agent's action at that time.

We note that, in Definition 1, it is possible to add constraints on the choice of actions available to the agent at any given time t, stemming from specifications on the agent's behavior. Thus, instead of allowing that a_t be any element of A at any time t, we may require $a_t \in A_t \subseteq A$. While we omit future references to such constraints in this section, we briefly discuss them in Section III.

Having defined a control objective and a nominal optimal control policy for an agent without the presence of an adversary, we now seek to formalize the adversary's role. In our framework, the adversary has two salient properties:

1) belief about the agent's intentions, which may change over time, and
2) influence of such a belief on the agent's actual collected reward.

We note that we have not concretized the meaning of a *belief* in 1), nor the meaning of an *intention*. Informally, we consider intention to be any property of the agent's task or agent's policy that is important to the adversary. For instance, it may be the agent's objective, agent's next action, or the agent's accumulated reward. A belief is, then, an assertion on the set of agent's intentions.

Formally, we define a *belief* B_t at time t as an element of some domain $\mathcal{B}$, which we refer to as the *belief space*. As stated above, $\mathcal{B}$ can be any set that in some way describes the elements of agent's behavior that are important to the adversary; a particular instantiation of $\mathcal{B}$ depends on the exact setting that we are dealing with. For instance, in the example of cops and deceptive robbers, one possible model — which we revisit later in the paper — is that $\mathcal{B}$ consists of all states that are the robbers' possible goals. Another model would define $\mathcal{B}$ as the set of probability distributions on the set of possible goals, thus allowing that the cops are uncertain about the robbers' goal.

Having formally described the adversary's beliefs, we now move to property 2) above. In our running example, the cops' knowledge about the robbers' intentions will change the payoff that the robbers receive for reaching the objective: instead of robbing the bank, they will be caught. We generalize this notion by introducing *belief-induced rewards*.

[1] A longer version [15] of this paper is available online.

Definition 2: A *belief-induced reward* function is a map given by $L : S \times \mathcal{B} \times A \to \mathbb{R}$.

While there is no requirement that the nominal reward R and belief-induced reward L from Definition 2 are in any way related, the motivating setting would imply that L is in some way a modification of R. However, we will not be a priori assuming any formal relationship between L and R in the theoretical results of this paper. Instead, we will be dealing with optimal design of the agent's policy with respect to a reward function L, while making use of the relationship between R and L when analyzing the effectiveness of deception for our running example, explored in Section V.

Analogously to the nominal optimal control policy (1), a belief-induced reward yields an optimal belief-induced control policy.

Definition 3: The *optimal belief-induced control policy* is given by

$$\underset{a_0,\ldots,a_T \in A}{\operatorname{argmax}} \sum_{t \in \mathcal{T}} L(s_t, B_t, a_t), \tag{2}$$

where s_t is the agent's state at time t, a_t is the agent's action at that time, and B_t is the adversary's belief at time t.

We now use the above definitions to more precisely define our running example. We assume that the agent has a single true goal $TG \in S$. The adversary possesses a set of possible agent's goals, but does not know which one is the true goal. If the adversary's belief of the agent's true goal is incorrect, i.e., the agent successfully fooled the adversary, then the agent collects a positive reward for reaching the true goal. On the other hand, if the adversary's belief is correct, then the agent collects a negative reward for reaching the true goal. The above setup is formalized as follows.

Example 4 (Cops and deceptive robbers): Let $\mathcal{M} = (S, A, P)$ be an MDP. Let $TG \in S$ be the agent's true goal, and $\{G_1, \ldots, G_k\} \subseteq S$, with $TG \in \{G_1, \ldots, G_k\}$, be the set of states that the adversary believes are the possible agent's objectives. Define $\mathcal{B} = \{G_1, G_2, \ldots, G_k\}$. Let the belief-induced reward $L : S \times \mathcal{B} \times A \to \mathbb{R}$ be defined by

$$L(s, B, a) = \begin{cases} L^+ & \text{if } s = TG, B \in \mathcal{B} \backslash \{TG\}, \\ L^- & \text{if } s = TG, B = TG, \\ 0 & \text{if } s \in S \backslash \{TG\}, \end{cases} \tag{3}$$

where $L^+ > 0$ and $L^- < 0$. The optimal belief-induced policy for the robbers is then given by (2).

III. Deception and Optimal Deceptive Policy

Having established the notion of an optimal belief-induced strategy, we now consider the problem of determining an optimal belief-induced policy (2) for the agent. Throughout the paper, we assume that the adversary does not actively make decisions during the system run; its beliefs change according to a predetermined mechanism. While the adversary's belief evolves on the belief space $\mathcal{B}$, the adversary has no goal that it attempts to reach.

As the adversary's beliefs may change over time, the MDP $\mathcal{M}$ along with adversary's belief dynamics defines a control system $\mathcal{M}_{\mathcal{B}}$ on state space $S \times \mathcal{B}$. The problem of finding an optimal belief-induced policy (2) can be understood as a reward maximization problem in $\mathcal{M}_{\mathcal{B}}$. At every time, the agent may commit an action with the purposeful desire to modify the adversary's belief in a way that will lead to an increase in the agent's collected reward. Thus, we define (2) to be the *optimal deceptive policy*. Generally, we define *deception* as any exploitation of prior or side information that the agent may have on L and dynamics of $\mathcal{B}$ to better design its control policy. For instance, in the context of our running example, robbers use deception if they do not merely go towards its true goal in the most direct path, but in some way exploit the knowledge that the cops are attempting to learn the robbers' goal, and may reduce the robbers' reward if they learn the goal correctly.

Without additional assumptions on belief dynamics, the problem of finding an optimal deceptive policy (2) is a general optimal control problem on a product space $S \times \mathcal{B}$, with a combination of MDP-induced motion of an agent in S, and belief dynamics in $\mathcal{B}$. In the remainder of the text, we assume that (i) set $\mathcal{B}$ is finite, and (ii) the adversary's learning mechanism is *memoryless*, i.e., given by $\mathbb{P}(B_{t+1} = B) = f(s_t, B_t, a_t, B)$, where function $f : S \times B \times A \times B \to [0, 1]$ satisfies $\sum_{B'} f(s, B, a, B') = 1$. Then, the problem of finding an optimal deceptive policy (2) is given as follows.

Problem 5 (Optimal deception): Let $\mathcal{M} = (S, A, P)$, $\mathcal{B}$, $f : S \times B \times A \times B \to [0, 1]$, and $L : S \times \mathcal{B} \times A \to \mathbb{R}$ be as above, and let $T \geq 0$, $s_0 \in S$, $B_0 \in \mathcal{B}$. Find a control policy π^* such that

$$\pi^* = (\pi_0^*, \ldots, \pi_T^*) = \underset{\pi}{\operatorname{argmax}} \mathbb{E}\left[\sum_{t=0}^{T} L\left(s_t, B_t, \pi_t\right)\right], \tag{4}$$

subject to

$$\mathbb{P}(s_{t+1} = s) = P(s_t, \pi_t, s), \tag{5}$$

$$\mathbb{P}(B_{t+1} = B) = f(s_t, B_t, \pi_t, B). \tag{6}$$

Problem 5 can be interpreted as a reward maximization problem on the MDP $\overline{\mathcal{M}} = (S \times \mathcal{B}, A, \overline{P})$, where $\overline{P}$ is given by (5)–(6). With such a model, it is well-known that the optimal policy π^* is memoryless, i.e., agent's action π_t^* at time t depends solely on s_t, B_t, and t, and Problem 5 is solvable by previously known and extensively discussed methods (see, e.g., [16] for a detailed study).

Remark 6: As mentioned, Problem 5 may be equivalently posed in the case where the set of all permissible policies π in (4) is the proper subset of the set of all policies taking values in A. Notably, such a constraint may come from a requirement that the agent follows a temporal logic specification; we direct the reader to a particularly detailed exposition given in [17]. A constraining specification may significantly lower the rewards that the agent is able to collect, making deception less effective. The extent to which a specification will make deception less effective is related to the extent to which it "clashes" with the behavior needed for an agent to successfully deceive the adversary. We illustrate

impact of different specifications through an example in Section V-C.

Remark 7: The assumption that belief dynamics are memoryless holds for a wide variety of estimation techniques, most notably, online inverse reinforcement learning [18]. In online inverse reinforcement learning, the parameter estimate $\hat{p}_t$ for a parametrized reward function R_p is updated at every time step by setting $\hat{p}_{t+1} = \hat{p}_t + \alpha \nabla l(\hat{p}_t)$, where $l(p)$ is the estimated log-likelihood of the agent performing action a_t at state s_t, if the reward function is given by R_p.

Having in mind the adversarial nature of our scenario, we now concentrate on developing deceptive policies in the case where the agent does not have full knowledge about the adversary's behavior.

IV. Lack of Knowledge

Problem 5 poses the question of designing the optimal deceptive policy as a problem of finding an optimal policy on an MDP. While such a problem can be solved by straightforward application of previously-known methods [16], its solution requires agent's full knowledge of the dynamics and the reward function L on the MDP $\overline{\mathcal{M}} = (S \times \mathcal{B}, A, \overline{P})$. Possession of such knowledge may not be realistic in adversarial settings: for instance, in the cops and robbers example, cops may not announce their current beliefs or operating procedures. Thus, it may be impossible for the agent to devise an objectively optimal deceptive policy; instead, the goal is to devise a deceptive policy that is optimal *given the agent's imperfect knowledge*.

In this section, we consider three fundamental categories in which the agent may lack precise knowledge about $\overline{\mathcal{M}}$ or L. In particular, the agent might have knowledge about:

1) *What the adversary thinks* — the adversary's belief B_t at every time step.
2) *How the adversary thinks* — the dynamics underlying B_t.
3) *What the adversary does* — how the true reward L depends on the adversary's beliefs B.

For the sake of exposition, we treat the above three limitations separately, with the understanding that it is naturally possible that the knowledge about $\overline{\mathcal{M}}$ is limited in more than one way at the same time. We also emphasize that we are presenting merely several typical cases, and not all possible variants of the agent's lack of knowledge. Let us now consider each of the above possible knowledge limitations.

A. Unknown Beliefs

If beliefs B_t are unknown to the agent, the system state (s_t, B_t), evolving in $S \times \mathcal{B}$, is partially observable. Namely, the agent knows s_t at every time, but may not know B_t. In our running example, example, lack of knowledge about the adversary's beliefs naturally arises from the robbers not knowing the cops' estimate of their goal.

Lack of agent's knowledge about the adversary's belief places the belief-induced MDP $\overline{\mathcal{M}}$ in the class of mixed-observability MDPs [19], where the entirely observable part of the system state is s_t, and the entirely unobservable part is B_t. We will not describe the mixed-observability MDP framework in detail; we refer the reader to [19] for a broad study. However, for the sake of phrasing the problem of designing an optimal deceptive policy, we note the fact that the agent cannot observe beliefs B_t does *not* mean that the agent has no knowledge of B_t whatsoever. Namely, if the agent possesses an initial probability distribution Pr_0 on possible $B_0 \in \mathcal{B}$, it may use the belief evolution (6) to obtain a probability distribution Pr_1 for $B_1 \in \mathcal{B}$, and by continuing onwards, distributions Pr_t for B_t, for all $t \geq 0$.

The initial probability distribution Pr_0 depends on the agent's knowledge about the adversary's initial belief. If the agent has no knowledge about B_0, i.e., finds all beliefs in $\mathcal{B}$ equally likely, the initial probability distribution Pr_0 is given by $Pr_0(B) = 1/|\mathcal{B}|$. On the other hand, if the agent knows that $B_0 = B'$ for a particular $B' \in \mathcal{B}$, Pr_0 is given by $Pr_0(B') = 1$, $Pr_0(B) = 0$ for all $B \neq B'$.

The problem of determining an optimal deceptive policy without belief observations is thus formalized as follows:

Problem 8 (Optimal deception with unknown beliefs): Let $\overline{\mathcal{M}}$, L, and a probability distribution $Pr_0 : \mathcal{B} \to [0, 1]$ be as defined previously, and let $T \geq 0$, $s_0 \in S$.

Find a control policy π^*, where π^*_t depends on $s_0, \ldots, s_t$ and $Pr_0, \ldots, Pr_t$, such that

$$\pi^* = \operatorname*{argmax}_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \sum_{B \in \mathcal{B}} Pr_t(B) L\left(s_t, B, \pi_t\right)\right], \qquad (7)$$

subject to

$$\begin{aligned} \mathbb{P}(s_{t+1} = s) &= P(s_t, \pi_t, s), \\ Pr_t(B) &= \sum_{B' \in \mathcal{B}} Pr_t(B') f(s_t, B', \pi_t, B). \end{aligned} \qquad (8)$$

We emphasize that in the above problem, unlike in the other problems considered throughout this paper, π^*_t needs to depend solely on the history of agent's positions and probability distributions on possible adversarial beliefs, as beliefs $B_0, \ldots, B_t$ are not directly known to the agent.

Problem 8 can be solved using algorithms for mixed-observability and partially observable MDPs (see, e.g., [19]–[21]). However, such algorithms are usually computationally infeasible [22], and algorithms for approximately optimal policies have been developed; in Section V we employ one such algorithm [23] in the context of our running example.

B. Uncertain Belief Dynamics

If the belief update mechanism f in (6) is not entirely known to the agent, $\overline{\mathcal{M}}$ is transformed into an MDP with uncertain transition probabilities [24]. In other words, the agent knows that

$$f \in \mathcal{F} = \{f^i : S \times \mathcal{B} \times A \times \mathcal{B} \to [0, 1] \mid i \in I\}, \qquad (9)$$

where I is an index set, and all f^i satisfy $\sum_{B' \in \mathcal{B}} f^i(s, B, a, B') = 1$ for all $s \in S$, $B \in \mathcal{B}$, $a \in A$. In such a case, the interest is to find a robust optimal policy, i.e., a policy that produces the best results for "worst-case" dynamics.

There are two variations of this problem: (a) the agent is aware that f, while uncertain, is the same at all times, and (b) the agent assumes that f may change over time, while remaining within $\mathcal{F}$. The latter version thus allows for the possibility that the adversary does not learn in an entirely Markovian way. For this reason, in this paper we choose to describe version (b), while emphasizing that problem (a) can be stated analogously.

Problem 9 (Robust optimal deception with uncertain belief dynamics): Let $\overline{\mathcal{M}}$ and L be as before, and let $T \geq 0$, $s_0 \in S$, $B_0 \in \mathcal{B}$. Let $\mathcal{F}$ be as defined in (9).

Find a control policy π^* such that

$$\pi^* = \operatorname*{argmax}_{\pi} \inf_{f_0,\ldots,f_{T-1} \in \mathcal{F}} \mathbb{E}\left[\sum_{t=0}^{T} L\left(s_t, B_t, \pi_t\right)\right], \quad (10)$$

subject to dynamics

$$\mathbb{P}(s_{t+1} = s) = P(s_t, a, s),$$
$$\mathbb{P}(B_{t+1} = B) = f_t(s_t, B_t, \pi_t, B).$$

Generally, Problem 9 is computationally easier to solve than version (a) of the same problem [25]. For a wide variety of uncertainty sets, algorithms for efficiently computing the solution to Problem 9 have been proposed. We turn the reader's attention to [24]–[26] for standard works that also deal with exploring the relationship between a solution to Problem 9 and a solution to the problem described in (a).

As in the case of unobservable beliefs, we present an example of an optimal policy for uncertain belief dynamics within our running example in Section V. Such a setting is naturally motivated by the robbers not knowing the exact mechanism that the cops use to learn about the robbers' intentions.

C. Uncertain Belief-Induced Reward

If the agent's knowledge of how the adversary's beliefs change the nominal reward R into L is not precise, $\overline{\mathcal{M}}$ is transformed into an MDP with uncertain rewards: it is known that

$$L \in \mathcal{L} = \left\{L^i : S \times \mathcal{B} \times A \to \mathbb{R} \mid i \in I\right\}, \quad (11)$$

where I is an index set. For $s \in S$, $B \in \mathcal{B}$, and $a \in A$, we denote $\mathcal{L}(s, B, a) = \{L^i(s, B, a) \mid i \in I\}$.

Analogously to the setting of uncertain belief dynamics, there are two basic cases: (a) the rewards, while unknown to the agent, are known to be fixed before the system run, and (b) the rewards are allowed to be time-varying, while staying within $\mathcal{L}$. The latter case appears more naturally in our running example: the robbers' gains may differ every time they rob a bank. We formalize it as follows.

Problem 10 (Robust optimal deception with uncertain rewards): Let $\overline{\mathcal{M}}$ be as before, and let $T \geq 0$, $s_0 \in S$, $B_0 \in \mathcal{B}$. For every $s \in S$, $B \in \mathcal{B}$, $a \in A$, let $\mathcal{L}$ be as defined in (11). Assume that all $\mathcal{L}(s, B, a)$ are bounded from below.

Find a control policy π^* such that

$$\pi^* = \operatorname*{argmax}_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \inf \mathcal{L}\left(s_t, B_t, \pi_t\right)\right],$$

subject to dynamics (5)–(6).

With an obvious proof, Problem 10 reduces to a problem of finding an optimal policy in an MDP.

Proposition 11: Control policy π^* is a solution of Problem 10 if and only if it is a solution of Problem 5, with L in Problem 5 replaced by $\inf \mathcal{L}$.

In contrast to Problem 10, variant (a) described above is not readily reducible to finding an optimal MDP policy. Its setting is known as an imprecise-reward MDP [27]. Robust optimal policies for imprecise-reward MDPs are usually based on solving a minimax regret optimization problem; we turn the reader's attention to [27], [28], and the references contained therein.

We illustrate the discussion of all the above scenarios of lack of knowledge in the subsequent section.

V. Cops and Deceptive Robbers

In this section, we provide a more thorough analysis of our running example, previously described in Fig. 1 and Example 4. While framed slightly differently, such an example is a deceptive variant of the well-known *heaven and hell* example [29], [30], where heaven becomes hell if the adversary finds out its location correctly.

We assume that the agent moves in a gridworld S, shown in Fig. 2. The actions available to the agent at any time are to go one tile north, south, east, west, or stay in place. (When the agent is at the edge of the grid state space, actions that would make it leave the state space are not available, or result in a prohibitively negative reward.) Given the agent's choice of action, the agent moves deterministically, i.e., moves in the desired direction with probability 1.

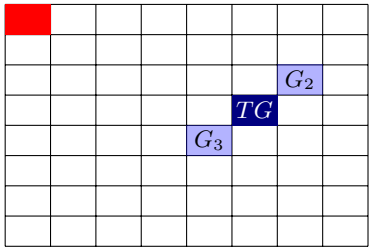

Fig. 2. An example of the agent's state space. The agent's true goal ($TG = G_1$) is marked in dark blue. The agent knows the position of the true goal, while the adversary only knows that the goal is one of the three blue tiles $\{G_1, G_2, G_3\}$ on the map. The red tile is the agent's starting position.

The agent's nominal reward R is given by

$$R(s, a) = \begin{cases} R^+ & \text{if } s = TG, \\ 0 & \text{otherwise,} \end{cases} \quad (12)$$

where $R^+ > 0$. Since the agent's goal is to maximize its reward over a time horizon, the nominal optimal policy π^{*N} of the agent is to take the shortest path to TG, and then remain at TG for the remaining portion of the system run.

The adversary knows that the agent's reward function R is given by (12). However, the adversary only knows of k candidate locations for TG, i.e., $G_1, \ldots, G_k$, where $TG \in$

$\{G_1, \ldots, G_k\}$. Such a situation is illustrated in Fig. 2, where $k = 3$. We assume that the adversary has the knowledge of the agent's position and action at all times.

As proposed in Section II, since the adversary is missing information about the exact location of the true goal, its belief space $\mathcal{B}$ can be given by $\mathcal{B} = \{1, \ldots, k\}$, where $B_t = i$ indicates that, at time t, the adversary believes that $TG = G_i$. The adversary uses the following memoryless mechanism for updating its beliefs:

$$\mathbb{P}(B_{t+1} = i | s_t, B_t, a_t) = m + \begin{cases} 0 & \text{if } B_t \neq i, \\ 1 - p & \text{if } B_t = i, \end{cases} \tag{13}$$

with

$$m = \begin{cases} 0 & \text{if } d(s_{t+1}, G_i) \geq d(s_t, G_i) \text{ and } s_{t+1} \neq G_i \\ & \quad \text{and } (B_t \neq i \text{ or } c_t \neq 0), \\ \frac{p}{c_t} & \text{if } d(s_{t+1}, G_i) < d(s_t, G_i) \text{ or } s_{t+1} = G_i, \\ p & \text{if } B_t = i \text{ and } c_t = 0, \end{cases} \tag{14}$$

where d is the 1-norm distance between two states, c_t is the number of all $i \in \{1, \ldots, k\}$ such that $d(s_{t+1}, G_i) < d(s_t, G_i)$ or $s_{t+1} = G_i$, and p is a fixed parameter in $[0, 1]$. We note that s_{t+1} is a deterministic function of (s_t, a_t), so the adversary's dynamics do not use any knowledge not available at the current time.

In plain words, (13)–(14) state that the adversary's belief remains the same with probability $1 - p$. The remaining p are divided equally among all goal candidates whose distance from the agent reduced in the last step. While such a learning method is simple, it guarantees that the adversary will eventually, with probability 1, correctly learn the position of the true goal if the agent uses a nominal optimal control policy.

Finally, we define the belief-induced reward L. It is modified from (12) along the lines of (3):

$$L(s, B, a) = \begin{cases} R^+ & \text{if } s = TG \text{ and } G_B \neq TG, \\ R^- & \text{if } s = TG \text{ and } G_B = TG, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

where R^+ is the same as in (12), and $R^- < 0$.

A. Optimal Deceptive Policy

As mentioned, belief update mechanism (13)–(14) ensures that, if the agent is following the nominal optimal control policy π^{*N}, the adversary will eventually correctly learn the position of the true goal with probability 1. Thus,

$$\lim_{T \to +\infty} \mathbb{E}\left[\sum_{t=0}^{T} L(s_t, B_t, \pi_t^{*N})\right] = -\infty.$$

As a consequence, there exists a need for determining the optimal belief-induced, or deceptive, policy π^{*O} that takes into account the adversary's beliefs B_t. As outlined in Section III, π^{*O} is an optimal control policy for an MDP $\overline{\mathcal{M}} = (S \times \mathcal{B}, A, \overline{P})$ given by deterministic dynamics of the agent on S, dynamics (13)–(14) on $\mathcal{B}$, and reward function (15). It can be easily constructed using any of the available algorithms for optimal control on MDPs. Fig. 3 presents the mean reward

$$\overline{L}(T') = \sum_{t=0}^{T'} \frac{L(s_t, B_t, \pi_t^{*O})}{T'} \tag{16}$$

obtained by the agent using such an optimal deceptive policy for $T' \leq T = 2000$, with parameters $p = 0.1$, $R^+ = 10$, $R^- = -10$, and the state space given as in Fig. 2.

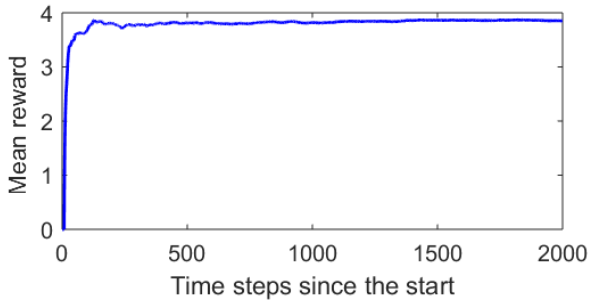

Fig. 3. The mean reward $\overline{L}$, collected by the agent when using the optimal deceptive policy, averaged over 100 system runs.

As Fig. 3 shows, the agent that uses the optimal deceptive policy π_O^* collects, on average, a positive reward. The exact value of such a reward depends on the value of the adversary's belief change probability p; nonetheless, the optimal deceptive policy always generates significant gains compared to the nominal optimal policy.

To illustrate the deceptive strategy π^{*O}, Fig. 4 shows the exact rewards that the agent collected during the first 100 time steps in one system run. The agent starts off by collecting a reward of 0 for the first 8 steps, until it reaches TG. It then proceeds to remain at this goal until the adversary correctly deduces that TG is indeed the agent's true goal. After the adversary realizes the true goal (and the agent collects a reward of -10), the agent leaves and attempts to trick the adversary by feigning that another one of the candidate goals is its goal. Once the adversary is tricked, the agent moves again to TG, and the process repeats. We remark that, on an intuitive level, such behavior is indeed deceptive: it relies on repeatedly attempting to instill incorrect beliefs in an adversary, while still receiving positive rewards as often as possible.[2]

B. Optimal Deception with Imperfect Knowledge

We proceed by developing deceptive policies for each of the cases of imperfect knowledge discussed in Section IV.

In the case when the agent does not have any knowledge of the adversary's beliefs B_t, we showed in Section IV-A that an optimal deceptive policy is given by an optimal control policy of a mixed-observability MDP. As mentioned, such policies are generally difficult to compute. In our simulation, we used a randomized approximation of an optimal policy

[2]A video illustrating a typical system run is available at `https://bit.ly/2ofmm32`.

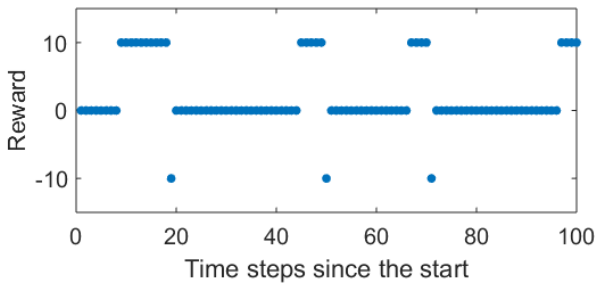

Fig. 4. Rewards obtained during the beginning of a system run.

based on combining optimal actions for MDPs where beliefs are known, with weights corresponding to the probability distribution of the beliefs [23]. The light blue graph in Fig. 5 describes attained mean rewards $\overline{L}$. As expected, the deceptive policy developed without belief observations performs worse than the optimal deceptive policy with perfect knowledge. However, it still yields positive rewards.

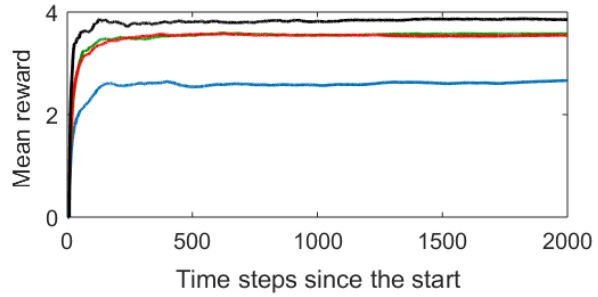

Fig. 5. The mean reward $\overline{L}$, averaged over 100 system runs, collected using optimal policies with imperfect knowledge. The light blue graph depicts the rewards obtained by an approximately optimal policy without belief observations. The red graph depicts the rewards obtained by an optimal policy for the case of an uncertain learning parameter. The green graph depicts the rewards obtained by an optimal policy for the case where the collected rewards are a priori uncertain. The black graph depicts the optimal deceptive policy with complete knowledge of the adversary.

Let us now present the simulation results for the case of uncertain transition probabilities and uncertain rewards. In the former scenario, the agent does not know the true value of learning parameter p, and knows that it is between 0.05 and 0.2; it expects that p may change at every time step, and designs the robust worst-case policy as a solution to Problem 9. In the simulation, p is set to constant 0.1, as before. In the latter scenario, the agent does not know the true reward that it will collect upon reaching TG if the adversary did not learn its goal correctly, and believes it to be anywhere between 1 and 20. The agent expects that this reward may change every time it reaches TG, and designs the robust worst-case policy as a solution to Problem 10. In the simulation, the collected reward is set to always equal 10, as before.

The rewards attained by worst-case optimal policies are shown in red and green, respectively, in Fig. 5. These policies perform worse than the optimal policy π^*_O designed with complete information. However, the difference is less stark than for the case of unknown beliefs. Such a property is a consequence of the simplicity of the reward function L; regardless of how quickly it believes the adversary is learning, or how large of a reward it may collect at goal, the agent has little motivation but to continue with the general behavior of reaching TG, waiting until the adversary learns of its goal, then moving away, tricking the adversary, and repeating the process.

C. Optimal Deception with Temporal Logic Specifications

Finally, let us briefly return to the setting where the agent is required to obey additional specifications while executing a deceptive policy. We introduce two specifications:

1) the agent is not allowed to visit either of the two false goals G_2 and G_3, described by light blue tiles in Fig. 2,
2) the agent is not allowed to visit G_2, but can visit G_3.

As described in Remark 6, such specifications yield constraints on the optimal deceptive policy (4). In this case, the constraints are simple, and we can easily compute the corresponding optimal deceptive policies π^{*O1} and π^{*O2}, respectively. Fig. 6 shows the mean rewards (16).

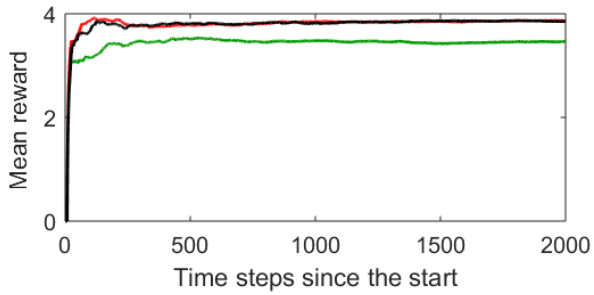

Fig. 6. The mean reward $\overline{L}$, averaged over 100 system runs, collected using optimal policies with additional specifications. The graph in the green describes the rewards obtained by policy π^{*O1}. The graph in red describes the rewards obtained by π^{*O2}. The graph in the black describes the rewards obtained by π^{*O} — an optimal deceptive policy without any specifications.

We note that the rewards obtained by policy π^{*O1} are lower than those obtained by π^{*O}. Such a difference arises from the fact that the specification underlying π^{*O1} is substantially interfering with the agent's ability to deceive the adversary. Not being able to go into either of the false goals makes it harder to convince the adversary that one of those goals is the agent's objective. On the other hand, the rewards obtained by π^{*O2} are effectively the same as the rewards for π^{*O}. While the specification underlying π^{*O2} restricts the agent's actions, it does not interfere with its ability to deceive the adversary — if the agent is unable to visit one of the false goals, it will simply try to convince the adversary that the other false goal is the true objective, without any loss in the quality of deception.

VI. Conclusions and Future Work

This paper provided a description of deception in an abstract optimal control scenario. The framework that we presented rests on the introduction of the belief space of an

adversary trying to learn agent's intentions, as well as on encoding the adversary's influence on the objective through belief-induced rewards. Assuming that the agent evolves on an MDP, and that the adversary's learning process is memoryless, the problem of optimal design of a deceptive policy is an optimal control problem in an MDP on a product state space. Such a result can be extended to describe the case when the agent lacks knowledge about the adversary as a problem on partially observable or uncertain MDPs. As discussed on a cops and robbers scenario, the proposed definitions of deception and deceptive policies correspond to a common intuition behind deceptive behavior.

We list several avenues of future work.

- *Deception of complex, realistic adversaries:* An adversary that operates without memory and has only finitely many potential beliefs can be easily deceived by a single "trick" performed time and time again during the system run, as exhibited in Section V.
- *Learning about the adversary:* Even if the agent might not know the adversary's belief, it can potentially partly infer it from the collected rewards. For instance, in our running example, if the agent positioned at TG collects a negative reward, it can immediately deduce that the adversary believes that its current location is its true goal. The agent should perform such deduction and, balanced against a possible short-term decrease in rewards, take actions to learn the adversary's beliefs.
- *Explainable deception:* Determining, and describing in understandable terms, the features of the agent's policy that cause deception to succeed will enable learning from deceptive behavior in one scenario in order to determine deceptive behavior in a similar scenario.
- *Broader control objectives:* Objectives encoded in temporal logic specifications are particularly intuitive for a variety of applications. In Remark 6 and Section V-C, we considered temporal logic specifications solely as an addition to the reward objective. Such a framework is a step towards the setting in which the agent's sole objective is given by a temporal logic specification.

ACKNOWLEDGMENT

The authors thank Steven Carr for coding and running the simulation of a deceptive strategy in the scenario where the observer beliefs are unknown within Section V-B.

REFERENCES

[1] T. E. Carroll and D. Grosu, "A game theoretic investigation of deception in network security," *Security and Communication Networks*, vol. 4, no. 10, pp. 1162–1172, 2011.

[2] Z.-H. Pang and G.-P. Liu, "Design and implementation of secure networked predictive control systems under deception attacks," *IEEE Transactions on Control Systems Technology*, vol. 20, no. 5, pp. 1334–1342, 2012.

[3] J. Shim and R. C. Arkin, "Robot deception and squirrel behavior: A case study in bio-inspired robotics," Georgia Institute of Technology, Tech. Rep. ADA608845, 2014.

[4] L. D. Whitley, "Fundamental principles of deception in genetic search," *Foundations of Genetic Algorithms*, vol. 1, pp. 221–241, 1991.

[5] W. McEneaney and R. Singh, "Deception in autonomous vehicle decision making in an adversarial environment," in *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2005.

[6] B. Whaley, "Stratagem: deception and surprise in war," Center for International Studies, Massachusetts Institute of Technology, Tech. Rep. C/69-9, 1969.

[7] S. Metts, "An exploratory investigation of deception in close relationships," *Journal of Social and Personal Relationships*, vol. 6, no. 2, pp. 159–179, 1989.

[8] Y. Yavin, "Pursuit-evasion differential games with deception or interrupted observation," *Computers & Mathematics with Applications*, vol. 13, no. 1, pp. 191–203, 1987.

[9] D. A. Castanon, M. Pachter, and P. R. Chandler, "A game of deception," in *43rd IEEE Conference on Decision and Control*, 2004, pp. 3364–3369.

[10] J. P. Hespanha, Y. S. Ateşkan, and H. H. Kızılocak, "Deception in non-cooperative games with partial information," in *2nd DARPA-JFACC Symposium on Advances in Enterprise Control*, 2000.

[11] A. R. Wagner and R. C. Arkin, "Robot deception: Recognizing when a robot should deceive," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 2009, pp. 46–54.

[12] R. C. Arkin, P. Ulam, and A. R. Wagner, "Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 571–589, 2012.

[13] D. Ettinger and P. Jehiel, "A theory of deception," *American Economic Journal: Microeconomics*, vol. 2, no. 1, pp. 1–20, 2010.

[14] R. Singh, "Deception in two-player zero-sum stochastic games: Theory and application to warfare games," Ph.D. dissertation, University of California, San Diego, 2006.

[15] M. Ornik and U. Topcu, "Deception in optimal control," *arXiv:1805.03090 [math.OC]*, 2018.

[16] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2005.

[17] C. Belta, B. Yordanov, and E. Aydin Gol, *Formal Methods for Discrete-Time Dynamical Systems*. Springer, 2017.

[18] K. Li and J. W. Burdick, "Online inverse reinforcement learning via Bellman Gradient Iteration," *arXiv:1707.09393 [cs.RO]*, 2017.

[19] S. C. Ong, S. W. Png, D. Hsu, and W. S. Lee, "POMDPs for robotic tasks with mixed observability," in *Robotics: Science and Systems*, 2009.

[20] D. Braziunas, "POMDP solution methods," University of Toronto, Tech. Rep., 2003.

[21] K. P. Murphy, "A survey of POMDP solution techniques," University of California, Berkeley, Tech. Rep., 2000.

[22] I. Chadès, J. Carwardine, T. G. Martin, S. Nicol, R. Sabbadin, and O. Buffet, "MOMDPs: A solution for modelling adaptive management problems," in *26th AAAI Conference on Artificial Intelligence*, 2012.

[23] S. Carr, N. Jansen, R. Wimmer, J. Fu, and U. Topcu, "Human-in-the-loop synthesis for partially observable Markov decision processes," in *2018 American Control Conference*, 2018.

[24] J. K. Satia and R. E. Lave, Jr., "Markovian decision processes with uncertain transition probabilities," *Operations Research*, vol. 21, no. 3, pp. 728–740, 1973.

[25] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[26] C. C. White, III and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 42, no. 4, pp. 739–749, 1994.

[27] E. Delage and S. Mannor, "Percentile optimization in uncertain Markov decision processes with application to efficient exploration," in *24th International Conference on Machine Learning*, 2007, pp. 225–232.

[28] K. Regan and C. Boutilier, "Robust online optimization of reward-uncertain MDPs," in *22nd International Joint Conference on Artificial Intelligence*, 2011, pp. 2165–2171.

[29] H. Geffner and B. Bonet, "Solving large POMDPs using real time dynamic programming," in *AAAI Fall Symposium on POMDPs*, 1998.

[30] T. Osogami, "Robust partially observable Markov decision process," in *32nd International Conference on Machine Learning*, 2015, pp. 106–115.