Explaining the Behavior of POMDP-based Agents Through the Impact of Counterfactual Information

Saaduddin Mahmud
University of Massachusetts Amherst
Amherst, MA, USA
smahmud@umass.edu

Marcell Vazquez-Chanlatte
Nissan Advanced Technology Center Silicon Valley
Santa Clara, CA, USA
marcell.vazquezchanlatte@nissan-usa.com

Stefan Witwicki
Nissan Advanced Technology Center Silicon Valley
Santa Clara, CA, USA
stefan.witwicki@nissan-usa.com

Shlomo Zilberstein
University of Massachusetts Amherst
Amherst, MA, USA
shlomo@cs.umass.edu

ABSTRACT

In this work, we consider AI agents operating in Partially Observable Markov Decision Processes (POMDPs)—a widely-used framework for sequential decision making with incomplete state information. Agents operating with partial information take actions not only to advance their underlying goals but also to seek information and reduce uncertainty. Despite rapid progress in explainable AI, research on separating information-driven vs. goal-driven behaviors remains sparse. To address this gap, we introduce a novel explanation generation framework called **Sequential Information Probing (SIP)**, to investigate the direct impact of state information, or its absence, on agent behavior. To quantify the impact we also propose two metrics under this SIP framework called **Value of Information (VoI)** and **Influence of Information (IoI)**. We then theoretically derive several properties of these metrics. Finally, we present several experiments, including a case study on an autonomous vehicle, that illustrate the efficacy of our method.

KEYWORDS

Explainability; POMDPs; Value of Information

ACM Reference Format:
Saaduddin Mahmud, Marcell Vazquez-Chanlatte, Stefan Witwicki, and Shlomo Zilberstein. 2024. Explaining the Behavior of POMDP-based Agents Through the Impact of Counterfactual Information. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

1 INTRODUCTION

In an array of applications spanning from autonomous driving [37] and communication networks [2] to healthcare [8], agents are increasingly being tasked with executing sequential decisions. Yet, in many of these scenarios, critical features essential for informed decision-making are not directly observed by the agent. Hence, the need for designing agents capable of reasoning about incomplete state information is paramount. The POMDP framework, with its capability to handle incomplete state information, has become the de facto standard for modeling sequential decision-making problems where some aspects of the world are not fully observable.

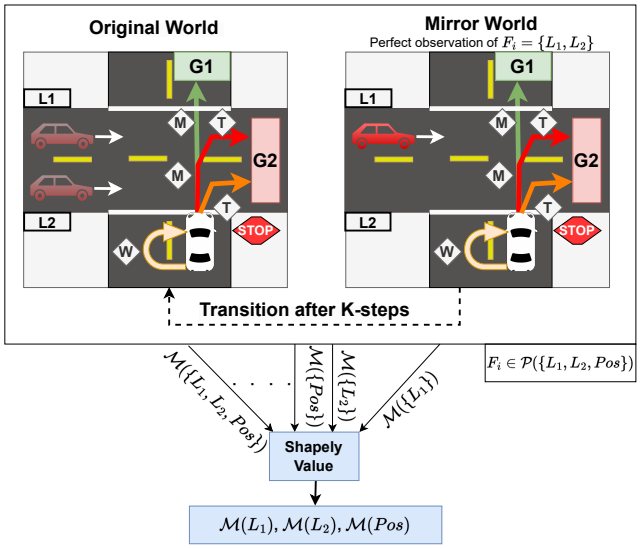


Figure 1: Example of sequential information probing (SIP) framework eliciting counterfactual behavior of an autonomous driving (AV) agent.

The complexities introduced by POMDPs often result in agents adopting diverse strategies that are harder to explain, compared with fully-observable settings. Consider the scenario in the original world in Figure 1 where the agent is trying to choose between the route $G1$ (more efficient) and the $G2$ while trying to reduce wait time. In the fully observable case, The agent waits $\{W\}$ for cars to pass on $L2$ and then moves $\{M\}$ into the intersection. If $L1$ gets blocked, it quickly takes a right turn $\{T\}$ to avoid a risky collision and ends up in $G2$; otherwise, it goes straight to $G1$. Now consider the partially observable case. The agent's sensor system takes a while to estimate oncoming cars in $L1$ and $L2$. With each step the agent chooses to wait, its estimates get better. On the other hand, with each move step $\{M$ or $T\}$ the estimate gets distorted. In this case, the agent will first wait to improve its estimate of $L2$ and cars

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

to pass (if any); then it will directly go to $G2$. It does not try to reach $G1$ because even if all the uncertainty goes away initially; after the agent takes the first move action the uncertainty about $L1$ will grow and it is risky to wait in the intersection to reduce uncertainty.

Now, if we reveal the information about $L1$ and $L2$ for 1 step, the AV will wait and let cars in $L1$ (if any) pass and make the right turn. Importantly, the wait time will be reduced because the AV does not have to fix its estimate. This will help the agent gain value by reducing the time to reach its goal. Furthermore, in this scenario the likelihood of observing the behavior $\{W, T\}$ goes down compared to the original scenario because the agent sometimes takes a direct right turn $\{T\}$. Now, if we reveal the status of $L1$ and $L2$ for multiple steps, the agent starts to behave similarly to the fully observable case, i.e., sometimes it goes to $G1$ depending on the status of $L1$ and thus can achieve even more value compared to a 1-step revelation. Under this scenario, the likelihood of $\{M, M\}$ and $\{M, T\}$ goes up and {T} goes down. Notice that if we revealed only $L1$ or $L2$ the impact will be different than revealing $L1 \wedge L2$.

The above example highlights two factors that play a key role in the resulting counterfactual behavior: 1) the duration of information probing (1-step or n-step), and 2) the type of information ($L1$ and/or $L2$). This underscores a few desired properties when developing a framework for estimating the impact of imperfect information on agents' behavior. First, a flexible option for probing information over a sequence of steps, and second, a fair mechanism for distributing the benefits among the individual pieces of information. This example also suggests useful metrics for analyzing agent behavior, such as quantifying the change in the likelihood of observing specific behaviors due to information probing and calculating the change in the value function to estimate the impact of missing information on agent performance. Despite the extensive research on explainable AI [10, 11, 34], a significant gap exists in the literature concerning the explainability of POMDP agents. While a handful of prior works (e.g., [19, 36]) have delved into explaining POMDP agents, to the best of our knowledge, none of these methods focus on analyzing how missing information affects an agent's behavior through multi-step information probing.

To bridge the existing gap, we introduce a novel framework called *Sequential Information Probing* (SIP) to analyze how imperfect information affects agents' behavior. SIP is a flexible framework that can probe information for varying durations and offer well-defined guarantees. In particular, we focus on three variants: 1) guaranteed K-step information, 2) probabilistic K-step information, and 3) myopic K-step information. Each of these methods is progressively cheaper to (pre-)compute. In the former two cases, the agent is aware of information probing and adjusts its behavior accordingly, while in the latter case, the agent utilizes information on the fly and does not adjust its strategy. The latter two can be applied for very large K while the former cannot. However, our experiment shows that the former can reveal counterfactual behaviors that require guaranteed observability for multiple time steps. In the above scenario, the agent will not make the move $\{M, M\}$ without the guarantee that $L2$ will be revealed for at least 2 steps.

SIP enables us to quantify different measurements that can help explain POMDP agent behavior. We focus here on two quantities: 1) *Value of Information* (VoI) and 2) *Influence of information* (IoI). VoI quantifies the value of different pieces of information by looking at the difference in the expected utility of the probed agent and the original agent. This quantity allows us to understand how the performance of the agent gets impacted due to missing information. Prior work on explaining MDP agents [29] has shown through large user studies that people overwhelmingly prefer explanations derived from a value function, offering further justification for studying this measure. On the other hand, IoI quantifies the changes in the negative log-likelihood of observing a particular behavior due to information probing. This helps us distinguish information-seeking and goal-oriented behavior as the likelihood of seeing an information-seeking behavior goes down when the agent already has the information. Finally, once these quantities are calculated, SIP fairly distributes the contribution of each individual piece of information toward the overall measure using the widely-used, game-theoretic Shapley value [31]. In the above scenario, revealing $L1 \wedge L2$ results in the highest value gain. The Shapley value allows us to find marginal contributions of $L1$ and $L2$ in that value gain.

Besides introducing the SIP framework, we also provide efficient methods for pre-computing key steps of VoI and IoI calculation for both discrete and continuous state spaces, which allows explanations to be displayed in real-time. Additionally, we derive several theoretical properties of VoI and IoI including a direct connection between these two quantities. Finally, Our empirical evaluation utilizes multiple POMDP models that simulate various scenarios in autonomous driving, including a study of a real autonomous vehicle (AV). In the AV setting, understanding the implications of incomplete information can offer real-world safety benefits. We report and analyze several quantitative measures including computational efficiency, consistency, and the similarities and differences among different types of information probing.

2 LITERATURE REVIEW

Explainable AI (XAI). To foster broad adoption of autonomous systems, it is crucial to establish user trust in these systems' capabilities [25, 33, 38]. It is well recognized that providing explanations can bolster trust [12, 18, 26]. In explainable AI, feature attributions are commonly used to elucidate how input features affect model outputs in tasks such as classification and regression. Notable feature attribution methods include LIME [30], SHAP [35], and Saliency Maps [1, 32]. In this study, similar to SHAP we too leverage Shapley value to estimate the marginal impact of information about different subsets of the features on the agent's decisions. However, prior methods generate explanations by single-step probing of the input; which as shown in the introduction is not sufficient to produce many nuanced behaviors in the POMDP setting. As the main focus of our paper is to explain POMDP-based agents we keep our discussion about broader XAI research short, however, some excellent reviews can be found in [4, 7, 10, 34].

Explanation for Sequential Decision Making. Research on explanations for stochastic planners has been limited, with a few notable studies. Elizalde et al. [14] highlighted crucial state factors by examining changes in the value function by altering state factors for a single step. On the other hand, works such as [6, 20, 22] explain the agent's behavior through the lens of the reward function. Additionally, broad classes of explanation methods for MDPs, such as

model reconciliation [11] and policy summarization [3], have also been proposed. Yet, a gap remains: none of these methods address the consequences of having partial information on agent decisions. Wang et al. [36] made an attempt to explain POMDP policies by conveying the comparative probabilities of various events or belief levels. Meanwhile, several techniques have been developed to explain deep reinforcement learning[19] that could potentially explain certain aspects of POMDP-based decision-making. Nevertheless, these methods neither delve into the effects of partial information on POMDP agents nor differentiate between information-seeking and goal-driven behaviors. Importantly, the idea of sequential probing is missing. Finally, ideas similar to the value of information have been explored in prior literature on active sensing [16, 39] and formal methods [9]. However, these works generally focus on optimizing the cost of observation rather than explaining POMDP agents. Further, SIP is a general framework that allows us to analyze the effect of counterfactual information on an agent's decisions through many different quantities beyond VoI.

Evaluation of Explanations. Several quantitative and qualitative metrics have been suggested for assessing automatically generated explanations. Prominent qualitative metrics encompass aspects like Complexity, Interactivity, and User Preference [28]. Conversely, frequently adopted quantitative metrics are Fidelity, Consistency, Coverage, Generalization, and Robustness [13, 28]. While both types of metrics offer important perspectives, the algorithmic nature of our study led us to prioritize the quantitative assessment of our approach.

3 BACKGROUND

This section provides a foundational understanding of the key concepts: Partially Observable Markov Decision Processes (POMDPs), solutions methods for POMDPs, and the Shapley value.

3.1 Partially Observable Markov Decision Processes (POMDPs)

A POMDP is a tuple $P = \langle S, A, O, T, \Omega, R, \gamma \rangle$ where:

- S is a finite set of states $\{s_1, s_2, s_3, ..., s_{|S|}\}$ where each state consist of a set of feature $F = \{f_1, f_2, f_3, ...f_{|F|}\}$.
- A is a finite set of actions $\{a_1, a_2, a_3, ..., a_{|A|}\}$.
- O is a finite set of observations $\{o_1, o_2, o_3, ..., o_{|O|}\}$.
- $T : S \times A \times S \to [0, 1]$ is the state transition function, where $T(s, a, s')$ represents the probability of moving from state s to state s' given action a.
- $\Omega : A \times S \times O \to [0, 1]$ is the observation function, where $O(a, s', o)$ is the probability of receiving observation o after taking action a and ending up in state s'.
- $R : S \times A \to \mathbb{R}$ is the reward function.
- $\gamma \in [0, 1)$ is the discount factor.

3.2 Solutions Methods for POMDPs

The objective of a POMDP agent is to maximize its expected cumulative discounted reward given by:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \lambda^t R(s_t, a_t)\right],$$

where s_t and a_t is the state and action at time t. The solution to a POMDP is a policy that maximizes this objective. A policy $\pi : B \to \Delta A$ maps a belief $b \in B$ to a distribution over actions in A where a belief $b \in B$ is a probability distribution over S. A value function V induced by a policy π specifies the expected total reward of executing policy π starting from belief b:

$$V(b) = R(b, \pi(b)) + \gamma \sum_{b'} Pr(b'|b, \pi(b))V(b')$$

where $Pr(b'|b, a)$ is the probability of transitioning to belief b' using the observation received after taking action a. Similarly, we can also derive a Q-function for a given policy π:

$$Q(b, a) = R(b, a) + \gamma \sum_{b'} Pr(b'|b, a)V(b').$$

It is known that V^*, the value function associated with the optimal policy π^*, can be approximated arbitrarily closely by a convex, piecewise-linear function [23]. Hence:

$$V(b) = \max_{\alpha \in \Gamma}(\alpha \cdot b); Q(b, a) = \max_{\alpha \in \Gamma \wedge \mathcal{A}(\alpha) = a}(\alpha \cdot b);$$

where Γ is a finite set of vectors called α-vectors and $\mathcal{A}$ maps each α-vector to an action $a \in A$. However, when dealing with a continuous state space, it is convenient to view the POMDPs as a belief-space MDP. In this case, the agent considers its belief space B as its state space S. For a belief-space MDP, we can rewrite the Q-function as follows:

$$Q(b, a) = R(b, a) + \gamma \mathop{\mathbb{E}}_{b' \sim T(b,a)}[V(b').]$$

Here, $T(\cdot, \cdot)$ is the belief transition model. Usually, we do not have access to $T(\cdot, \cdot)$, and therefore, it is convenient to apply a model-free reinforcement learning algorithm to solve them. In such settings, the Q function is usually represented using a neural network.

3.3 Shapley Values

The Shapley value [31] is a concept from cooperative game theory used to fairly distribute a collective reward among players based on their individual contributions. For a game with characteristic function $v : \mathcal{P}(N) \to \mathbb{R}$, where N is the set of players and $\mathcal{P}(F)$ is the power set of N, the Shapley value $C_i(v)$ of player i is defined as:

$$C_i(v) = \frac{1}{|N|!} \sum_{pr \in Pr_N} \left[v(G_i^{pr} \cup \{i\}) - v(G_i^{pr})\right]. \quad (1)$$

Here, Pr_N is the set of all permutations of N, and G_i^{pr} is the set of players preceding i in permutation pr. The Shapley value essentially measures the average marginal contribution of a player across all possible permutations. In our case, VoI or IoI can be considered as the characteristic function v and each player represents information related to a particular feature.

4 SIP: SEQUENTIAL INFORMATION PROBING

In this section, we first give an introduction to the general framework of Sequential Information Probing (SIP). We then define two important measurements under this framework, namely VoI and IoI. In SIP (Figure 1), we have two POMDP models: the original and the mirror. The original and the mirror are equivalent in every aspect except that in the mirror, the agent can perfectly observe $F_i \in \mathcal{P}(F)$.

For example in the scenario from the introduction we can construct a mirror world where the agent could have perfect information about $\{L1\}$, $\{L2\}$, or $\{L1 \wedge L2\}$. To isolate the impact of sequential information probing, we take two measurements of a quantity: one by placing the agent in the original world, and another by placing the agent in a mirror world, allowing it to take a sequence of steps and then transporting it back to the original world. The difference between these two measurements is the impact of sequential information probing on the corresponding quantity. We repeat this for different subsets of the features and finally marginalize it using the Shapley value to calculate the fair impact of each individual feature.

How the agent is transported back to the original world has computational implications when calculating measurements such as VoI or IoI and the type of counterfactual behavior they can induce. We consider 3 transportation approaches:

(1) $[KS]$: The agent stays exactly K steps in the mirror world and adapts its policy to exploit the available information.
(2) $[GE]$: The agent stays K steps in expectation with $K \sim Geometric(1-\lambda)$ in the mirror world. Under this transportation approach, the agent can transport back to the original world at each step with probability $1-\lambda$. Similarly, the agent adapts its policy to exploit the available information.
(3) $[MY]$: The agent remains in the mirror world for K steps. However, the agent does not adapt its policy to utilize the additional information available in the mirror world but rather keeps on using the policy computed for the original world, leading to a myopic use of the available information. Computationally this strategy is much simpler to compute and may be preferable in certain settings.

4.1 Value of Information

The Value of Information (VoI) tries to quantify how imperfect information about a subset of features affects agents' performance. More specifically, VoI quantifies the expected utility the agent loses due to the lack of information about the subset of feature $F_i \subseteq \mathcal{P}(F)$ for the next K time steps. If certain features have high VoI then a utility-maximizing agent is more likely to seek that information and vice-versa. Hence, VoI can be used as an indicator of that agent's propensity to seek information in the near future.

We now formally define VoI. Consider the value function that gives the expected utility the agent could achieve from current belief b, and true state s_t, if the perfect information about $F_i \subseteq \mathcal{P}(F)$ is given for the next K steps with KS strategy, $V_{F_i,s_t}(b,K)$:

$$R(b^{s_t}_{F_i}, \pi^{F_i}(b^{s_t}_{F_i})) + \gamma \sum_{b'} Pr(b'|b^{s_t}_{F_i}, \pi^{F_i}(b^{s_t}_{F_i}))V_{F_i,s_{t+1}}(b', K-1) \quad (2)$$

Here, $b^{s_t}_{F_i}$ is the updated belief b we get using information about $F_i = F_i(s)$. $b^{s_t}_{F_i}$ can be written as $b^{s_t}_{F_i} = normalized(\overline{b^{s_t}_{F_i}})$ where:

$$\overline{b^{s_t}_{F_i}}(s) = \begin{cases} b(s), & \text{if } K > 0 \wedge F_i(s) = F_i(s_t), \\ 0, & \text{if } K > 0 \wedge F_i(s) \neq F_i(s_t), \\ b(s), & \text{if } K \leq 0, \end{cases}$$

Also, here π^{F_i} is the optimal policy when the feature set F_i is revealed. Now, we can similarly define V_{F_i,s_t} function for MY strategy simply by replacing π^{F_i} with $\pi^{\emptyset}$ (i.e. original POMDP policy). Finally, We can write $V_{F_i,s_t}(b,M,\lambda)$ for GE strategy:

$$R(b^{s_t}_{F_i}, \pi^{F_i}(b^{s_t}_{F_i})) + \gamma \sum_{b'} Pr(b'|b^{s_t}_{F_i}, \pi^{F_i}(b^{s_t}_{F_i}))dV_\lambda, \quad (3)$$

where $dV_\lambda = \left[\lambda V_{F_i,s_{t+1}}(b', M) + (1-\lambda)V_{F_i,s_{t+1}}(b', \bar{M})\right]$, M indicates the mirror world and $\bar{M}$ indicates the original world. Here, b_{F_i,s_t} has a similar definition as above with M being equivalent to $K > 0$ and $\bar{M}$ being equivalent to $K \leq 0$. Based on this we can define the value of the sequence of information in KS strategy as the expected difference between the value function without information probing V and V_{F_i,s_t}:

$$VoI(b, F_i, K) = E_{s_t \sim b}[V_{F_i,s_t}(b,K) - V(b)] \quad (4)$$

In reality, we do not have access to the true state and hence expectation is taken over the agent's belief of the true state. We can similarly define VoI for GS strategy $VoI(b, F_i, M, \lambda)$. Finally, in order to get the marginal value of each feature we apply the Shapley Value Framework:

$$C_i(VoI) = \frac{1}{|F|!} \sum_{pr \in Pr_N} \left[VoI(F^{pr}_{F_i} \cup \{F_i\}, b, K) - VoI(F^{pr}_{F_i}, b, K)\right]$$

We now describe some of the important theoretical properties[1] of the VoI.

Property 1. *Null 1 :* $VoI(F_i, b, 0) = 0$; $VoI(F_i, b, M, 0) = 0$.

Property 2. *Null 2 :* $VoI(\emptyset, b, \cdot) = 0$.

Property 3. *Efficiency:* $VoI(F, b, \cdot) = \sum_i^{|F|} C_i(VoI)$.

Property 4. *Relation between KS and MY :*
$VoI_{MY}(F_i, b, K) \leq VoI_{KS}(F_i, b, K)$

Property 5. *Bounded:* $0 \leq VoI(F_i, b, K) \leq QMDP(b) - V^*(b)$. *where QMDP is defined as in [17].*

While other feature attribution methods can be used one of the main reasons for selecting the Shapley value is Property 3. By combining properties 3 and 5, it can be shown that VoI can be can be normalized to a scale of [0,1]. This allows easy comparison of the value of a different set of features within a specific POMDP and across different POMDPs.

4.2 Influence of Information

While VoI quantifies how imperfect information affects an agent's behavior in the near future it does provide a direct explanation of a particular behavior. Therefore, we also introduce the Influence of Information (IoI) to quantify how the likelihood of observing a behavior $\tau = \{b, s_0, a_0, o_1, s_1, a_1, ..., o_T, s_T, a_T\}$ changes when probed with sequential information. To do this, IoI tries to calculate the Negative-log likelihood (NLL) ratio of observing τ under different subsets of perfect information and original POMDP. Note that IOI's unit is in bits when used with base-2 logarithm. This measure can be interpreted as the surprise associated with the observed behavior given the sequence of information. The NLL of behavior τ is as

[1]Proofs can be found in the extended version of the paper at https://sequential-information-probing.github.io.

follows:

$$-\log P(\tau|\pi) = -\sum_{t=0}^{T}\log\pi(a_t|b_t) - \sum_{t=1}^{T}\log T(s_t|a_{t-1},s_{t-1}) - \sum_{t=1}^{T}\log O(o_t|a_{t-1},s_t) \quad (5)$$

If we consider an entropy regularized [15] policy π then the probability distribution over the actions A is:

$$\pi(a|b) = \frac{e^{Q(b,a)}}{\sum_{a^j\in A} e^{Q(b,a^j)}}$$

On the other hand, if we consider a deterministic policy π then we can consider ϵ-smooth probability distribution over the actions A:

$$\pi(a|b) = \begin{cases} 1-\epsilon, & \text{if } a = argmax_{a\in A}Q(b,a), \\ \frac{\epsilon}{|A|-1}, & \text{otherwise} \end{cases}$$

Finally, we define the influence of sequential information probing on the behavior τ as NLL-Ratio under the original policy π and π^{F_i}.

$$IoI(\tau-\tau_s, F_i) = E_{\tau_s\sim\mathcal{D}}[-\log\frac{P(\tau_{F_i}|\pi^{F_i})}{P(\tau|\pi)}] \quad (6)$$

Here, τ_s is the sequence of state from τ. Since we do not observe the states IoI is defined with $\tau-\tau_s$. $\mathcal{D}$ is a distribution[2] from which we sample τ_s. Finally, τ_{F_i} is the feature set F_i augmented trajectory meaning for a belief b_t from τ we use $b^{s_t}_{t,F_i}$. Now, it can be shown that the T-terms and O-terms cancel each other and the $IoI_{F_i}(\tau-\tau_s,F_i)$ becomes simplified:

$$E_{\tau_s\sim\mathcal{D}}[\sum_{t=0}^{T}\log\pi(a_t|b_t) - \sum_{t=0}^{T}\log\pi^{F_i}(a_t|b^{s_t}_{t,F_i})] \quad (7)$$

Finally, for MY strategy we can write $IoI_{F_i}(\tau-\tau_s,F_i)$ as follows:

$$E_{\tau_s\sim\mathcal{D}}[\sum_{t=0}^{T}\log\pi(a_t|b_t) - \sum_{t=0}^{T}\log\pi(a_t|b^{s_t}_{t,F_i})] \quad (8)$$

Property 6. *Null ($F_i=\emptyset$): $IoI(\tau-\tau_s,F_i)=0$.*

Property 7. *Efficiency of marginal IoI: $IoI(\tau-\tau_s,\emptyset) = \sum_i^{|F|} C_i(IoI)$.*

Property 8. *Relation to VoI: With $\mathcal{D}(\tau_s) = \prod_{t=0}^{T} p(s_t|b_t)$ and entropy regularized policy the following relation holds,*

$$|\leq IoI(\tau-\tau_s,F_i) - \sum_{t=0}^{T}[VoI(b_t,F_i) - QoI(b_t,F_i)]| \leq log(|A|)$$

.

Here, QoI can be derived by replacing the value function with the Q-function in Equation 4.

5 CALCULATING VOI AND IOI

In this section, we describe methods for computing VoI and IoI for both discrete and continuous state spaces. It consists of a pre-computation step where we calculate all the probed value functions for $F_i\in\mathcal{P}(F)$. After that explanations can be generated efficiently during the deployment of the system in real time.

[2]The natural choice for distribution is the one induced by the observation in τ.

5.1 Discrete State Space

We can pre-compute the main component of VoI, by generating a combined POMDP model $\overline{P_{F_i}} = \langle\overline{S},A,\overline{O},\overline{T},\overline{\Omega},R,\gamma\rangle$ of the original and mirror worlds. For the KS strategy we can define $\overline{P_{F_i}}$ as follows:

- $\overline{S}$ is an augmented state space with $\overline{F} = \{f_1,f_2,f_3,...f_{|F|},time\}$. Here, $time\in\{0,1,...,K\}$ and indicates the number of steps till information injection ends.
- $\overline{O}$ is a finite set of observations by adding the true value of features in $F_i\subseteq F$ and $Time$. For example, instead of receiving o_i at $time=t$, the agent receives $\{o_i,t,F_i(s)\}$ when $t>0$ and $\{o_i,0,\emptyset\}$ otherwise.
- $\overline{T}:\overline{S}\times A\times\overline{S}\to[0,1]$ is the augmented state transition function defined as follows:

$$\overline{T}(\overline{s},a,\overline{s}') = \begin{cases} T(s,a,s'), & \text{if } time(s)=time(s')=0, \\ T(s,a,s'), & \text{if } time(s)=time(s')+1, \\ 0, & \text{otherwise} \end{cases}$$

- $\overline{\Omega}: A\times\overline{S}\times\overline{O}\to[0,1]$ is the augmented observation function, defined as follows:

$$\overline{\Omega}(a,\overline{s'},\overline{o}) = \begin{cases} \Omega(a,s',o), & \text{if } Time(s)=0\wedge\overline{o}=\{o,0,\emptyset\}, \\ \Omega(a,s',o), & \text{if } Time(s)>0\wedge\overline{o}=\{o,Time(s),F_i(s)\}, \\ 0, & \text{otherwise} \end{cases}$$

For the GE strategy we can define $\overline{P_{F_i}}$ as follows:

- $\overline{S}$ is an augmented state space with $\overline{F} = \{f_1,f_2,f_3,...f_{|F|},M\}$ where M is an indicator variable showing whether the agent is in the mirror world or not.
- $\overline{O}$ is a finite set of observations by adding the true value of features in $F_i\subseteq F$ similar to the KS strategy.
- $\overline{T}:\overline{S}\times A\times\overline{S}\to[0,1]$ is the augmented state transition function defined as follows:

$$\overline{T}(\overline{s},a,\overline{s}') = \begin{cases} T(s,a,s'), & \text{if } M(s)=M(s')=False, \\ (1-\lambda)*T(s,a,s'), & \text{if } M(s)=True\wedge M(s')=False, \\ \lambda*T(s,a,s'), & \text{if } M(s)=M(s')=True, \\ 0, & \text{if } M(s)=False\wedge M(s')=True \end{cases}$$

- $\overline{\Omega}: A\times\overline{S}\times\overline{O}\to[0,1]$ is the augmented observation function, defined as follows:

$$\overline{\Omega}(a,\overline{s'},\overline{o}) = \begin{cases} \Omega(a,s',o), & \text{if } M(s')=False\wedge\overline{o}=\{o,\emptyset\}, \\ \Omega(a,s',o), & \text{if } M(s')=True\wedge\overline{o}=\{o,F_i(s)\}, \\ 0, & \text{otherwise} \end{cases}$$

Notice that the combined POMDP of GE has $K/2$ times the smaller number of states and K times the smaller number of observations. Also, notice that the MY strategy only requires that we estimate the $V_{F_i,s_t}(b,K)$ under the original POMDP policy. This can be estimated with a straightforward Monte-Carlo sampling strategy.

5.2 Continues State Space

Many practical problems require modeling with continuous state spaces. In such scenarios, explicit representation of T and O is not available and it is common to train the agent using a simulation of the environment. Further, large or infinite state spaces can make it impossible to apply α-vector policy representation. To solve this, the predominant approach is to use deep reinforcement learning on the belief state. Further, if an explicit belief update is not available,

Algorithm 1 Meta-CDQL

Require: $Q_{\overline{\theta}}, B_{\overline{\phi}}$, Transition strategy TS

1: $Q^M_\theta \leftarrow$ Initialize randomly
2: $B^M_\phi \leftarrow$ Initialize randomly
3: Replay_Buffer $\leftarrow \emptyset$
4: **while** Condition not met **do**
5: $h_0, s_0 \sim simulate(Q_{\overline{\theta}}, B_{\overline{\phi}})$
6: $F_i \sim \mathcal{P}(F)$
7: $D \leftarrow$ Collect_Transitions$(Q^M_\theta, B^M_\phi, h_0, s_0, F_i, TS)$
8: Replay_Buffer $\leftarrow$ Update(Replay_Buffer, D)
9: Update Q^M_θ, B^M_ϕ using Eq. 9, 10, 11, 12
10: **end while**
11: **return** $Meta - Q_{\phi,\theta}$

one can jointly learn the belief update function along with the policy. We now extend VoI and IoI to such scenarios.

For VoI and IoI we need to estimate a set of counterfactual values and policies corresponding to each $F_i \in F$. To address this, a simple meta counterfactual Q-value estimation algorithm (Meta-CDQL) is designed that jointly learns the Q-value function for each $F_i \in F$. Note that we assume that the Q-function for the policy that will be used during deployment is given. The choice for using Deep-Q Learning [27] is purely due to simplicity and the estimation could be done with other value-function-based Deep RL algorithms [5].

A sketch of the algorithm is given in Algorithm 1. The algorithm takes input Q-function and belief update function $Q_{\overline{\theta}}, B_{\overline{\phi}}$ for the original POMDP. The process starts by initializing the a meta-Q function Q^M_θ, a meta-belief function B^M_ϕ and a replay buffer (Lines $1-3$). There are three key differences in the training process compared to the standard implementation of deep Q-learning. First, the starting history and state distribution are defined by the original policy (Line 5). This is because during deployment we are required to compute VoI or IoI starting from different history under the original policy. Second, the Collect_Transition function takes input a start state, history, and the feature set that will have perfect information and generate transitions for training. The important detail is that with λ probability at each step, the simulation will stop giving information about F_i for GE strategy. For KS it will stop after K steps. Also, there should not be any transition that goes from the original world to the mirror world in D. Finally, we train the Q-function using the standard Deep-Q learning loss function:

$$\mathcal{L}(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}(r_t + \gamma Q^M_\theta(B^M_\phi(h^{t+1}), a_{max}) - Q^M_\theta(B^M_\phi(h^t), a^t))^2 \tag{9}$$

Here, N is the batch size and a_max is the action associated with the max Q-value. Importantly, we introduce 3 regularization to enforce Q^M_θ, B^M_ϕ to emulate $Q_{\overline{\theta}}, B_{\overline{\phi}}$ when $F_i = \emptyset$. This is desirable for several reasons including speeding up training, faithfulness to the original policy, and necessary for the MY strategy. First, we want the difference between Q-estimates to be low when $F_i = \emptyset$

$$\mathcal{L}^{QD}(\theta,\phi) = (Q^M_\theta(B^M_\phi(h), a) - Q_{\overline{\theta}}(B_{\overline{\phi}}(h), a))^2 \tag{10}$$

Second, we want the representation to be similar when $F_i = \emptyset$.

$$\mathcal{L}^{RD}(\phi) = (B^M_\phi(h) - B_{\overline{\phi}}(h))^2 \tag{11}$$

Finally, we also want the two belief representations to induce similar Q value in the original policy $Q_{\overline{\theta}}$ when $F_i = \emptyset$.

$$\mathcal{L}^{RD2}(\theta,\phi) = (Q_{\overline{\theta}}(B^M_\phi(h), a) - Q_{\overline{\theta}}(B_{\overline{\phi}}(h), a))^2 \tag{12}$$

All the loss functions can be jointly optimized as:

$$\mathcal{L}_(\theta,\phi) + \beta_1\mathcal{L}^{QD}(\theta,\phi) + \beta_2\mathcal{L}^{RD}(\theta) + \beta_3\mathcal{L}^{RD2}(\theta,\phi) \tag{13}$$

Once Q^M_θ and B^M_ϕ are learned we can calculate $V^\pi_{F_i}(b)$ as:

$$\mathbb{E}_{s\sim\mathcal{G}(h)}\left[\max_{a\in A} Q^M_\theta(B^M_\phi(h_{F_i,s}), a) - \max_{a\in A} Q_{\overline{\theta}}(B_{\overline{\phi}}(h, a))\right] \tag{14}$$

Here, $\mathcal{G}(h)$ is a generator model that captures the distribution over states given history.

6 EVALUATION

In this section, we experimentally analyze SIP, dividing our discussion into two subsections. First, we conduct a quantitative analysis of different variants of SIP, rigorously examining its computational performance of pre-compute step, consistency, similarity, predictive power, and faithfulness. Note that consistency is an important measurement because the explanations are generated by solving a set of surrogate POMDP models. Therefore, depending on the solution the generated explanations might be different which is undesirable. In the following subsection, we illustrate a case study, of deploying SIP to enhance the decision-making transparency of an actual autonomous vehicle.

6.1 Quantitative Evaluation

Environments. Our quantitative analysis is conducted across four POMDP environments, each of which encapsulates interactions of an autonomous vehicle (AV) under distinct scenarios:

(1) **LEFT:** A model where the AV makes decisions about an unprotected left turn.
(2) **CROSS:** A scenario in which the AV is situated at the intersection of two crossroads, as depicted in the introduction.
(3) **TL:** A controlled intersection where the interaction between the AV and an oncoming vehicle is mediated by a traffic light.
(4) **PED:** A model representing the interaction between the AV and a pedestrian near an intersection.

The models encompass 12, 12, 24, and 36 states, respectively. Generally, each model assigns a substantial penalty for placing the AV in hazardous situations, such as collisions or near-collisions. Moreover, agents are incentivized to attain their objective—typically traversing the intersection—as promptly as possible.

Evaluation Metrics. We evaluate our method on several quantitative dimensions. First, we examine the computational performance of KS, and GE during the pre-computation steps using average time in milliseconds over 30 different runs. Subsequently, we evaluate the consistency of the explanation generation method by conducting the pre-computation step 30 times, generating explanations for a single scenario from the 30 different pre-computed values/policies. We repeated this for M scenarios. The measure of consistency is derived from the average distance from the mean, as

	LEFT	CROSS	TL	PED
KS	0.88	0.48	0.78	0.53
GE	**0.89**	**0.53**	**0.85**	**0.69**
MY	0.81	0.42	0.77	0.50

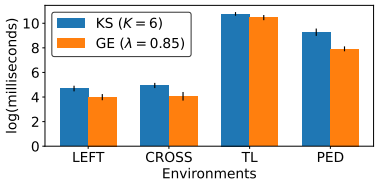
Table 1: Correlation between VoI and IoI

	LEFT		CROSS		TL		PED	
	VoI	IoI	VoI	IoI	VoI	IoI	VoI	IoI
KS	4.0	14.0	2.9	3.7	**8.2**	2.9	**8.2**	6.5
GE	**35.5**	35.5	**35.4**	35.4	7.3	3.0	4.6	5.4
MY	2.2	∞	1.1	∞	2.1	∞	2.0	∞

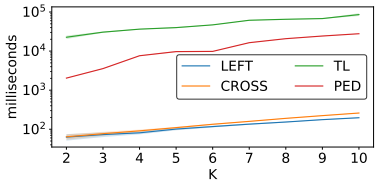
Table 2: Consistency score of different SIP method (higher score means higher consistency)

	LEFT		CROSS		TL		PED	
	VoI	IoI	VoI	IoI	VoI	IoI	VoI	IoI
KS vs. GE	0.86	0.91	**0.80**	0.95	**0.81**	0.77	0.31	0.61
KS vs. MY	0.82	0.91	0.60	0.96	0.73	0.79	0.19	0.59
GE vs. MY	**0.89**	**0.98**	0.64	**0.97**	0.80	**0.97**	**0.61**	**0.90**

Table 3: Similarity of explanations using different SIP methods (higher means more similar)



(a) Pre-computation Time



(b) K vs. Time for SIP(KS)

Figure 2: Computational load for different SIP methods

	LEFT		CROSS		TL		PED	
	VoI	IoI	VoI	IoI	VoI	IoI	VoI	IoI
KS	2.4	2.6	1.2	1.1	1.3	1.4	1.1	1.2
GE	2.6	2.7	1.3	1.3	1.5	1.7	1.1	1.3
MY	**1.8**	∞	**1.1**	∞	**1.8**	∞	**1.7**	∞

Table 4: Consistency score of different SIP method using Meta-CDQL (higher score means more consistent)

	LEFT		CROSS		TL		PED	
	L^{QD}	$L^{\not QD}$	L^{QD}	$L^{\not QD}$	L^{QD}	$L^{\not QD}$	L^{QD}	$L^{\not QD}$
KS	**97.2**	95.2	**94.5**	91.3	**91.3**	88.4	**93.8**	88.0
GE	**98.7**	96.5	**98.9**	96.3	**94.7**	90.4	**95.1**	92.1

Table 5: Faithfulness to Policy induced by $Q_{\bar{\theta}}$

expressed below:

$$\text{Consistency} = -\log\left[\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N}\sum_{j=1}^{N}\left(E_i^j - \overline{E}\right)^2\right] \tag{15}$$

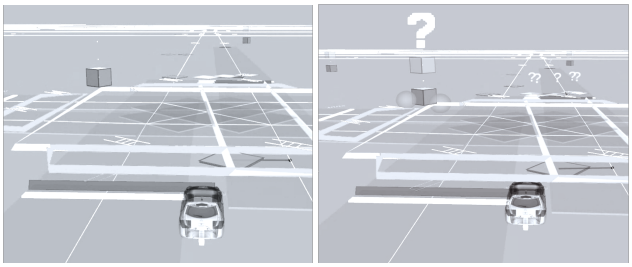
Here,E_i^j is the explanation computed for scenario i using the pre-computed value from run j, This metric discerns the disparity among different generated explanations. Additionally, to explore the similarities and differences among different SIP methods, we employ Kendall's rank correlation [21]. The reasoning behind utilizing rank correlation for similarity stems from the fact that both VoI and IoI rank features from the most to the least valuable/impactful, and hence a metric that assesses the alignment between two rankings is apt. Kendall's rank correlation, denoted as K_τ, is a commonly employed metric with $K_\tau \in [-1, 1]$. A K_τ greater than zero indicates a positive correlation, and vice versa.

Computational Performance. The computational proficiency during the pre-computation phase for various SIP transportation approaches is illustrated in Figure 2. The models were solved utilizing the Julia implementation of the widely-used SARSOP algorithm [24]. Anticipatedly, *KS* (Figure 2(a)) demands the lengthiest computation time relative to *GE*. Furthermore, as observed from Figure 2(b), increasing K for *KS* results in a non-linear escalation in computation time, indicating *KS* may be less apt for long horizon probing under computational budget constraints. Nonetheless, given that it is part of the pre-computation step, time restrictions may not pose a significant challenge in numerous practical contexts. During the explanation computation phase, *KS* and *GE* showcase comparable performance, whereas *MY* (only for VoI), necessitating sampling, requires a longer time. This may bear implications, particularly in the settings that require real-time explanation generation; making *KS* and *GE* more desirable in such settings.

Consistency Analysis. The summarized results of the consistency of the explanations, generated utilizing SARSOP and Meta-CDQL, are presented in Tables 3 and 4. The emergence of the pattern $IoI \leq VoI$ is attributed to the fact that VoI is only calculated with initial belief whereas IoI is calculated with a trajectory of belief. Notice that for *MY* it is a little bit different. For IoI, we do not calculate any new policy and get infinitely consistence explanations[3]. For VoI, the inconsistency comes from sampling error. Finally, Meta-CDQL displays less consistency in comparison to SARSOP, which can be attributed to the intrinsic instability of deep reinforcement learning algorithms and the stochastic nature of neural network training. Although a more stable RL algorithm could improve the results, it was not explored in our experiments. Overall, we found the level of consistency in the tabular case for all SIP approaches to be highly reliable. Our suggestion for continuous state space would be to use an ensemble of a learned Meta Q-function instead of one to mitigate the consistency issue. As Meta Q-function will be pre-computed for most applications; the additional computational load can be reasonable.

Exploring the Predictive Capability of VoI. Next, we examine the relationship between VoI and IoI. Given a behavior τ, the VoI is calculated at the initial belief of τ, while the IoI is determined for the entirety of the behavior τ. We then compare the ranking induced by VoI with the absolute value of IoI—given its potential to be either positive or negative. VoI ranks features with respect to their future

[3]Note that we only train Q_θ^M.



(a) Waiting at a controlled intersection (b) Sudden change in traffic light

Figure 3: Deployment of VoI on real-world AV system

performance from the initial belief, whereas IoI identifies which feature indeed influenced the behavior. It can be hypothesized that features deemed most valuable are likely to exert an impact on future behavior, essentially positioning VoI as a predictor for IoI prior to the behavior's instantiation. Observations from Table 1 reveal a persistent positive correlation between these two metrics, and in numerous instances, the correlation is notably strong. The variances observed in the results can be ascribed to differing degrees of stochasticity in the observation function O and transition T in different environments.

Similarity. As alluded to in the introduction, the duration of information probing exerts an impact on agent behavior, as does the assurance regarding the duration of the information probing. For instance, *GE*, even with an expected probing surpassing 6, cannot induce the $\{M, M\}$ behavior in the agent discussed in the introduction (i.e., the *CROSS* environment). This sets the stage for an exploration of the similarity between the methods. Insights from Table 3 indicate that *GE* and *MY* bear more resemblance to each other than to *KS*. In the context of *GE*, the agent remains unaware of the duration for which the information will be probed, whereas, in *MY*, the agent is oblivious to the probing of information altogether. Both scenarios deprive the agent of a long-term guarantee of accurate information, leading to the adoption of different

Impact of Regularization. Finally, we examine the influence of employing L^{DQ} regularization within the Meta-CDQL in terms of how faithfully it predicts actions of the original POMDP policy. Table 5 reveals that L^{DQ} is effective in improving faithfulness. Additionally, we observed an expedited convergence of the Meta-CDQL as a beneficial side effect of its application.

Overall, our experiments demonstrate that SIP can be made to generate very consistent explanations and differences in computational requirements of different SIP methods. Furthermore, the similarities across various SIP methods, and between IoI and VoI, offer insight into the nuance of the different methods we present.

6.2 Application of SIP in a Real AV System

We implemented SIP within a real autonomous vehicle (AV) system, specifically within a decision-making module, which is tasked with executing decisions such as initiating motion (go), halting (stop), edging, and turning. The decision-making mechanism within the AV system is modeled using POMDPs. Typical vehicular interactions during navigation (e.g., with a pedestrian, other vehicle, traffic control, lane change, etc.) are represented through separate POMDPs. These decision models work together continually throughout the AV's operation to navigate it safely toward the goal.

We employ Value of Information (VoI) to identify how different features are influencing the AV's behavior at each decision point. Subsequently, the VoI is visualized on the AV's developer interface. For visualization of the VoI, a "?" symbol is superimposed atop each feature modeled in the set of POMDP models. The opacity of this "?" symbol is modulated by the corresponding VoI value. An example of this is presented in Figure 4, which illustrates a real-world scenario wherein the AV halts at a regulated intersection during a test drive with a safety driver in an urban environment, awaiting the traffic light to transition to green (Figure 4a). While stationary, the AV maintains an accurate estimation of the light's status therefore we do not see any "?" symbol. After the light transitions to green, the AV remains stopped even though it could start moving. This delay is attributed to the agent's existing uncertainty regarding the light's status after a recent change. During this interval, a "?" symbol illuminates atop the traffic light (Figure 4b), indicative of its high VoI. Lesser bright "?" symbols are also observed on an oncoming vehicle as after the car crosses the traffic light they will become important for navigation. Note that the scene is complex and includes multiple other cars, pedestrians, and traffic control elements. We focus on the simple traffic light interaction for ease of illustration.

Leveraging the efficient pre-computation method presented in this paper, we are able to generate explanations from approximately 20 POMDPs at a rate of 10 explanations per second per POMDP. Within this system, VoI assists developers in exploring potential rationales behind various observed behaviors in real-time. Moreover, it facilitates developers in debugging the system by revealing situations where features that should have had a significant impact, were not highlighted by the VoI, and vice versa.

7 CONCLUSIONS AND FUTURE WORK

We presented SIP, a new framework for analyzing POMDP-based agent behaviors by sequentially probing counterfactual perfect information and observing behavioral changes. Metrics like the change in the value function (VoI) or change in the likelihood of observing a given behavior (IoI) were proposed to quantify the impact of the given information, while Shapley values help isolate individual features' contributions. Three SIP variants and methods for calculating VoI and IoI were explored, suitable for both discrete and continuous states, with a thorough examination of their theoretical and computational properties. Finally. SIP's practical application was demonstrated in an autonomous vehicle (AV), showcasing its role in enhancing transparency. Future work will focus on improving SIP's pre-computation efficiency and consistency in deep learning systems and conducting extensive user studies.

ACKNOWLEDGMENTS

This work was supported in part by the Alliance Innovation Lab Silicon Valley, and by NSF Grants 1954782 and 2321786.

REFERENCES

[1] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Bianchi-Berthouze. 2020. Evaluating saliency map explanations for convolutional

neural networks: A user study. International Conference on Intelligent User Interfaces (2020).
[2] Eitan Altman. 2000. Applications of Markov decision processes in communication networks: A survey. Technical Report, INRIA (2000).
[3] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2019. Summarizing agent strategies. Autonomous Agents and Multi-Agent Systems (2019).
[4] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion (2019).
[5] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine (2017).
[6] Josh Bertram and Peng Wei. 2018. Explainable deterministic MDPs. arXiv preprint arXiv:1806.03492 (2018).
[7] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. IJCAI-2017 workshop on explainable AI (XAI) (2017).
[8] Richard J. Boucherie and Nico M. van Dijk. 2017. Markov decision processes in practice. Springer.
[9] Peter Bulychev, Franck Cassez, Alexandre David, Kim Guldstrand Larsen, Jean-François Raskin, and Pierre-Alain Reynier. 2012. Controllers with minimal observation power (application to timed systems). International Symposium on Automated Technology for Verification and Analysis (2012).
[10] Nadia Burkart and Marco Huber. 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research (2021).
[11] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2020. The emerging landscape of explainable automated planning & decision making. International Joint Conference on Artificial Intelligence (2020).
[12] Jessie YC Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science (2018).
[13] Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. 2022. Framework for evaluating faithfulness of local explanations. International Conference on Machine Learning (2022).
[14] Francisco Elizalde, Enrique Sucar, Julieta Noguez, and Alberto Reyes. 2009. Generating explanations based on Markov decision processes. Mexican International Conference on Artificial Intelligence (2009).
[15] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. 2019. A Theory of Regularized Markov Decision Processes. International Conference on Machine Learning (2019).
[16] Eric A. Hansen, Andrew G. Barto, and Shlomo Zilberstein. 1996. Reinforcement Learning for Mixed Open-loop and Closed-loop Control. Neural Information Processing Systems (1996).
[17] Milos Hauskrecht. 2000. Value-Function Approximations for Partially Observable Markov Decision Processes. Journal of artificial intelligence research (2000).
[18] Bradley Hayes and Julie A. Shah. 2017. Improving robot controller transparency through autonomous policy explanation. ACM/IEEE International Conference on Human-Robot Interaction (HRI) (2017).
[19] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz Rodríguez. 2020. Explainability in Deep Reinforcement Learning. Knowledge-Based Systems (2020).
[20] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. 2019. Explainable reinforcement learning via reward decomposition. IJCAI/ECAI Workshop on Explainable Artificial Intelligence (2019).
[21] M. G. Kendall. 1938. A New Measure of Rank Correlation. Biometrika (1938).
[22] Omar Khan, Pascal Poupart, and James Black. 2009. Minimal sufficient explanations for factored Markov decision processes. International Conference on Automated Planning and Scheduling (2009).
[23] Mykel J. Kochenderfer, Tim A. Wheeler, and Kyle H. Wray. 2022. Scalable Gradient Ascent for Controllers in Constrained POMDPs. MIT Press.
[24] Hanna Kurniawati, David Hsu, and Wee Sun Lee. 2009. SARSOP: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces. Robotics: Science and Systems (2009).
[25] Michael P. Linegang, Heather A. Stoner, Michael J. Patterson, Bobbie D. Seppelt, Joshua D. Hoffman, Zachariah B. Crittendon, and John D. Lee. 2006. Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2006).
[26] Joseph E. Mercado, Michael A. Rupp, Jessie YC Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human-agent teaming for Multi-UxV management. Human Factors (2016).
[27] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. NeurIPS Deep Learning Workshop (2013).
[28] Christoph Molnar. 2022. Interpretable Machine Learning: A Guide For Making Black Box Models Explainable. Lulu.com.
[29] Samer B. Nashed, Saaduddin Mahmud, Claudia V. Goldman, and Shlomo Zilberstein. 2023. Causal Explanations for Sequential Decision Making Under Uncertainty. International Conference on Autonomous Agents and Multiagent Systems (2023).
[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. (2016).
[31] Lloyd S. Shapley. 1988. A Value for n-person Games. Classics in Game Theory (1988).
[32] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv preprint arXiv:1312.6034 (2014).
[33] Kristen Stubbs, Pamela J. Hinds, and David Wettergreen. 2007. Autonomy and common ground in human-robot interaction: A field study. IEEE Intelligent Systems (2007).
[34] Erico Tjoa and Cuntai Guan. 2019. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems (2019).
[35] Erik trumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems (2014).
[36] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. International Conference on Autonomous Agents and Multiagent Systems (2016).
[37] Kyle Hollins Wray, Stefan J Witwicki, and Shlomo Zilberstein. 2017. Online decision-making for scalable autonomous systems. International joint conference on artificial intelligence (2017).
[38] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. ACM Conference on Fairness, Accountability, and Transparency (2020).
[39] Shlomo Zilberstein and Stuart J. Russell. 1993. Anytime Sensing Planning and Action: A Practical Model for Robot Control. International Joint Conference on Artificial Intelligence (1993).