A Unifying Framework for Observer-Aware Planning and its Complexity

Shuwa Miura[1] **Shlomo Zilberstein**[1]

[1]College of Information and Computer Sciences , University of Massachusetts Amherst

Abstract

Being aware of observers and the inferences they make about an agent's behavior is crucial for successful multi-agent interaction. Existing works on observer-aware planning use different assumptions and techniques to produce observer-aware behaviors. We argue that observer-aware planning, in its most general form, can be modeled as an Interactive POMDP (I-POMDP), which requires complex modeling and is hard to solve. Hence, we introduce a less complex framework for producing observer-aware behaviors called Observer-Aware MDP (OAMDP) and analyze its relationship to I-POMDP. We establish the complexity of OAMDPs and show that they can improve interpretability of agent behaviors in several scenarios.

1 INTRODUCTION

Reasoning about the beliefs of observers is ubiquitous in our daily lives. For example, consider a scenario where an autonomous vehicle (AV) and a pedestrian are approaching a crosswalk. The AV may optimize travel time and approach the crosswalk at high speed before stopping. The pedestrian, however, may feel unsafe when the vehicle is approaching the crosswalk at high speed. If the AV is aware of the perspective of the pedestrian, it may slow down further away from the crosswalk to assure the pedestrian that it plans to stop. We call this kind of behavior an *observer-aware* behavior. Observer-aware behaviors include explicit communication to convey intentions, for example, using hand gestures as well as implicit communication through behaviors. In this paper, we focus on developing a general, disciplined approach for *observer-aware* planning.

Several existing frameworks offer different approaches to produce different kinds of observer-aware behaviors. The AV example illustrates *legible* behavior (e.g. Dragan et al., 2013), which implicitly conveys intentions via the choice of actions. Similarly, *explicable* behaviors (e.g. Zhang et al., 2017) conform to observers' expectations. *Deceptive* behaviors (e.g. Dragan et al., 2015; Masters and Sardina, 2017) hide agents' intentions or actively deceive observers. *Predictable* behaviors enable observers to predict future actions (e.g. Fisac et al., 2020). Agents can also express their *(in)capability* via the choice of their actions (e.g. Kwon et al., 2018). While there have been several attempts to combine different kinds of observer-aware behaviors (Dragan and Srinivasa, 2013; Strouse et al., 2018; Chakraborti et al., 2019; Kulkarni et al., 2019), there is no unifying framework that reveals the relationships among the approaches and the complexity of the problem.

In this paper, we introduce a unified framework for observer-aware planning called OAMDP and illustrate that OAMDPs can produce useful forms of observer-aware behavior (Section 3). OAMDPs might seem similar to Partially Observable Markov decision processes (POMDPs) (Kaelbling et al., 1998), in the sense that both formulations operate on agents' beliefs. We clarify the differences between OAMDPs and POMDPs, most notably that OAMDPs operate on the (assumed) beliefs of observers instead of the beliefs of the acting agent (Section 4). To further motivate the study of OAMDP, we argue that observer-aware planning, in its most general form, can be formulated using a multi-agent model called interactive POMDP (I-POMDP) (Gmytrasiewicz and Prashant, 2005). We then identify the set of assumptions that allow us to reduce an I-POMDP to an OAMDP (Section 5).

We analyze the complexity of OAMDP when the observer is Bayesian (Section 6), showing that it is PSPACE-complete (Theorem 1 and 2) and that it remains NP-hard even when restricted to stationary policies or deterministic environments (Theorem 3). While this places OAMDP in a provably lower complexity class relative to I-POMDP, it also confirms the intractability of the problem. Hence, we show how standard algorithms like UCT (Kocsis and Szepesvári, 2006) can be used to solve OAMDPs and report initial results on several problems of interest (Section 7).

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:610–620.

2 BACKGROUND

MDP A Markov decision process (MDP) models sequential decision making in environments with stochastic effects. An MDP is described by a tuple $M = \langle S, A, T, R, \gamma, \iota \rangle$. S is a set of states. A is a set of actions. $T(s_t, a_t, s_{t+1})$ is the probability of $S_{t+1}{=}s_{t+1}$ when $A_t{=}a_t$ and $S_t{=}s_t$. R is a conditional distribution of reward given s_t, a_t. γ is a parameter called the discount factor. ι is the initial state (we assume WLOG one initial state). The absorbing terminal state always transitions back to itself with zero reward.

A policy (π) describes how to act. We use the following two kinds of policies in the paper. A *stationary policy* is a conditional distribution of actions given a state. When π is deterministic, it is a mapping from S to A. A *history-dependent policy* is a conditional distribution of actions given a history, where a history h_{t+1} is a sequence of state-action pairs up to time t and the last visited state s_{t+1}. The return of a history is the discounted sum of rewards. An optimal policy for an MDP is a policy that maximizes expected return. For a particular state, a value function V_H^π represents the expected return given a policy π up to time step H. When H is finite, we call it a value function for a finite horizon.

3 OBSERVER-AWARE MDP

We define an Observer-Aware Markov Decision Process (OAMDP) as an extension of an MDP that allows the reward to depend on the assumed belief of an observer.

Definition. An OAMDP is a tuple $M = \langle S, A, T, \gamma, \iota, \Theta, B, R \rangle$ where:

- S, A, T, γ, and ι are as in an MDP.
- Θ is a set of *types*, representing a characteristic of the agent such as possible goals, intentions, or capabilities.
- $B : H^* \rightarrow \Delta^{|\Theta|}$ represents the assumed belief of the observer given a history. H^* is the set of all finite histories and $\Delta^{|\Theta|}$ is a simplex on Θ.
- $R : S \times A \times \Delta^{|\Theta|} \rightarrow \mathbb{R}$ describes how desirable it is to take an action given a state and a belief $b \in \Delta(\Theta)$. When the reward depends only on $\Delta^{|\Theta|}$, we abuse the notation slightly and treat R as $\Delta^{|\Theta|} \rightarrow \mathbb{R}$. Note that the reward depends on histories through the beliefs.

A value function for OAMDP is defined as follows:

$$V_H^\pi(s) = \mathbb{E}[\sum_{t=0}^{H} \gamma^t R(B(h_t)) | S_0 = s, \pi].$$

Intuitively, OAMDPs assume a model of how the observer interpret their behaviors (B) and what interpretations are desirable (R).

For the rest of the paper, we assume S,A, and Θ are finite.

3.1 BELIEF UPDATES IN OAMDP

In principle, the belief update B in the definition of OAMDP can be performed by *any* function that maps histories to beliefs. Dragan and Srinivasa (2013) used as B the ratio between the optimal trajectory and the optimal trajectory constrained to include the current position. MacNally et al. (2018) used a common goal recognition formula (Ramírez and Geffner, 2010) as B. Strouse et al. (2018) proposed a reinforcement learning technique to show/hide goals by maximizing/minimizing mutual information between actions and goals given states. While the formulation does not update beliefs explicitly, it can be viewed as implicitly defining beliefs through mutual information.

Despite the generality of B, however, allowing any arbitrary function as B could make the problem intractable as the number of possible histories is exponential in the number of states and actions. Therefore, we need restrictions on the form of B. For example, we can restrict B to be a Bayesian belief update function:

$$Pr(\theta|h_{t+1}) \propto \hat{Pr}(s_{t+1}, a_t|\theta, s_t) Pr(\theta|h_t) \qquad (1)$$

where $\hat{Pr}(s_{t+1}, a_t|\theta, s_t)$ is represented in a tabular fashion. Intuitively, $\hat{Pr}(s_{t+1}, a_t|\theta, s_t)$ represents the probability that the observed agent takes the action a_t and ends up in s_{t+1} given s_t and θ (according to the observer's model). We use $\hat{Pr}$ instead of Pr as this is an assumed model of the observer.

Definition. An OAMDP$_{BU}$ is a special case of OAMDP, where B is performed by Bayesian belief updating (i.e., $M = \langle S, A, T, \gamma, \iota, \Theta, \hat{Pr}, R \rangle$).

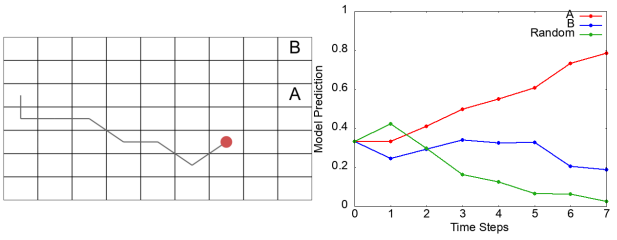
OAMDP with BST Update To illustrate how OAMDP works, we now describe a particular instantiation of OAMDP where the belief update function is according to Baker et al. (2009), referred to as BST belief update.

Baker et al. (2009) examined the relationship between Bayesian reasoning and human goal understanding, showing that human ratings of possible goals correlate well with the posteriors derived using Bayes' rule:

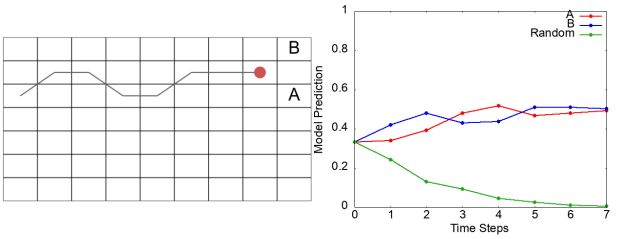
$$Pr(\theta|h_{t+1}) = Pr(\theta|s_t, a_t, s_{t+1}, b_t)$$
$$= \frac{\hat{T}(s_t, a_t, s_{t+1}|\theta)\hat{\pi}(s_t, a_t|\theta)b_t(\theta)}{\sum_{\theta'} \hat{T}(s_t, a_t, s_{t+1}|\theta')\hat{\pi}(s_t, a_t|\theta')b_t(\theta')}.$$

where $\theta{\in}\Theta$ is the type of the agent, b_t is the previously held belief on each θ, and $\hat{T}$ ($\hat{\pi}$) is an assumed transition (policy) given a type. We denote an MDP corresponding to each type θ as M_θ.

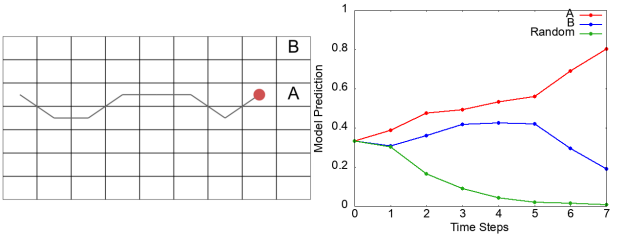
For example, Figure 1 shows a Maze World from Baker et al. (2009) where the agent can take 9 different actions: *Stay, North, South, East, West, NorthEast, NorthWest, SouthEast*



(a) Legibility



(b) Explicability



(c) Legibility + Explicability

Figure 1: Three traces of different observer-aware behaviors.

and *SouthWest*. However, when the agent takes an action it can veer left or right with probability 0.15, respectively. The agent's goal is to reach either one of the possible goals $\{A, B\}$. The rewards for the actions are proportional to the negative distance traveled by the action. The domain can be described as an MDP with possible types $\Theta = \{A, B\}$. All MDPs (M_A and M_B) share the same transition and reward dynamics except at goal locations; that is, performing Stay when $\theta^* = A$ and the agent is at A leads to the absorbing terminal state for M_A but not for M_B.

Note that the equation above assumes that attempted actions are observable to the observer. If that is not the case, we can marginalize over actions as follows.

$$Pr(\theta|h_{t+1}) = Pr(\theta|s_t, s_{t+1}, b_t)$$
$$= \frac{\sum_{a'_t \in A} \hat{T}(s_t, a'_t, s_{t+1}|\theta)\hat{\pi}(s_t, a'_t|\theta)b_t(\theta)}{\sum_{\theta'}\sum_{a'_t \in A} \hat{T}(s_t, a'_t, s_{t+1}|\theta')\hat{\pi}(s_t, a'_t|\theta')b_t(\theta')}.$$

While we can use any $\hat{\pi}$ in principle for the belief update, BST belief update makes the following assumption:

$$\hat{\pi}(s_t, a_t|\theta) \propto exp(\beta Q^*(s_t, a_t|\theta)) \quad (2)$$

That is, at each time step, an agent takes an action with a probability exponentially proportionate to how good the action is at the current state (based on the optimal Q-value $Q^*(s_t, a_t|\theta)$). The hyper-parameter β presents the agent's level of rationality.

For example, in Figure 1a, when $\beta = 1$, $\hat{\pi}(s_o, South|A) \approx 0.08$ while $\hat{\pi}(s_0, South|B) \approx 0.05$. This means that by going south, the observed agent can make the posterior belief on A higher. Figure 1a shows that the posterior belief on A is slightly higher than the other goal after taking the first action, and how it evolves over time.

3.2 BELIEF-DEPENDENT REWARDS IN OAMDP

Now, we show how OAMDP can produce various observer-aware behaviors proposed in the literature by changing R.

Legibility Legible behaviors convey intentions via the choice of actions. Legible behaviors are often modeled as maximizing the beliefs on the true intention/goal (θ^*) of the agent. To accomplish that, R could be the negative Euclidean distance between the current belief and the target belief ($b(\theta^*)$=1 for the true type $\theta^* \in \Theta$), or the negative Kullback-Leibler divergence. Dragan and Srinivasa (2013) assigned higher legibility for trajectories that make the posterior on the true goal higher. MacNally et al. (2018) used as metric the cost before the belief on the true goal reaches a certain threshold. Our earlier ourk (Miura et al., 2021) extended legible planning to stochastic settings.

For example, in Figure 1a, the agent makes a detour to the south to clarify to the observer that it is not going to B.

Figure 2 shows another example of an agent legibly stacking blocks in Stochastic Blocks World—a stochastic domain in which picking up a block always succeeds with probability 1, while putting down a block fails with probability 0.1 (the block falls on the table). Each action has a negative reward of -1. In Figure 2, starting from the initial state, there are two possible goals, spelling "ARMS" or "RAMS". Suppose that the agent's true goal is to spell "ARMS". The optimal policy in terms of underlying rewards is to first unstack the block "S", but this is also part of an optimal policy to spell "RAMS". The maximally legible policy first stacks the block "R" on top of "A". This makes it more costly to spell "RAMS" than"ARMS", but it makes the intention clearer.

Explicability Explicable behaviors conform to observers' expectations. Works on explicability initially used the distance between plans as metric (Zhang et al., 2017), which does not translate very well to OAMDP. Sreedharan et al. (2020) later proposed a Bayesian account of explicability, where explicability is proportional to $\sum_{\theta \neq \theta_0} Pr(\theta|h_t)$. θ_0 represents a random agent. Intuitively, explicable agents want to avoid being interpreted as completely random. To accomplish that, we can define R so that the agent gets punished for having high belief on θ_0 ($-b(\theta_0)$).

For example, in Figure 1b, the agent makes progress toward

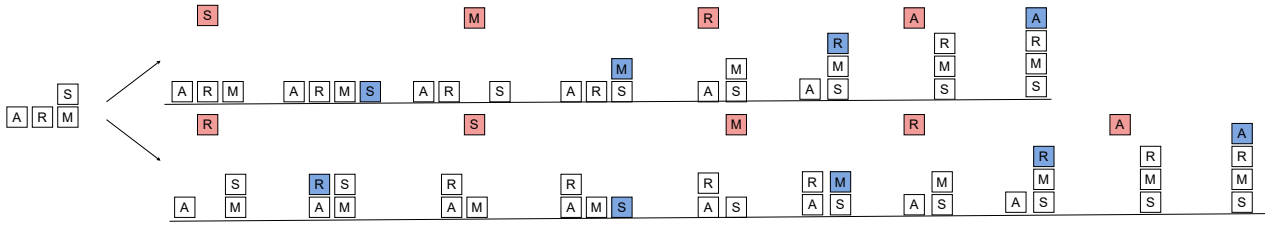


Figure 2: Dissimulation behavior (top) and legible behavior (bottom) in Stochastic Blocks World. Red blocks are the ones the agent is holding. Blue blocks represent blocks that were just put down. The possible goals are "ARMS" or "RAMS".

its goal, showing to the observer that it is not random.

Deceptive Deceptive agents adversarially manipulate the beliefs of observers. Two specific kinds of deceptive behaviors are *simulation* and *dissimulation* (Bell, 2003; Masters and Sardina, 2017). A simulation agent is interested in making its observer believe in a false goal. Much like what we proposed for legibility, we can reward agents for making the belief on the deceptive goal/intention higher. A dissimulation agent is interested in making its intention obscure. To that end, we can, for example, reward agents for inducing beliefs with high entropy. Note, however, that the agent can simply do nothing to stay obscure. To ensure progress with the agent's task, the rewards need to incorporate the underlying rewards of the task.

Care must be taken, however, when defining B for deceptive agents. Most of the previous works (Baker et al., 2009; Ramírez and Geffner, 2010) on mapping histories to belief operate under the *keyhole* setting, where the acting agent is not aware that it is being observed. However, if the observer is aware of the acting agent trying to manipulate its belief, the keyhole assumption is no longer valid. For example, imagine a scenario with two possible goals X and Y. The acting agent can go to Y to manipulate the belief, but if the observer is aware of the manipulation, it might infer that the true goal is X. Note, however, that OAMDP is not tied to any particular B. If B incorporates recursive reasoning about the agents, OAMDP can plan according to that model.

Works on goal recognition design (Keren et al., 2019) investigate how to design environments so that agents cannot conceal their goals by minimizing the *worst-case distinctiveness* measure, which is relevant to deceptive OAMDPs.

Predictability Predictable agents take actions so that their future actions (rather than goals) are easier to predict. Dragan et al. (2013) proposed to model the predictability of a trajectory as simply proportional to the value (negative cost) of a trajectory. OAMDP can maximize predictability according to this definition by simply maximizing underlying rewards. Fisac et al. (2020) proposed to model t-predictable agents (in deterministic settings) as maximizing $Pr(a_{t+1}, \cdots, a_T | a_1, \cdots, a_t)$. OAMDP can, in principle, model t-predicability by having a type for each possible trajectory. However, this would require an exponential number of types. Fisac et al. (2020) proposed an approximate algorithm to maximize predictability by only considering l-least cost plans.

(In)Capability Agents can also express their (in)capabilities through behaviors (e.g. Kwon et al., 2018). While not a direct generalization of their work, we can have OAMDPs with two possible types $\theta_{capable}$ and $\theta_{incapable}$, and maximize the beliefs on them accordingly.

For example, consider a variation of Stochastic Blocks World, where the block "R" has a different shape than the other blocks. The observer is not sure whether the agent can handle that block or not. Assume, for example, that picking up "R" has a negative reward of -3 given $\theta_{incapable}$ (instead of -1). Then an OAMDP agent would first pick up "R" and put it down, showing to the observer that it can handle "R".

Combining Different Observer-Aware Behaviors The definition of OAMDP is general enough to combine different notions of interpretability, although different notions of interpretability can be at odds with each other (Dragan and Srinivasa, 2013; Sreedharan et al., 2020). For example, in Figure 1a, moving away from goals made the posterior that it is a random agent higher. If we introduce θ_0 and combine legibility and explicability, the agent makes a smaller detour (Figure 1c). Similarly, if we do not consider explicability of behaviors, the agent can keep picking up "R" to show its capability. By combining capability and explicability, the agent can balance these competing objectives.

Explicit Communication While observer-aware agents in this paper primarily focus on manipulating the observer's belief through implicit behaviors, OAMDP can also model explicit communication (e.g. Renoux et al., 2020). For example, in the car example earlier, the driver can use hand gestures to communicate their intention (although that may be more costly relative to implicit communication). In this case, after making the hand gesture, B wwould increase the belief on θ^*.

OAMDP does not generalize frameworks which allow for partial observability (e.g. Kulkarni et al., 2019). Introducing partial observability amounts to relaxing Assumption 3

in Section 5. OAMDP also cannot express joint transition functions as in (Sadigh et al., 2016; Lo et al., 2020), which would violate Assumption 1 in Section 5.

4 RELATIONSHIP TO POMDP

Despite substantial similarities between OAMDPs and POMDPs Kaelbling et al. (1998), they do not subsume each other. Both formulations operate on agents' beliefs. However, while POMDP keeps track of the acting agent's belief, the belief in OAMDP is the assumed belief of the observer. Moreover, while rewards in POMDP are defined in terms of underlying states, rewards in OAMDP depend on beliefs of observers. Although there exist extensions of POMDP that allow belief-dependent rewards (Mauricio et al., 2010; Spaan et al., 2015), the beliefs are still over underlying states unlike in OAMDPs.

Most importantly, while POMDPs only allow Bayesian belief updates through observations, OAMDPs assume full observability and do not use observations to update beliefs. To produce observer-aware behaviors, one might try formulating the problem as POMDP, where the states are pairs of S and Θ and observations are pairs of S and A. Then letting $O(\langle s_{t+1}, a_t\rangle|\langle s_{t+1}, \theta\rangle, a_t) = \hat{Pr}(s_{t+1}, a_t|\theta, s_t)$ may seem similar to the belief update in Equation 1. However, as observing different states other than the current one should be impossible ($O(\langle s'_{t+1}, a'_t\rangle|\langle s_{t+1}, \theta\rangle, a_t) = 0$ for $\langle s'_{t+1}, a'_t\rangle \neq \langle s_{t+1}, a_t\rangle$) , O is not a valid probability distribution (it does not sum to one). Although MacNally et al. (2018) call their similar formulation POMDP, it also does not perform belief updates through observations.

While OAMDPs allow more general belief updates, they do not generalize POMDPs because the transitions depend on the *actual* current state. We get the following rather obvious property directly from the definition of OAMDPs.

Proposition 1. *Transitions in OAMDPs are Markovian given the last observation.*

This is in contrast to belief MDPs induced by POMDPs. Transitions between beliefs in a POMDP is given by $Pr(b'|a, b) = \sum_{o\in\Omega} Pr(b'|a, b, o)Pr(o|a, b)$. $Pr(o|a, b)$ depends on the previous belief b, which is a summary of the complete history up to the current time. Thus, POMDPs do not generally satisfy the property.

Yet another difference is that OAMDPs do not necessarily have α-vector representation, which is the basis of many existing POMDP planning algorithms such as Value Iteration (Sondik, 1978) and PBVI (Pineau et al., 2003). In α-vector representations, the value function can be expressed as a set of $|S|$-dimentional vectors ($\Gamma = \{\alpha_1, \cdots \alpha_n\}$). The value of a belief $b \in \Delta^{|S|}$ then is $V(b) = \max_{\alpha\in\Gamma} b\cdot\alpha$. α-vector representation depends on immediate rewards in POMDPs being a linear functions of beliefs ($R(b, a) = \sum_{s\in S} R(s, a)$). This is not necessarily the case in OAMDPs, where R can be, for example, the negative Euclidean distance.

OAMDP can be seen as a special case of Decision Process with non-Markovian Reward (NMRDP) (Bacchus et al., 1996; Thiébaux et al., 2006), where rewards are non-Markovian. Existing works on NMRDP (Bacchus et al., 1996; Thiébaux et al., 2006; Littman et al., 2017; Brafman et al., 2018), unlike OAMDPs, employ temporal logic to describe rewards over histories.

A number of existing works tackled the problem of inferring other agents' intentions from observations (Baker et al., 2009; Ramírez and Geffner, 2010), and how to react in response (e.g. Macindoe et al., 2012; Broz et al., 2013; Fern et al., 2014; Freedman and Zilberstein, 2017). In this paper, we focus on the orthogonal problem of choosing actions so as to make an agent's intentions more interpretable.

5 OAMDP AS A SUBCLASS OF I-POMDP

In this section, we show that an OAMDP can be derived as a special case of I-POMDP (Gmytrasiewicz and Prashant, 2005), which is an extension of POMDP to multi-agent settings. In an I-POMDP, agents maintain beliefs over models of other agents as well as physical states. As observer-aware planning handles (pseudo) multi-agent settings with observed and observing agents, it can be captured as an I-POMDP. In fact, Lo et al. (2020) use a formulation similar to I-POMDP.

Multi-agent formulations (e.g. Sadigh et al., 2016, 2018; Zhu et al., 2017; Lo et al., 2020) are arguably more general. For example, the multi-agent formulation allows agents to reason about what other agents may do in response. When being legible does not help other agents, the agent may choose not to be legible (Lo et al., 2020). However, multi-agent formulations are notoriously harder to solve (Seuken and Zilberstein, 2008) and require the full joint transition function. We argue that when the possible interactions among agents are limited, agents do not necessarily need a full multi-agent formulation. In this section, we spell out the set of assumptions needed to reduce an I-PODMP to the simpler OAMDP.

I-POMDP An Interactive POMDP (I-POMDP) for the observed agent (i) and observing agent (j) is described as a tuple $\langle IS^i, A^{ij}, \Omega^i, T^{ij}, O^i, R^i\rangle$. $IS^i = S \times M^j$ is the set of interactive states. M^j is the set of possible models of the other agent. The set of possible models are often subdivided into *subintentional* and *intentional* models. Subintentional models are relatively simple models such as fictitious play or finite-state controllers. In intentional models, the observing agent is an I-POMDP agent itself, and the agents recursively model each other to a finite depth. We use notations for subintentional models in this section. Each model $m^j \in M^j$ is a tuple $m^j = \langle O^j, z^j, f^j\rangle$, where O^j is an observation function of j, $z^j \in Z^j$ is a history of j's actions and obser-

vations, and $f^j \in F^j : Z^j \to \Delta^{|A^j|}$ is an assumed behavior of j. We call $\langle O^j, f^j \rangle$ a frame of j. $A^{ij} = A^i \times A^j$ is the set of joint actions for the two agents. Ω^i is the set of observations for the agent i. $T^{ij} : S \times A^{ij} \times S \to [0,1]$ is the stochastic transition function. $O^i : A^{ij} \times S \times \Omega^i \to [0,1]$ is the stochastic observation function. $R^i : IS^i \times A^{ij} \to \mathbb{R}$ is the reward function. The belief update is given by:

$$b^i_{t+1}(is^i_{t+1}) = Pr(is^i_{t+1}|a^i_t, \omega^i_t, b^i_t) = \eta \sum_{is^i_t} b^i_t(is^i_t)$$
$$\sum_{a^j_t \in A^j} f^j(z^j, a^j) O^i(a^i_t, a^j_t, s_{t+1}, \omega^i_{t+1}) Pr(is^i_{t+1}|is^i_t, a^i_t, a^j_t)$$

where is^i_t ranges over interactive states sharing the frame with is^i_{t+1}, η is a normalizing constant, and $Pr(is^i_{t+1}|is^i_t, a^i_t, a^j_t)$ is the transition probability between interactive states:

$$Pr(is^i_{t+1} = \langle s_{t+1}, m^j_{t+1} \rangle | is^i_t = \langle s_t, \langle O^j_t, z^j_t, f^j_t \rangle \rangle, a^i_t, a^j_t)$$
$$= T^{ij}(s_t, a^i_t, a^j_t, s_{t+1}) \sum_{\omega^j_t \in \Omega^j} O^j_t(a^i_t, a^j_t, s^j_{t+1}, \omega^j_t) \delta(z^j_{t+1}, z^j_t a^j_t \omega^j_t)$$

where δ is Kronecker delta and $z^j_t a^j_t \omega^j_t$ represents the result of concatenating an action and observation to a history.

OAMDP Assumptions OAMDPs form a special case of I-POMDPs under the following five assumptions:

1. The observing agent is passive. Formally, we can represent this assumption by having only one action for the observing agent $A^j = \{noop\}$.
2. The frame of the observing agent is known to the observed agent. We refer to this frame as $\langle O^j_*, f^j_* \rangle$.
3. The observing agent can fully observe the underlying states (S) and actions performed by the observed agent (A^i), i.e. $\Omega^j = S \times A^i$ and $O^j_*(a^i_t, a^j_t, s_{t+1}, \omega^j_t) = 1$ if $\omega^j_t = \langle s_{t+1}, a^i_t \rangle$, and 0 otherwise.
4. The observed agent can fully observe the underlying states (S), i.e. $\Omega^i = S$ and $O^i(a^i_t, a^j_t, s_{t+1}, \omega^j_t) = 1$ if $\omega^j_t = s_{t+1}$, and 0 otherwise.
5. Rewards are described in terms of B and R.

Proposition 2. *Under Assumptions* 1-5*, I-POMDPs and OAMDPs are equivalent.*

Proof Sketch. With these assumptions, the transition function between interactive states simplifies to:

$$Pr(is^i_{t+1} = \langle s_{t+1}, m^j_{t+1} \rangle | is^i_t = \langle s_t, \langle O^j_t, z^j_t, f^j_t \rangle \rangle, a^i_t, a^j_t)$$
$$= T^{ij}(s_t, a^i_t, a^j_t, s_{t+1}) \sum_{\omega^j_t \in \Omega^j_t} O^j_t(a^i_t, a^j_t, s_{t+1}, \omega^j_t) \delta(z^j_{t+1}, z^j_t a^j_t \omega^j_t)$$
$$= T^{ij}(s_t, a^i_t, a^j_t, s_{t+1}) \delta(z^j_{t+1}, z^j_t a^j_t a^i_t s_{t+1}) \quad \text{by Assumption 3}$$
$$= T^{ij}(s_t, a^i_t, s_{t+1}) \delta(z^j_{t+1}, z^j_t a^i_t s_{t+1}) \quad \text{by Assumption 1}$$

This is exactly the same as the transition function of an OAMDP when $T^{ij} = T$.

Now we argue that under Assumptions 1-5 above, there is exactly one interactive state is^i_t such that $\langle O^j_t, f^j_t \rangle = \langle O^j_*, f^j_* \rangle$ and $b^i_t(is^i_t) = 1$ for every time step t. We can show this by induction on the number of time steps. The claim is trivially true for the first time step because there is no uncertainty about the frame of the observing agent according to Assumption 2. And using Assumption 4, we can easily show that there is exactly one is^i_{t+1} with $b(is^i_{t+1}) = 1$. Assuming $b^i_t(is^i_t) = 1$ for exactly one is^i_t, we get:

$$b^i_{t+1}(is^i_{t+1}) = Pr(is^i_{t+1}|a^i_t, \omega^i_t, b^i_t)$$
$$= \beta \sum b^i_t(is^i_t) \sum_{a^j_t \in A^j} f^j_t(z^j_t, a^j_t) \quad O^i(a^i_t, a^j_t, s_{t+1}, \omega^i_t)$$
$$Pr(is^i_{t+1}|is^i_t, a^i_t, a^j_t)$$
$$= \beta \sum b^i_t(is^i_t) \sum_{a^j_t \in A^j} f^j_t(z^j_t, a^j_t) \delta(\omega^i_t, s_{t+1})$$
$$T^i(s_t, a^i_t, s_{t+1}) \delta(z^j_{t+1}, z^j_t a^i_t s_{t+1}) \text{ by Assumption 4}$$
$$= \beta \sum b^i_t(is^i_t) T^i(s_t, a^i_t, s_{t+1}) \delta(\omega^i_t, s_{t+1}) \delta(z^j_t, z^j_t a^i_t s_{t+1})$$
$$= \delta(O^j_{t+1}, O^j_*) \delta(f^j_{t+1}, f^j_*) \delta(\omega^i_t, s_{t+1}) \delta(z^j_t, z^j_t a^i_t s_{t+1})$$

which implies that there is exactly one is^i_{t+1} with $b(is^i_{t+1}) = 1$. This justifies the fact that OAMDPs do not keep a belief over a model of the observing agent. □

Note that the AV examples seemingly violate Assumption 1 that the observing agent is passive. But, while the pedestrians interact with the observed AV, they are assumed to be agnostic to what the AV does. This assumption is not strictly true, but given the leader-follower nature of the problem (the car needs to adapt to the pedestrian, not the other way around), we argue that the assumption is reasonable for the purpose of planning. The detailed description of the domain is provided in Section 7.2.

6 THE COMPLEXITY OF OAMDP$_{BU}$

We next show several complexity results for OAMDP$_{BU}$, a special case of OAMDP where the observer is Bayesian. As complexity classes are defined in terms of decision problems, we consider the (finite-horizon) value problem: Given an OAMDP$_{BU}$, a planning horizon H, and a threshold K, does the OAMDP$_{BU}$ have a (finite-horizon history-dependent) policy with value equal to or greater than K?

Theorem 1. *The finite-horizon value problem for OAMDP$_{BU}$ is PSPACE as long as R can be evaluated using polynomial space.*

Proof. The proof that OAMDP is PSPACE is almost identical to the proof that POMDP is PSPACE (Papadimitriou and Tsitsiklis, 1987). Given a policy, the possible outcomes within the finite-horizon can be expressed as a tree of that depth. As long as R can be evaluated using polynomial

space, checking if the policy achieves an expected return greater than a threshold can also be done using polynomial space via tree traversal[1]. As NPSPACE=PSPACE, OAMDP$_{BU}$ is PSPACE. □

The result implies that OAMDP$_{BU}$ is less complex than a finitely-nested (intentional) I-POMDP, whose complexity is considered to be doubly exponential in the input size (Seuken and Zilberstein, 2008).

The result also implies that there is a polynomial time reduction from OAMDP$_{BU}$ to POMDP (in the sense that we can reduce OAMDP$_{BU}$ to QSAT and then to POMDP). However, as we discussed above (Section 4), there seems to be no obvious direct translation between the two frameworks. Thus, reducing OAMDP$_{BU}$ to POMDP and solving the resulting POMDP would not be fruitful.

Theorem 2. *The finite-horizon value problem for OAMDP$_{BU}$ is PSPACE-hard.*

Proof. By reduction of QSAT to an OAMDP. Given a QBF $Q_1x_1Q_2x_2\cdots Q_nx_n\phi(x_1,x_2,\cdots,x_n)$, where Q_i is i-th quantifier ($\forall$ or $\exists$), we reduce it to a value problem with OAMDP$_{BU}$ (M^ϕ) and $K=1$. The key idea is to use $\Delta^{|\Theta|}$ to record the history:

- S consists of t_i, f_i, the initial state ι , the terminal state s_∞, and s'_∞. Intuitively, $S_i = t_i/f_i$ means x_i is true/false.
- There are two actions, corresponding to assigning true and false to variables. When $Q_{i+1}=\forall$, the actions lead to t_{i+1} and f_{i+1} with equal probability. When $Q_{i+1}=\exists$, the actions make deterministic transitions leading to t_{i+1} and f_{i+1}, correspondingly. Taking any action from t_n and f_n leads to s_∞.
- Θ consists of θ_i for each x_i, θ_{dummy}, and θ_∞. Intuitively, $b(\theta_i)$ represents the truth values for x_i. θ_{dummy} is only needed to make sure that $\Delta^{|\Theta|}$ sums to 1. θ_∞ represents whether all the variables have been assigned. We have the uniform prior $b_0(\theta)=\frac{1}{|\Theta|}$ for all $\theta\in\Theta$.
- As for BU, $\hat{Pr}(t_i,a_t|\theta_i,s_t) > 0$ and $\hat{Pr}(f_i,a_t|\theta_i,s_t)=0$ for all $a_t\in A$ and $s_t\in S$. This means that after transitioning to f_i, $b(\theta_i)=0$.
- $\hat{Pr}(s_\infty,a_t|\theta_\infty,s_t)=0$ and $\hat{Pr}(s'_\infty,a_t|\theta_\infty,s_t)=1$. This means that after transitioning to s_∞, $b(\theta_\infty)=0$.
- $R(b)=0$ when $b(\theta_\infty)>0$. When $b(\theta_\infty)=0$, R considers the corresponding assignments for x_i ($x_i=[b(\theta_i)>0]$). $R(b)=1$ if ϕ is true under this assignment. Otherwise ,$R(b)=0$.

[1]As in (Papadimitriou and Tsitsiklis, 1987), we assume that the planning horizon is polynomial in the input size.

We claim that there exists a policy with value 1 iff the QBF is true. Suppose such a policy (π) exists. Then for all h_{n+1} possible under π, we have $R(h_{n+1})=1$, which means ϕ is true under the assignments defined by h_{n+1}. By construction of M^ϕ, this is exactly the set of assignments possible when all the existential quantifies are assigned the same way as π. Therefore, the QBF is true.

Conversely, suppose the QBF is true. We can construct a policy that takes corresponding actions at existentially quantified variables as the assignment that makes the QBF true. For all h_{n+1} possible, we have $R(h_{n+1})=1$, which means the policy has the value 1. □

Corollary 1. *The finite-horizon value problem for OAMDP$_{BU}$ is PSPACE-complete when R can be evaluated using polynomial space.*

We next show that even if we restrict our attention to stationary policies, the worst-case complexity of OAMDP$_{BU}$ remains intractable.

Theorem 3. *The value problem of stationary policy for deterministic OAMDP$_{BU}$ is NP-hard.*

Proof. The proof is similar to the NP-hardness proof for finding a stationary policy for POMDP (Littman, 1994; Lusena et al., 2001). To show NP-hardness, we reduce 3SAT to OAMDP$_{BU}$. Given a 3CNF formula $\phi(x_1,\cdots,x_n)=C_1\wedge\cdots\wedge C_m$, we have an OAMDP$_{BU}$ M'_ϕ such that M'_ϕ has three states for each appearance of a variable (x_{ij}) in the formula. x_{ij} represents an appearance of x_i in a clause C_j (we assume without loss of generality that a variable appears only once in each clause). For each x_{ij}, we have a decision state (d_{ij}), where there are two actions corresponding to assigning true/false to the variable. The actions lead to t_{ij}/f_{ij} deterministically. When the assignment makes the clause true, the process transitions to the first decision state of the next clause (for the last variable, it transitions to the terminal state s_∞). Otherwise, the process transitions to the next variable in the same clause (if it is the last variable, the process transitions to a sink state s_{bad}).

Intuitively, transitioning to s_∞ means that ϕ is satisfiable. However, a policy can assign different actions (truth values) to different appearances of the same variable, which leads to a contradicting assignment. To work around the issue, M'_ϕ has a reward that ensures that different appearances of the same variable ($x_{ij}=x_{ik}$ where $j\neq k$). Same as M_ϕ in the previous proof, M'_ϕ uses $\Delta^{|\Theta|}$ to record the assignments to each (appearance of) variable. Then we can make $R(b)=[\bigwedge_{j<k}x_{ij}=x_{ik}\wedge s_\infty]$, where b defines the assignments to variables. Note that the size of the formula describing the reward is linear in the size of the 3CNF. It is easy to see that M'_ϕ has a value of 1 if and only if the 3CNF is satisfiable. □

7 EXPERIMENTS

In this section, we describe our initial approach for solving OAMDPs and evaluate it on the examples presented earlier in the paper. The purpose of the experiments is to assess the feasibility of (approximately) solving OAMDPs in practice. Further work is needed to refine the solution methods and evaluate them more rigorously.

7.1 SOLUTION METHODS

Most solution methods for MDPs (e.g., Value Iteration) cannot be applied to OAMDPs. The reason is that these methods rely on the Markov property of rewards, which OAMDPs do not satisfy. Similarly, most solution methods for POMDPs cannot be applied to OAMDPs because these methods rely on α-vector representation of the value function, which OAMDPs do not necessarily have (Section 4). Solution methods for I-POMDPs (e.g. Doshi and Perez, 2008; Doshi and Gmytrasiewicz, 2009) are applicable, but seem overly complex as it is unnecessary to consider the uncertainty over interactive states (Section 5) to solve OAMDPs.

Given these considerations, we use methods that work for any general (acyclic) AND/OR graph: AO^* (Nilson, 1980) and UCT (Kocsis and Szepesvári, 2006). From the root node, representing the initial state, AO^* gradually builds a solution graph by expanding tip nodes of the current best partial solution graph. It utilizes heuristic values to estimate how good newly expanded nodes are, and propagates the new information back to the root. When an admissible heuristic (lower-bound estimate) is used, AO^* is guaranteed to return an optimal solution. However, finding an admissible heuristic function for different kinds of observe-aware behaviors is hard. Hence, we did not use any heuristics in our experiments.

UCT uses a sequence of stochastic simulations from the root node. The algorithm chooses actions according to the UCB1 (Auer et al., 2002) formula:

$$Q(a,s,d) + C\sqrt{2\log N(s,d)/N(s,a,d)}$$

where $Q(a,s,d)$ is the estimated Q-value, $N(s,d)$ and $N(s,a,d)$ are counters of the number of times the simulations encountered the node (s,d) and (s,a,d), and C is a constant that controls the degree of exploration, respectively. In our experiments, C is set to the current $Q(a,s,d)$, which is common in the literature (e.g. Bonet and Geffner, 2012). When a new node is added to the explicit graph, an accumulated discounted reward is sampled by simulating a base policy π (averaged over 10 episodes). We used an optimal policy for the underlying MDP as the base policy. Note that although UCT eventually explicates the whole graph and finds the optimal policy, policies returned after a fixed number of simulations are not guaranteed to be optimal.

7.2 DOMAINS

We solved instances of the problems presented earlier: Maze World (MW), Blocks World (BW), and Autonomous Vehicle (AV). The detailed descriptions of Maze World and Blocks World are in Section 3. We used BST belief update and set $\beta = 1$ in Equation 2 as in Baker et al. (2009) and Ramírez and Geffner (2010), except in AV, where we used $\beta = 0.25$ as the difference in values is much higher in that domain. We assumed that actions are observable, except in Maze World problems. As R, we used the negative Euclidean distance from the target belief.

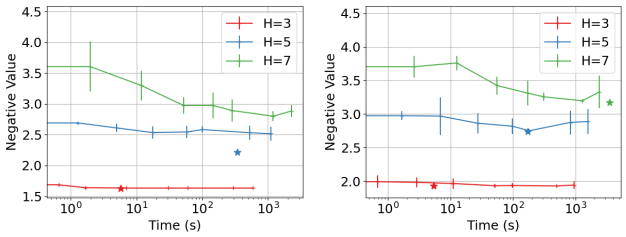
AV at Crosswalk (Aware/NotAware): We now describe the simple scenario introduced earlier of an autonomous vehicle (AV) approaching a crosswalk. We want to compute a policy for the AV that clearly conveys its intention.

States are described by the current configuration of the car and the position of the pedestrian. The current configuration of the vehicle is represented by how far ahead the vehicle is along the intended trajectory $0 \leq \phi \leq 30$ and its velocity $0 \leq \dot{\phi} \leq 4$. Available actions are $A = \{-2, -1, 0, +1, +2\}$, which change the velocity of the vehicle by the corresponding value. The current position of the pedestrian is represented by $0 \leq \xi \leq 10$. At each time step the pedestrian increases its position by 1 with probability 0.9 and stays at the current position with probability 0.1. Our model assumes that the vehicle wants to keep its speed in the range $2 \leq \dot{\phi} \leq 3$, for which there is a negative reward of -1. Otherwise, the reward is -5.

The two possible types, $\Theta = \{\textit{Aware}, \textit{NotAware}\}$, indicate whether the vehicle is aware of the pedestrian or not. The vehicle and the pedestrian would collide with each other at the crosswalk when $4 \leq \xi \leq 8$ and $17 \leq \phi \leq 21$. In this case, the agent will receive a significant penalty of -1000. When $\theta^* = \textit{NotAware}$, the agent is oblivious to this penalty. The policy maximizing legibility for $\theta^* = \textit{Aware}$ would start hitting the brake immediately until the vehicle completely stops.

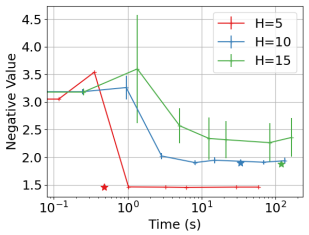
AV at Crosswalk (Close/Far) is a variant of the previous AV scenario in which the crosswalk and the pedestrian are sufficiently far apart that the vehicle does not need to decelerate. The states and available actions remain the same.

There two parameter values, $\Theta = \{\textit{Close}, \textit{Far}\}$, indicate where the crosswalk is located. When $\theta^* = \textit{Close}$, the vehicle and the pedestrian would collide with each other at the crosswalk when $4 \leq \xi \leq 8$ and $17 \leq \phi \leq 21$. When $\theta^* = \textit{Far}$, the crosswalk is farther away from the pedestrian and the vehicle and the pedestrian would collide with each other at the crosswalk when $7 \leq \xi \leq 10$ and $17 \leq \phi \leq 21$. The policy maximizing legibility for $\theta^* = \textit{Far}$ would start accelerating before the crosswalk. This is because when $\theta^* = \textit{Close}$, accelerating is not very efficient. The vehicle will probably have to stop for the pedestrian later.

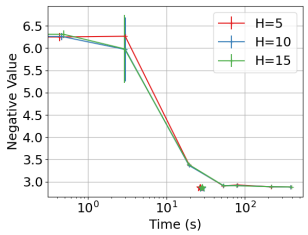


(a) MW (L)

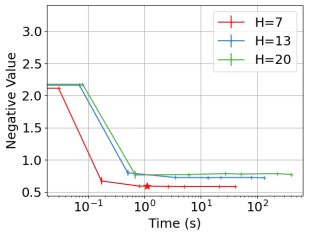
(b) MW (EL)



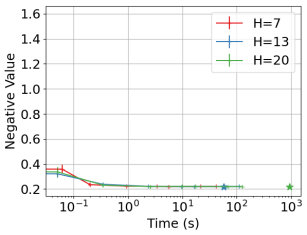
(c) BW (L)



(d) BW (CE)



(e) AV (Aware/NotAware)



(f) AV (Close/Far)

Figure 3: Planning time and (negative) values for each task: Maze World (MW), Blocks World (BW), and Autonomous Vehicle (AV). Problems with different notions of observability and their combinations were considered: Legibility (L), Explicability (E), and Capability (C). The starred dots represent the planning time and value for AO*. When a starred dot is missing, the problem was not solved within the limit (60 mins and 2 GB memory). H is the planning horizon.

7.3 RESULTS

We experimented with both AO^* and UCT, applying them to the examples introduced earlier in the paper. For UCT, the values of the policies were obtained by averaging over 100 episodes. When UCT has not seen a state, the base policy was used to select an action. Figure 3 shows the results. The points for UCT correspond to 0, 10, 100, 1k, 5k, 10k, 50k, 100k iterations (rollouts), averaged over 10 runs. UCT with 0 iterations corresponds to the optimal policy for the underlying MDP.

As suggested by Theorem 2, the exact solution method for OAMDP (AO^*) failed to produce policies within a given time limit for some problems (MW (L) and AV (Aware/NotAware)). AO^*, however, performed better in BW, where significant portion of the search space are pruned. UCT was able to improve interpretability compared to the base policy for all problems except MW(E), where the base policy turned out to be optimal for explicability as well. Increasing H made the problems harder to solve in general. For some problems (BW (CE) and AV (Close/Far)), however, the agent was able to change the observer's belief sufficently during the first several time steps that increacing H did no make the problems harder. Note that Equation 2 in the BST belief update requires the optimal Q-values for every state, action, and type. The time needed to compute the optimal Q-values is not included in the reported runtimes.

8 CONCLUSION

We propose OAMDP—a general framework for observer-aware planning—and illustrate through examples how OAMDP can model a wide range of observer-aware planning problems from the literature (Section 3). While previous works have identified different kinds of observer-aware behaviors and proposed different techniques to optimize them, there is much to be gained by exposing the connections between these works through the lens of our unifying framework. Furthermore, as a general framework, OAMDP can naturally combine different notions of interpretability—an important objective that has been previously recognized (Dragan and Srinivasa, 2013; Sreedharan et al., 2020).

To properly place the OAMDP model in the context of previous work, we show that OAMDP can be derived from I-POMDP using five assumptions (Section 5). Despite the close connection, the link between observer-aware planning and I-POMDP has not been previously investigated. We argue that OAMDP is preferable to I-POMDP for formulating observer-aware planning problems in two ways. First, OAMDP$_{BU}$ is less complex than I-POMDP (Theorem 1). While OAMDP$_{BU}$ itself is PSPACE-hard, it is preferable to solving an exponential number of POMDPs as in I-POMDP. Second, OAMDP does not require the full multi-agent model of the environment (observation functions for both agents and the joint transition function).

We analyze the complexity of solving OAMDP$_{BU}$ optimally and prove that several variants of the problem are intractable (Theorems 2-3). With the exception of Kulkarni et al. (2019) who consider a slightly different setting (with partial observability), no previous work discussed the complexity of observer-aware planning. While solving OAMDP$_{BU}$ exactly is intractable, for the problems we considered, UCT was able to return good approximate policies (Section 7).

This work opens up several interesting directions for future research, including empirical evaluations of various tasks with human subjects and approximation algorithms that further exploit the structure of OAMDPs.

Acknowledgements

This research was supported by the National Science Foundation grant number IIS-1724101.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Fahiem Bacchus, Craig Boutilier, and Adam Grove. Rewarding behaviors. In *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence*, pages 1160–1167, 1996.

Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113: 329–349, 2009.

J. Bowyer Bell. Toward a Theory of Deception. *International Journal of Intelligence and CounterIntelligence*, 16(2):244–279, 2003.

Blai Bonet and Hector Geffner. Action selection for MDPs: anytime AO* versus UCT. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1749–1755, 2012.

Ronen Brafman, Giuseppe De Giacomo, and Fabio Patrizi. LTLf/LDLf non-Markovian rewards. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1771–1778, 2018.

Frank Broz, Illah Nourbakhsh, and Reid Simmons. Planning for human–robot interaction in socially situated tasks: The impact of representing time and intention. *International Journal of Social Robotics*, 5(2):193–214, 2013.

Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. Balancing explicability and explanations in human-aware planning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1335–1343, 2019.

Prashant Doshi and Piotr J. Gmytrasiewicz. Monte Carlo sampling methods for approximating interactive POMDPs. *Journal of Artificial Intelligence Research*, 34 (1):297–337, 2009.

Prashant Doshi and Dennis Perez. Generalized point based value iteration for interactive POMDPs. In *Proceedings of the Twenty-Third National Conference on Artificial Intelligence*, pages 63–68, 2008.

Anca Dragan and Siddhartha Srinivasa. Generating legible motion. In *Proceedings of Robotics: Science and Systems*, 2013.

Anca Dragan, Rachel Holladay, and Siddhartha Srinivasa. Deceptive Robot Motion: Synthesis, Analysis and Experiments. *Auton. Robots*, 39(3):331–345, 2015.

Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *Proceedings of Eighth ACM/IEEE International Conference on Human-Robot Interaction*, pages 301–308, 2013.

A. Fern, S. Natarajan, K. Judah, and P. Tadepalli. A Decision-Theoretic Model of Assistance. *Journal of Artificial Intelligence Research*, 50:71–104, 2014.

Jaime F. Fisac, Chang Liu, Jessica B. Hamrick, Shankar Sastry, J. Karl Hedrick, Thomas L. Griffiths, and Anca D. Dragan. Generating plans that predict themselves. In *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*, pages 144–159, 2020.

Richard G. Freedman and Shlomo Zilberstein. Integration of planning with recognition for responsive interaction using classical planners. In *Proceedings of the Thirty-First Conference on Artificial Intelligence*, pages 4581–4588, 2017.

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

Leslie Pack Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Sarah Keren, Avigdor Gal, and Erez Karpas. Goal recognition design in deterministic environments. *Journal of Artificial Intelligence Research*, 65:209–269, 2019.

Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *European Conference on Machine Learning*, pages 282–293, 2006.

Anagha Kulkarni, Siddharth Srivastava, and Subbarao Kambhampati. A unified framework for planning in adversarial and cooperative environments. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 2479–2487, 2019.

Minae Kwon, Sandy H. Huang, and Anca D. Dragan. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95, 2018.

Michael L. Littman. Memoryless policies: Theoretical limitations and practical results. In *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pages 238–245, 1994.

Michael L. Littman, Ufuk Topcu, Jie Fu, Charles Isbell, Min Wen, and James MacGlashan. Environment-independent task specifications via GLTL. *arXiv:1704.04341*, 2017.

Shih-Yun Lo, Elaine Schaertl Short, and Andrea L. Thomaz. Planning with partner uncertainty modeling for efficient information revealing in teamwork. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 319–327, 2020.

Christopher Lusena, Judy Goldsmith, and Martin Mundhenk. Nonapproximability results for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 14:83–103, 2001.

Owen Macindoe, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. POMCoP: Belief space planning for sidekicks in cooperative games. In *Proceedings of the Eighth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 38–43, 2012.

Aleck M. MacNally, Nir Lipovetzky, Miquel Ramirez, and Adrian R. Pearce. Action selection for transparent planning. In *Proceedings of the Seventeenth International Conference on Autonomous Agents and MultiAgent Systems*, pages 1327–1335, 2018.

Peta Masters and Sebastian Sardina. Deceptive path-planning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4368–4375, 2017.

Araya-López Mauricio, Olivier Buffet, Vincent Thomas, and François Charpillet. A POMDP extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems*, pages 64–72. 2010.

Shuwa Miura, Andrew Cohen, and Shlomo Zilberstein. Maximizing legibility in stochastic environments. In *Proceedings of the Thirtieth IEEE International Conference on Robot and Human Interactive Communication*, 2021.

Nils Nilson. *Principles of Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., 1980.

Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

Joelle Pineau, Geoff Gordon, and Sebastian Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 1025–1030, 2003.

Miquel Ramírez and Hector Geffner. Probabilistic plan recognition using off-the-shelf classical planners. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1121–1126, 2010.

Jennifer Renoux, Tiago Veiga, Pedro Lima, and Matthijs Spaan. A unified decision-theoretic model for information gathering and communication planning. In *Proceedings of the Twenty-nineth IEEE International Conference on Robot and Human Interactive Communication*, pages 67–74, 2020.

Dorsa Sadigh, Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*, 2016.

Dorsa Sadigh, Nick Landolfi, Shankar Sastry, Sanjit Seshia, and Anca Dragan. Planning for cars that coordinate with people: Leveraging effects on human actions for planning and active information gathering over human internal state. *Autonomous Robots*, 42(7):1405–1426, 2018.

Sven Seuken and Shlomo Zilberstein. Formal models and algorithms for decentralized decision making under uncertainty. *Autonomous Agents and Multi-Agent Systems*, 17(2):190–250, 2008.

Edward J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, 1978.

Matthijs T. J. Spaan, Tiago S. Veiga, and Pedro U. Lima. Decision-theoretic planning under uncertainty with information rewards for active cooperative perception. *Autonomous Agents and Multi-Agent Systems*, 29(6):1157–1185, 2015.

Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, David E. Smith, and Subbarao Kambhampati. A Bayesian account of measures of interpretability in human-AI interaction. *arXiv:2011.10920*, 2020.

D. J. Strouse, Max Kleiman-Weiner, Josh Tenenbaum, Matt Botvinick, and David Schwab. Learning to share and hide intentions using information regularization. In *Advances in Neural Information Processing Systems*, pages 10270–10281, 2018.

Sylvie Thiébaux, Charles Gretton, John Slaney, David Price, and F. Kabanza. Decision-theoretic planning with non-Markovian rewards. *Journal of Artificial Intelligence Research*, 25:17–74, 2006.

Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *International Conference on Robotics and Automation*, pages 1313–1320, 2017.

Huaijiang Zhu, Volker Gabler, and Dirk Wollherr. Legible action selection in human-robot collaboration. In *Proceedings of the Twenty-Sixth IEEE International Symposium on Robot and Human Interactive Communication*, pages 354–359, 2017.