Observer-Aware Probabilistic Planning under Partial Observability

Extended Abstract

Salomé Lepers
Univ. de Lorraine, CNRS, Inria, LORIA
F-54 000 Nancy, France
firstname.lastname@loria.fr

Vincent Thomas
Univ. de Lorraine, CNRS, Inria, LORIA
F-54 000 Nancy, France
firstname.lastname@loria.fr

Olivier Buffet
Univ. de Lorraine, CNRS, Inria, LORIA
F-54 000 Nancy, France
firstname.lastname@loria.fr

ABSTRACT

We are interested in planning problems where the agent is aware of the presence of an observer, and where this observer is in a partial observability situation. The agent has to choose its strategy so as to optimize the information transmitted by observations. Building on observer-aware Markov decision processes (OAMDPs), we propose a framework to handle this type of problems and thus formalize properties such as legibility, explicability and predictability. This extension of OAMDPs to partial observability can not only handle more realistic problems, but also permits considering dynamic hidden variables of interest. We discuss theoretical properties of PO-OAMDPs and, experimenting with benchmark problems, we analyze HSVI's convergence behavior with dedicated initializations and study the resulting strategies.

KEYWORDS

Probabilistic Planning; Partial Observability; Legibility; Explicability; Predictability

ACM Reference Format:
Salomé Lepers, Vincent Thomas, and Olivier Buffet. 2025. Observer-Aware Probabilistic Planning under Partial Observability : Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

As explained by Klein et al. [9], efficient and safe human-agent collaboration requires behaviors that carry information such as intentions, abilities, current status or upcoming actions (see also [1, 3, 6–8, 14, 15]).

Here we consider an agent (robot or otherwise) observed by a passive human, as in Figure 1 (left). In this setting, Chakraborti et al. [4, 5] derive a taxonomy of the concepts behind such information communication through the behavior. In particular, they distinguish between (1) transmitting information, with properties such as *legibility* (legible behaviors convey intentions, *i.e.*, actual task at hand, via action choices), *explicability* (explicable behaviors conform to observers' expectations, *i.e.*, they appear to have some purpose), and *predictability* (a behavior is predictable if it is easy to guess the end of an on-going trajectory); or (2) hiding information, as through *obfuscation*, when the agent tries to hide its actual goal. They propose a general framework for such problems while assuming deterministic dynamics, and work mostly with plans (a sequence of actions, which induces a sequence of states). In their approach, the human is modeled by the robot as having a model of the robot+environment system (including the robot's possible tasks), and is thus able to predict the robot's behavior.

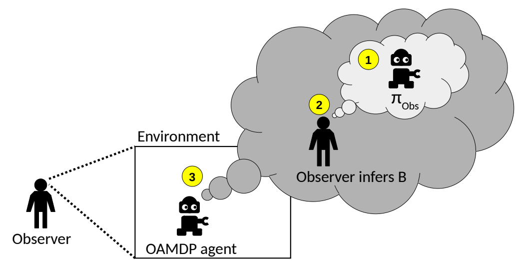


Figure 1: An OAMDP agent (3) assumes that the observer expects (2) the agent to behave so as to achieve some task (1).

Adopting a similar approach as Chakraborti et al. [5], Miura and Zilberstein [12] build a unifying framework while assuming stochastic transitions, namely *observer-aware Markov decision processes* (OAMDPs), illustrated in Figure 1. Among other things, their work also covers legibility, explicability, and predictability. The present paper proposes a formalism that can handle problems with partial observability. The observer has only access to a transition-dependent observation, while the agent has access to all information, *including* the observer's observation. The PO-OAMDP formalism is introduced in Sec. 2, before discussing theoretical properties of PO-OAMDPs and an example solving algorithm in Sec. 3.

2 OAMDPS WITH PARTIAL OBSERVABILITY

This section introduces the PO-OAMDP formalism, shows how the observer's belief about the target variable is maintained, and looks at some typical use cases.

Formalism. We describe the key ingredients of the PO-OAMDP framework before providing a formal definition. (1) Within the PO-OAMDP framework, a set of observations and an observation function are added to the OAMDP formalism. (2) A generic *target variable* is introduced whose value at each time step is a function of the transition followed by the system. (3) The agent has access not only to the complete state of the system, but also to the observations received by the observer (this is realistic in particular if the

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

observation process is deterministic). The agent can thus build the mental state of the observer during the execution of its behavior.

Formally, a PO-OAMDP is defined by a tuple $\langle \mathcal{S}, s_0, \mathcal{A}, T, \gamma, \mathcal{S}_f, \Psi, \Omega, O, B, R_{\text{AG}}, \phi \rangle$, where: $\langle \mathcal{S}, s_0, \mathcal{A}, T, \gamma, \mathcal{S}_f \rangle$ is an MDP with initial state s_0 but *no reward function*; Ψ denotes both the (dynamic) *target variable* and the finite set of values it can take; $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \Psi$ returns the value of the target variable given the transition: $\psi_t = \phi(s_t, a_t, s_{t+1})$; Ω is the finite set of observations; $O : \mathcal{A} \times \mathcal{S} \times \Omega \to \mathbb{R}$ is the observation function; $O(a, s', o)$ is the probability of emitting observation o if state s' is reached while performing a; $B : \Omega^* \to \Delta^{|\mathcal{S}|}$ gives the observer's belief on the state given an *observation* history; the belief on the target variable can be deduced from that *state belief*, denoted b; $R_{\text{AG}} : \mathcal{S} \times \Delta^{|\Psi|} \times \mathcal{A} \times \mathcal{S} \times \Delta^{|\Psi|} \to \mathbb{R}$ is the agent's reward function under its most general form: $R_{\text{AG}}(s_t, \beta_t, a_t, s_{t+1}, \beta_{t+1})$, where β denotes a *target belief.*

BST belief state update & target belief computation. Following Miura and Zilberstein, we employ the BST Bayesian belief update rule [2], thus introduce a reward function $R_{\text{OBS}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ assumed to be the agent's reward function according to the observer. Then, the observer models the agent's behavior for a given task through an MDP by: (1) solving the MDP with R_{OBS}; and (2) deriving a softmax policy π_{OBS}. Given the dynamics of the PO-OAMDP and presumed policy π_{OBS} of the agent, the observer faces a hidden Markov model (HMM) [13] and solves a *filtering* problem, using observation history $o_{1:t}$ to estimate her belief on the state s_t. One can then easily derive (1) the belief β on the value that will be taken by target variable Ψ_t, and (2) the expected reward.

The PO-OAMDP model allows us to generate different behaviors by changing Ψ and R, and to work with different types of problems. An important property, formally demonstrated by Lepers et al. [10], shows that PO-OAMDPs are at least as expressive as OAMDPs.

PROPOSITION 2.1. *Any OAMDP $\mathcal{M}$ with BST belief update can be turned into an equivalent PO-OAMDP $\mathcal{M}'$,* i.e., *such that an optimal solution to one problem is optimal for the other problem.*

The proof relies on turning the static *type* of an OAMDP into a (hidden) *target* state variable. For illustrative purposes, [10] show how to formulate legibility, explicability, and predictability.

3 RESOLUTION

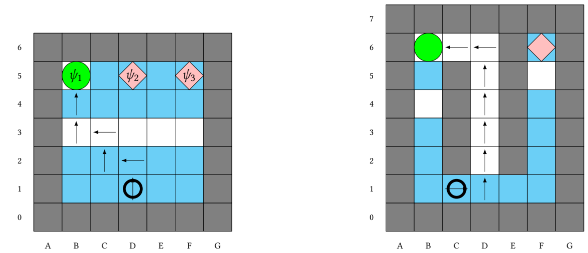
Sequential Decision-Making Problem. Similar to OAMDPs [11, 12], a PO-OAMDP can be turned into an equivalent MDP using the state-action-observation history $\langle s_{0:t}, a_{1:t}, o_{1:t} \rangle$ (*i.e.*, all the raw information available to the agent at t) as information state, or the state-belief (over target) pair $\langle s, b \rangle$ when using the BST update. Formally, we obtain an MDP $\langle \mathcal{I}, i_0, \mathcal{A}, T', R', \gamma, \mathcal{I}_f \rangle$, and Bellman's optimality operator is thus written

$$V^*(i) = \max_a \sum_{i' \in \text{nxt}(i,a)} T'(i, a, i') \cdot [R'(i, a, i') + \gamma V^*(i')],$$

with $\text{nxt}(i, a)$ the (finite) set of next state-belief pairs under (i, a).

SSPs. When setting $\gamma = 1$, The following proposition ensures the problem is a valid SSP while infinitely many states are reachable from initial state $\langle s_0, - \rangle$, where $-$ denotes the empty history.

PROPOSITION 3.1. *Assuming that R_{AG} is bounded from above by $R_{\text{AG}}^{\max} < 0$ (in non-terminal states), the PO-OASSP is a valid SSP.*



(a) Noiseless obs. **(b) Noisy obs.**

Figure 2: Two PO-OAMDP trajectories for *legibility* tasks, with hidden cells in blue, $p_{\text{OBS}} = 1$ (left) and $p_{\text{OBS}} = 0.5$ (right).

A simple trick to retrieve a valid SSP is to combine the invalid R_{AG} with a valid R using $R'_{\text{AG}} = R_{\text{AG}} + \lambda \cdot R$ for some $\lambda > 0$.

Complexity. Proposition 2.1 tells us that PO-OAMDPs cover a larger class of problems than OAMDPs. We establish that PO-OAMDPs inherit the same main complexity results as OAMDPs, results which require assuming *Bayesian updates* for the observer's belief, what we denote by PO-OAMDP$_{BU}$. Such results are obtained considering the *value problem*, *i.e.*, determining whether a policy exists that can achieve some pre-defined value.

HSVI. We propose solving discounted PO-OAMDPs ($\gamma < 1$) using a variant of Smith and Simmons's *heuristic search value iteration* (HSVI) algorithm [16–18]. HSVI is generally used to solve POMDPs, maintaining an upper and a lower bound of V^*, respectively denoted $\overline{V}$ and $\underline{V}$, and whose representations exploit V^*'s convexity in belief space. Differences between POMDPs and PO-OAMDPs lead to several differences in HSVI. (1) $\underline{V}$ and $\overline{V}$ are expressed in information space $\mathcal{I} \equiv \mathcal{S} \times \Delta^{|\mathcal{S}|}$, not in $\Delta^{|\mathcal{S}|}$ alone. (2) PO-OAMDPs inherit local discontinuities in $\Delta^{|\mathcal{S}|}$ from OAMDPs [11, Sec. 3.2], so that we only rely on pointwise representations. (3) Usual bound initializations do not apply.

Initializing Bounds. In [10], we propose dedicated bound initializations (significantly speeding up convergence compared to trivial values) that rely on (1) separating the reward in two terms: one belief-dependent and one belief-independent; and (2) bounding the belief-dependent term using the minimum number of time steps to a terminal state.

4 CONCLUSION

PO-OAMDPs allow formalizing planning problems where an agent accounts for an external observer with partial and noisy observability, allowing to model legibility, explainability and predictibility tasks in a wider class of scenarios than previously. We demonstrate how to solve such problems by turning them into abstract MDPs and adapting a standard algorithm such as HSVI. Experiments [10] (source code available here: https://gitlab.inria.fr/po-oamdp/po-oamdp_aamas25) illustrate resulting non-trivial behaviors, such as avoiding cell $(D, 3)$ in Fig. 2 (left) to lower the probability of goal ψ_2, or taking a longer, but more visible path, in Fig. 2 (right).

REFERENCES

[1] Georgios Angelopoulos, Alessandra Rossi, Claudia Di Napoli, and Silvia Rossi. 2022. You Are In My Way: Non-verbal Social Cues for Legible Robot Navigation Behaviors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*. IEEE, 657–662. https://doi.org/10.1109/IROS47612.2022.9981754

[2] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113, 3 (12 2009), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

[3] Michael Beetz, Freek Stulp, Piotr Esden-Tempski, Andreas Fedrizzi, Ulrich Klank, Ingo Kresse, Alexis Maldonado, and Federico Ruiz-Ugalde. 2010. Generality and legibility in mobile manipulation. *Auton. Robots* 28, 1 (2010), 21–44. https://doi.org/10.1007/S10514-009-9152-9

[4] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. 2018. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. *CoRR* abs/1811.09722 (2018). arXiv:1811.09722 http://arxiv.org/abs/1811.09722

[5] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. 2019. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In *Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling (ICAPS)*. https://ojs.aaai.org/index.php/ICAPS/article/view/3463

[6] Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and predictability of robot motion. 301–308.

[7] Anca D. Dragan and Siddhartha S. Srinivasa. 2013. Generating Legible Motion. In *Robotics: Science and Systems IX, Technische Universität Berlin, Berlin, Germany, June 24 - June 28, 2013*, Paul Newman, Dieter Fox, and David Hsu (Eds.). https://doi.org/10.15607/RSS.2013.IX.024

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

[9] Gary Klein, David D. Woods, Jeffrey M. Bradshaw, Robert R. Hoffman, and Paul J. Feltovich. 2004. Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intell. Syst.* 19, 6 (2004), 91–95. https://doi.org/10.1109/MIS.2004.74

[10] Salomé Lepers, Vincent Thomas, and Olivier Buffet. 2024. Observer-Aware Probabilistic Planning under Partial Observability. *CoRR* abs/2502.10568 (2024). arXiv:2502.10568 https://arxiv.org/abs/2502.10568

[11] Shuwa Miura, Olivier Buffet, and Shlomo Zilberstein. 2024. Approximation Algorithms for Observer Aware MDPs. In *The 40th Conference on Uncertainty in Artificial Intelligence*. https://openreview.net/forum?id=UXsERjAZy8

[12] Shuwa Miura and Shlomo Zilberstein. 2021. A unifying framework for observer-aware planning and its complexity. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Vol. 161. 610–620. https://proceedings.mlr.press/v161/miura21a.html

[13] L.R. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* 77, 2 (February 1989), 257–286.

[14] Bob R. Schadenberg, Dennis Reidsma, Dirk K. J. Heylen, and Vanessa Evers. 2021. “I See What You Did There”: Understanding People’s Social Perception of a Robot and Its Predictability. *J. Hum.-Robot Interact.* 10, 3, Article 28 (jul 2021), 28 pages. https://doi.org/10.1145/3461534

[15] Phani-Teja Singamaneni, Pilar Bachiller-Burgos, Luis J. Manso, Anaís Garrell, Alberto Sanfeliu, Anne Spalanzani, and Rachid Alami. 2024. A survey on socially aware robot navigation: Taxonomy and future challenges. *Int. J. Robotics Res.* 43, 10 (2024), 1533–1572. https://doi.org/10.1177/02783649241230562

[16] Trey Smith. 2007. *Probabilistic Planning for Robotic Exploration*. Ph.D. Dissertation. The Robotics Institute, Carnegie Mellon University.

[17] Trey Smith and R.G. Simmons. 2004. Heuristic Search Value Iteration for POMDPs. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.

[18] Trey Smith and Reid G. Simmons. 2005. Point-Based POMDP Algorithms: Improved Analysis and Implementation. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. 542–549.