# Reinforcement Learning for Human-AI Collaboration via Probabilistic Intent Inference

**Anonymous authors**
Paper under double-blind review

**Keywords:** Reinforcement Learning, Probabilistic Intent Inference, Human-AI Collaboration, Belief-Space Planning, Decision-Making Under Uncertainty.

## Summary

Effective collaboration between humans and AI agents is increasingly essential as autonomous systems take on critical roles in domains like disaster response, healthcare, and robotics. However, achieving robust human-AI collaboration remains challenging due to the uncertainty, complexity, and unpredictability of human behavior, which is often difficult to convey explicitly to AI agents. This paper presents a belief-space reinforcement learning framework that enables AI agents to implicitly and probabilistically infer latent human intentions from behavioral data and integrate this understanding into robust decision-making. Our approach models human behavior at both the action (low) and subtask (high) levels, combining these with human and agent state information to construct a comprehensive belief state for the AI agent. We demonstrate that this belief state follows the Markov property, enabling the derivation of an optimal Bayesian policy under human and task uncertainty. Deep reinforcement learning is used to train an offline Bayesian policy across a wide range of human and task uncertainties, allowing real-time deployment to support effective human-AI collaboration. Numerical experiments demonstrate the effectiveness of the proposed policy in terms of cooperation, adaptability, and robustness.

## Contribution(s)

1. We develop a decision-making framework that represents the human behavioral model at two levels—low-level actions and high-level subtasks—allowing the AI agent to anticipate long-term human goals and adapt to changing task priorities in real-time.
   **Context:** Unlike prior models that focus on human rationality at a single (action) level, our approach incorporates hierarchical intent modeling, enhancing goal-aware human-AI collaboration and improving adaptability to dynamic environments.
2. We propose a structured belief state that captures state information alongside the posterior distribution of human intent, serving as a sufficient statistic for optimal Bayesian decision-making in human-AI collaboration.
   **Context:** Unlike existing Partially Observable Markov Decision Process (POMDP)-based frameworks that maintain beliefs over partially observable states, our belief state explicitly models uncertainty in high-level human intent, leading to more informed and adaptive decision-making under uncertainty.
3. We develop a deep reinforcement learning (DRL) approach that optimizes the AI agent's decision-making over the belief space, enabling dynamic adaptation to inferred human intent for effective long-term human-AI collaboration.
   **Context:** Unlike existing methods that optimize AI agents for pre-specified human tasks or rely on explicit feedback, our approach leverages a belief-space policy trained on human behaviors. This policy captures the AI's belief about human intent—including uncertainty in their goals and actions (theory of mind)—to optimize decision-making accordingly. This enables efficient real-time adaptation without requiring explicit human feedback.

# Reinforcement Learning for Human-AI Collaboration via Probabilistic Intent Inference

**Anonymous authors**
Paper under double-blind review

## Abstract

Effective collaboration between humans and AI agents is increasingly essential as autonomous systems take on critical roles in domains like disaster response, healthcare, and robotics. However, achieving robust human-AI collaboration remains challenging due to the uncertainty, complexity, and unpredictability of human behavior, which is often difficult to convey explicitly to AI agents. This paper presents a belief-space reinforcement learning framework that enables AI agents to implicitly and probabilistically infer latent human intentions from behavioral data and integrate this understanding into robust decision-making. Our approach models human behavior at both the action (low) and subtask (high) levels, combining these with human and agent state information to construct a comprehensive belief state for the AI agent. We demonstrate that this belief state follows the Markov property, enabling the derivation of an optimal Bayesian policy under human and task uncertainty. Deep reinforcement learning is used to train an offline Bayesian policy across a wide range of human and task uncertainties, allowing real-time deployment to support effective human-AI collaboration. Numerical experiments demonstrate the effectiveness of the proposed policy in terms of cooperation, adaptability, and robustness.

## 1 Introduction

**Motivation:** The rapid advancement of autonomous systems has enabled AI agents to take on increasingly complex roles in healthcare, manufacturing, disaster response, and robotics (Zhang et al., 2021; Hauptman et al., 2023; Berretta et al., 2023). However, achieving effective human-AI collaboration remains an open challenge due to the ambiguity and unpredictability of human intent (Bhatt et al., 2021; Charalampous et al., 2017). Unlike multi-agent settings, human-AI collaboration presents unique challenges due to the dynamic and context-dependent nature of human decision-making, which is influenced by cognitive biases, task priorities, workload constraints, and environmental factors. This inherent complexity necessitates real-time adaptation for AI agents, particularly in scenarios lacking explicit supervision. While explicit feedback mechanisms have been explored to mitigate uncertainty (Arulkumaran et al., 2017; Kiran et al., 2021; Abeyruwan et al., 2023), these approaches are often impractical in high-stakes environments where human input is limited, delayed, or unavailable (Kaluarachchi et al., 2021).

**Prior Work:** Existing methods for human-AI teaming fall into two broad categories: human-guided learning and human-inferred intent modeling (Ambhore, 2020; Obaigbena et al., 2024; Teng et al., 2023). Human-guided approaches, including human-in-the-loop reinforcement learning (HITL) (Retzlaff et al., 2024; Ho & Griffiths, 2022; Lu, 2019), reward shaping (Hu et al., 2020), imitation learning (Zare et al., 2024) and inverse reinforcement learning (IRL) (Ziebart et al., 2008; Arora & Doshi, 2021), rely on human feedback to shape AI behavior. While effective, these methods require continuous human demonstrations, corrections, or explicit reward signals. In contrast, human-inferred approaches, such as behavior modeling, Bayesian inference, and intent recognition (Hoffman et al., 2024; Singh et al., 2020; Ni et al., 2023; Nasernejad et al., 2021), attempt to infer intent from past interactions. These methods typically estimate current human intent but fail to model the rationality behind long-term intent evolution, making them impractical for cooperation planning that requires predicting future human intent over extended horizons.

Existing Partially Observable Markov Decision Process (POMDP)-based models (Hadfield-Menell et al., 2016; Mai et al., 2025) model uncertainty using belief distributions over latent states. This approach is employed in frameworks such as Cooperative Inverse Reinforcement Learning (CIRL) (Hadfield-Menell et al., 2016) and human-robot task allocation techniques (Ali et al., 2022; Lee et al., 2022). However, these methods fail to explicitly model hierarchical intent structures. Hierarchical IRL (Nair et al., 2018; Sun et al., 2018; Chen et al., 2023) are also developed for learning structured human behavior by decomposing decision-making into multiple levels, typically to achieve a more interpretable representation of human policies. These methods are designed to extract hierarchical task structures or policies from demonstrations, assuming a stationary decision-making process without considering the influence of AI agents' decisions. While effective for modeling human behavior in a structured manner, they are not designed for real-time adaptation or for capturing non-stationary, evolving human intent that may emerge in interactive settings.

**Proposed Framework:** This paper presents a belief-space reinforcement learning framework that enables AI agents to infer and respond to high-level human intent in real time, without requiring explicit feedback or retraining. Cooperation is structured around a set of pre-specified subtasks that human and AI agents must collaboratively complete to achieve a shared objective. Unlike existing belief-planning models, which primarily focus on uncertainty in state transitions, our approach introduces a structured belief-state representation that jointly models low-level actions and high-level task objectives, allowing the AI agent to dynamically adjust its strategy as human intent evolves. By explicitly modeling uncertainty in human decision-making, our method supports long-term planning and adaptive cooperation.

At the core of our framework is a hierarchical belief state, which serves as a sufficient statistic for optimal decision-making by encapsulating the agent state, human state, and posterior human intent distribution. We prove that this belief state satisfies the Markov property, allowing for the derivation of an optimal Bayesian policy that accounts for long-term uncertainties in both state transitions and evolving human intent. To efficiently approximate this policy, we introduce a deep reinforcement learning (DRL) framework that pre-trains the AI agent on diverse human behaviors and task uncertainties. This offline-trained policy enables real-time deployment, allowing the AI agent to implicitly infer human intent and adapt dynamically without retraining. Extensive numerical experiments demonstrate the effectiveness of our framework, particularly in environments where human intent evolves dynamically and task uncertainty is high.

## 2 Background - A Markov Decision Process

A Markov decision process (MDP) representing human-agent collaboration is defined by the 4-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R \rangle$. Here, $\mathcal{S} = \mathcal{S}^A \times \mathcal{S}^H$ denotes the state space, comprising all possible agent and human states, while $\mathcal{A} = \mathcal{A}^A \times \mathcal{A}^H$ represents the joint action space of the agent and human. The state transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ defines the probability of reaching a new state $\mathbf{s}'$ given the current state $\mathbf{s}$ and joint action $\mathbf{a} = (\mathbf{a}^A, \mathbf{a}^H)$. Without loss of generality, the state transition function can be factorized into independent components for the human and agent: $\mathcal{P}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = \mathcal{P}^H(\mathbf{s}^H, \mathbf{a}^H, \mathbf{s}'^H) \cdot \mathcal{P}^A(\mathbf{s}^A, \mathbf{a}^A, \mathbf{s}'^A)$, where $\mathcal{P}^H(\mathbf{s}^H, \mathbf{a}^H, \mathbf{s}'^H)$ represents the transition dynamics of the human, and $\mathcal{P}^A(\mathbf{s}^A, \mathbf{a}^A, \mathbf{s}'^A)$ describes the agent's state transition. This factorization simplifies modeling by treating human and agent dynamics independently while still capturing their joint interaction within the MDP framework. The reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ assigns a real-valued reward for each state-action pair. Specifically, $R(\mathbf{s}, \mathbf{a})$ captures the cooperative reward earned when the joint action $\mathbf{a} = (\mathbf{a}^A, \mathbf{a}^H)$ is taken in state $\mathbf{s}$.

## 3 Implicit Probabilistic Human Intent Inference

In many collaborative domains, humans and AI agents must work together without direct communication or explicit feedback, making it critical to accurately capture human intent for seamless coordination. We consider an AI agent capable of probabilistically inferring human intent using limited behavioral data. This inference is conducted with respect to a set of $N$ predefined subtasks, denoted as $T^1, T^2, \ldots, T^N$, which may vary in priority and can be performed by either the human or the AI agent. When the AI agent infers that the human is moving toward a given subtask, it can prioritize complementary subtasks, thereby playing a supportive role in human-AI teaming.

Our framework leverages probabilistic intent learning to capture the evolution of human intent over time. Let $p_0 = [p(T_0 = T^1), \ldots, p(T_0 = T^N)]$ represent the prior probability distribution over the subtasks, where $\sum_{j=1}^{N} p_0(j) = 1$. As the agent observes the sequence of human states $\mathbf{s}^H_{0:k} = (\mathbf{s}^H_0, \ldots, \mathbf{s}^H_k)$, it updates its understanding of the human's current intent using the posterior distribution:

$$p_k = \Big[p(T_k = T^1 \mid \mathbf{s}^H_{0:k}), \ldots, p(T_k = T^N \mid \mathbf{s}^H_{0:k})\Big]^T. \tag{1}$$

AI agents often lack access to the specific human actions that produce the observed sequence of human states $\mathbf{s}^H_{0:k}$, limiting the applicability of intent inference techniques that rely on observing human state-action pairs. In the following paragraphs, we introduce the human model and recursive approach to probabilistically capture human intent posterior using only state sequences.

**Human Modeling:** Let $\eta_k \in \{0,1\}^N$ be a subtask tracker at time step $k$, which keeps track of subtask completion. Specifically, $\eta_k(j) = 0$ indicates that the $j$-th subtask has not yet been completed by either the human or the AI agent, while $\eta_k(j) = 1$ indicates that the subtask has been completed. A subtask tracker with $\eta_k = [0, \ldots, 0]$ implies that no subtasks have been completed, whereas $\eta_k = [1, \ldots, 1]$ indicates that all subtasks have been completed. Let $\{\mathcal{G}^1, \ldots, \mathcal{G}^N\}$ denote the terminal states for the subtasks, where $\eta_k(j)$ switches from 0 to 1 when either the human or the agent first reaches state $\mathcal{G}^j$. The subtask tracker $\eta_k$ is updated based on the observed states of the human and agent up to time step $k$, or recursively as:

$$\eta_{k+1}(j) = \sum_{r=0}^{k+1} 1_{\mathbf{s}^A_r = \mathcal{G}^j \text{ or } \mathbf{s}^H_r = \mathcal{G}^j}, \quad \eta_{k+1}(j) = \begin{cases} 1 & \text{if } \mathbf{s}^A_{k+1} = \mathcal{G}^j \text{ or } \mathbf{s}^H_{k+1} = \mathcal{G}^j \\ \eta_k(j) & \text{otherwise} \end{cases}, \quad j = 1, \ldots, N, \tag{2}$$

where $\mathcal{G}^j$ denotes the terminal state of the $j$-th subtask.

Assuming the human is the primary agent, acting based on its state and the remaining subtasks, the optimal human policy at any given state $\mathbf{s}^H$ and subtask tracker $\eta$ can be computed as:

$$\pi^{*,H}(\mathbf{s}^H, \eta) = \operatorname*{argmax}_{\pi^H} \mathbb{E}\Big[\sum_{t=0}^{h} \gamma^t r_t \mid \mathbf{s}^H_0 = \mathbf{s}^H, \eta_0 = \eta, \mathbf{a}^H_{0:h} \sim \pi^H\Big], \quad \text{for all } \mathbf{s}^H \in \mathcal{S}^H \text{ and } \eta \in \{0,1\}^N, \tag{3}$$

where the expectation is taken with respect to stochastic transitions and rewards, $\gamma \in (0, 1]$ is a discount factor, and $h$ is the horizon. This formulation considers the human as the sole actor in the environment, modeling a human-AI setting where the AI agent plays a supportive role. We model the human as a sub-optimal reinforcement learning agent, with behavior approximated by a stochastic form of the optimal policy in (3) as:

$$\pi^H(\mathbf{a}^H | \mathbf{s}^H, \eta) := P(\mathbf{a}^H | \mathbf{s}^H, \eta) = \begin{cases} q + \frac{1-q}{|\mathcal{A}^H|} & \text{If } \mathbf{a}^H = \pi^{*,H}(\mathbf{s}^H, \eta) \\ \frac{1-q}{|\mathcal{A}^H|} & \text{If } \mathbf{a}^H \neq \pi^{*,H}(\mathbf{s}^H, \eta) \end{cases}, \text{ for } \mathbf{a}^H \in \mathcal{A}^H, \mathbf{s}^H \in \mathcal{S}^H, \eta \in \{0,1\}^N, \tag{4}$$

where $\pi^H(\mathbf{a}^H \mid \mathbf{s}^H, \eta)$ indicates the probability that the human takes action $\mathbf{a}^H$ at state $\mathbf{s}^H$ and subtask tracker $\eta$, and $q \in [0, 1]$ represents the human's rationality at the action level, referred to as the low-level rationality rate. Higher values of $q$ (close to 1) correspond to more rational behavior, while lower values (close to 0) reflect more random behavior.

**Recursive Inference of Human Intent:** Using the human model in Equation (4), we can infer the posterior distribution of human intent, as described in Equation (1). Let $p_k(j) = P(T_k = T^j \mid \mathbf{s}^H_{0:k}, \eta_k)$ denote the $j$-th element of the posterior of human intent at time step $k$, where $\eta_k$ represents the subtask completion status based on the prior experiences of both the human and the AI agent. When a new human state $\mathbf{s}^H_{k+1}$ is observed and both the agent and the human are not at a terminal goal state, the posterior of human intent can be updated recursively as follows:

$$\begin{aligned} p_{k+1}(j) = P(T_{k+1} = T^j \mid \mathbf{s}^H_{0:k+1}, \eta_{k+1}) &\propto \sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1}, \mathbf{a}^H_k = \mathbf{a}^H, T_{k+1} = T^j \mid \mathbf{s}^H_{0:k}, \eta_{k+1}) \\ &= \sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_k, \mathbf{a}^H_k = \mathbf{a}^H) P(\mathbf{a}^H_k = \mathbf{a}^H \mid \mathbf{s}^H_k, T_{k+1} = T^j) P(T_{k+1} = T^j \mid \mathbf{s}^H_{0:k}, \eta_{k+1}) \\ &= \sum_{\mathbf{a}^H \in \mathcal{A}^H} \mathcal{P}^H(\mathbf{s}^H_k, \mathbf{a}^H, \mathbf{s}^H_{k+1}) \pi^H(\mathbf{a}^H \mid \mathbf{s}^H_k, \eta = \mathbf{e}^j)\, p_k(j), \end{aligned} \tag{5}$$

where $\mathbf{e}^j$ is a binary vector of size $N$ with all elements set to 1, except for the $j$-th element, which is 0. This expression provides a fully recursive update for the posterior distribution of human intent. Note that in this scenario, no subtask is completed at time step $k+1$, implying that $\eta_{k+1} = \eta_k$.

If the human reaches the terminal state of the $i$th subtask at time step $k+1$, i.e., $\mathbf{s}^H_{k+1} = \mathcal{G}^i$, the posterior of human intent no longer depends on past human state sequence. Under these conditions, predicting human intent is equivalent to quantifying the probability of human performing the next subtask given its current state and the subtask tracker, that is, $P(T_{k+1} = T^j \mid s^H_{k+1} = \mathcal{G}^i, \eta_{k+1})$. We estimate the next human intent by propagating human states using the low-level action policy $\pi^H$ from Equation (4) until the next subtask is completed (i.e., $\eta_t \neq \eta_{t-1}$). This process can be achieved through the Monte Carlo method, allowing the approximation of the predictive distribution $p_{k+1} \approx p_{s^H_{k+1},\eta_{k+1}}$, where $p_{s^H_{k+1},\eta_{k+1}}(l)$ represents the ratio of trajectories that ended up at state $\mathcal{G}^l$ as the next subtask. To account for variability in human rationality during the subtask selection, we introduce:

$$p_{k+1} = \alpha \cdot p_{\mathbf{s}^H_{k+1},\eta_{k+1}} + \frac{1-\alpha}{\|\mathbf{1}_N - \eta_{k+1}\|_1} \cdot (\mathbf{1}_N - \eta_{k+1}), \tag{6}$$

where $\alpha \in [0,1]$ represents the human's rationality level at the subtask level, referred to as the high-level rationality rate. Higher values of $\alpha$ indicate a more rational subtask selection, while lower values suggest a more random decision.

## 4 Adaptive Planning with Implicit Learning of Human Intention

In human-AI teaming, the optimal cooperative policy for an AI agent can be derived using the standard Markov Decision Process formulation from Section 2, given that the AI agent has complete knowledge of human intent (i.e., the agent is fully aware of the human intention). However, in the absence of explicit interactions, human intent is uncertain and the AI agent must make decisions given the uncertainty in the inferred human intention. This section introduces a framework that enables the AI agent to make optimal decisions under partial and probabilistic knowledge of human intent.

**Belief State:** We introduce the concept of a belief state, which encompasses all relevant information the AI agent needs to make informed decisions at each time step. Let $\mathbf{s}_k = [\mathbf{s}^A_k, \mathbf{s}^H_k]^T$ denote the joint state of the agent and human, $\eta_k$ represent the status of subtasks completed by both entities, and $p_k$ indicate the posterior probability of human intent inferred by the AI agent at time step $k$. The *belief state* at time step $k$ is then defined as:

$$\mathbf{b}_k = [\mathbf{s}_k, \eta_k, p_k]^T, \tag{7}$$

where $\mathbf{b}_k$ is a vector of size $|\mathbf{s}_k| + 2N$. This belief state $\mathbf{b}_k$ resides in an uncountable belief space $\mathcal{B} = \mathcal{S} \times 2^N \times \Delta_N$, where $\mathbf{s}_k \in \mathcal{S}$ is the joint state space of the agent and human, $\eta_k \in \{0,1\}^N$ is the subtask tracker, and $p_k$ is a posterior sample from the simplex $\Delta_N$ of size $N$.

**Belief Transitions:** In this part, we derive the belief transitions and demonstrate that they satisfy the Markov property. Let $(\mathbf{b}_0, \mathbf{a}^A_0, \ldots, \mathbf{a}^A_{k-1}, \mathbf{b}_k)$ represent the sequence of agent actions and belief states up to time $k$. If the agent takes action $\mathbf{a}_k^A$, the next belief state transition is given by:

$$\begin{aligned}
&P(\mathbf{b}'=[\mathbf{s}',\eta',p'] \mid \mathbf{b}_0,\mathbf{a}_0^A,\dots,\mathbf{a}_{k-1}^A,\mathbf{b}_k,\mathbf{a}_k^A)\\
&=P(\mathbf{s}_{k+1}=\mathbf{s}',\eta_{k+1}=\eta' \mid \mathbf{s}_0,\eta_0,p_0,\mathbf{a}_0^A,\dots,\mathbf{s}_k,\eta_k,p_k,\mathbf{a}_k^A)\\
&\times P(p_{k+1}=p' \mid \mathbf{s}_{k+1}=\mathbf{s}',\eta_{k+1}=\eta',\mathbf{s}_0,\eta_0,p_0,\mathbf{a}_0^A,\dots,\mathbf{s}_k,\eta_k,p_k,\mathbf{a}_k^A)\\
&=P(\mathbf{s}_{k+1}=\mathbf{s}' \mid \mathbf{s}_k,\eta_k,p_k,\mathbf{a}_k^A)P(\eta_{k+1}=\eta' \mid \eta_k,\mathbf{s}_{k+1}=\mathbf{s}')P(p_{k+1}=p' \mid \mathbf{s}_{k+1}^H=\mathbf{s}'^H,\eta_{k+1}=\eta',\mathbf{s}_k^H,\eta_k,p_k),
\end{aligned} \tag{8}$$

where the last line indicates that the belief transition depends only on the most recent belief state and agent action, satisfying the Markov property: $P(\mathbf{b}' \mid \mathbf{b}_0,\mathbf{a}_0^A,\dots,\mathbf{a}_{k-1}^A,\mathbf{b}_k,\mathbf{a}_k^A)=P(\mathbf{b}' \mid \mathbf{b}_k,\mathbf{a}_k^A)$.

Since the state transitions of the human and the AI agent are mutually independent, the first term in the last expression of Equation (8) can be expanded as:

$$P(\mathbf{s}_{k+1}=[\mathbf{s}'^A,\mathbf{s}'^H] \mid \mathbf{s}_k,\eta_k,p_k,\mathbf{a}_k^A)=\mathcal{P}(\mathbf{s}_k^A,\mathbf{a}_k^A,\mathbf{s}'^A)\left[\sum_{j=1}^{N}p_k(j)\sum_{\mathbf{a}^H\in\mathcal{A}^H}\mathcal{P}^H(\mathbf{s}_k^H,\mathbf{a}^H,\mathbf{s}'^H)\pi^H(\mathbf{a}^H \mid \mathbf{s}_k^H,\mathbf{e}^j)\right]. \tag{9}$$

The second term can be further expressed according to Equation (2) as:

$$P(\eta_{k+1}=\eta' \mid \eta_k,\mathbf{s}_{k+1}[\mathbf{s}'^A,\mathbf{s}'^H])=\prod_{j=1}^{N}\left[1_{\eta_k(j)=0 \text{ and } \eta'(j)=1}\,1_{\mathbf{s}'^H=\mathcal{G}^j \text{ or } \mathbf{s}'^A=\mathcal{G}^j}+1_{\eta_k(j)=\eta'(j)}\right], \tag{10}$$

The last term in (8) returns 1 only if $p'$ matches the next posterior of human intent computed based on $\mathbf{s}'^H,\eta',\mathbf{s}_k^H,\eta_k$, and $p_k$ using Equations (5) and (6), otherwise returns 0.

**Reward in Belief Space:** After the human and the AI agent take actions, the cooperative reward they obtain through collaboration is determined by their current states, the actions taken, and the subtask tracker, defined as $R(\mathbf{s}=[\mathbf{s}^A,\mathbf{s}^H],\eta,\mathbf{a}^A,\mathbf{a}^H)$. The existence of $\eta$ ensures that the reward corresponding to a specific subtask can only be obtained by one entity only once throughout the whole task. Within the current framework, the human action $\mathbf{a}^H$ is not observable at the agent's side, but the human's future behavior can be inferred using the recursively updated probability distribution $p$, and the completion status of the subtasks can also be tracked in real-time by $\eta$. Using the belief state, the reward function can be rewritten as:

$$R^B(\mathbf{b}=[\mathbf{s},\eta,p],\mathbf{a}^A)=E_{\mathbf{s},\eta,\mathbf{a}^H\mid\mathbf{b}}\left[R(\mathbf{s},\eta,\mathbf{a}^A,\mathbf{a}^H)\right]=\sum_{j=1}^{N}p(j)\sum_{\mathbf{a}^H\in\mathcal{A}^H}\pi^H(\mathbf{a}^H \mid \mathbf{s}^H,\mathbf{e}^j)R(\mathbf{s},\eta,\mathbf{a}^A,\mathbf{a}^H), \tag{11}$$

where the intention posterior $p(j)$ and human policy $\pi^H$ are used to predict the unobservable human action $\mathbf{a}^H\in\mathcal{A}^H$. The reward structure in (11) explicitly incorporates inferred human intent, enabling the AI agent to optimize cooperation and decision-making given the uncertainty in the inferred human intent.

Based on the previous definitions and explanations, the decision-making and action-taking process of the AI agent can be modeled as a MDP as described in Section 2, defined using a 4-tuple $\langle\mathcal{B},\mathcal{A}^A,\mathcal{P}^B,R^B\rangle$, where $\mathcal{B}$ is the belief space, $\mathcal{A}^A$ is the action space of the AI agent, and $\mathcal{P}^B:\mathcal{B}\times\mathcal{A}^A\times\mathcal{B}\to[0,1]$ is the belief transition function derived in the previous section, defined as $\mathcal{P}^B(\mathbf{b},\mathbf{a}^A,\mathbf{b}')=P(\mathbf{b}' \mid \mathbf{b},\mathbf{a}^A)$.

**Optimal Belief State Policy:** This part introduces the agent policy that accounts for the uncertainty in human intent to make optimal cooperative decisions. Therefore, an agent must make decisions based not only on its own state but also on the human's state, the subtask tracker, and the inferred intent posterior, all of which are reflected in the belief state.

We define a deterministic policy in the belief space as $\mu^A:\mathcal{B}\to\mathcal{A}^A$, mapping an agent action to any given belief state $\mathbf{b}\in\mathcal{B}$. The optimal Bayesian policy for the agent can be expressed as:

$$\mu^{*,A}(\mathbf{b})=\underset{\mu^A}{\operatorname{argmax}}\,\mathbb{E}\left[\sum_{t=0}^{h}\gamma^t R^B(\mathbf{b}_t,\mathbf{a}_t^A) \mid \mathbf{b}_0=\mathbf{b},\mathbf{a}_{0:h}^A\sim\mu^A\right],\quad \text{for all } \mathbf{b}\in\mathcal{B}, \tag{12}$$

where the maximization is over all possible deterministic policies within the belief space. The expectation in Equation (12) accounts for uncertainties in both agent and human state transitions, as well as the posterior distribution of human intent. This optimal Bayesian policy $\mu^{*,A}$ provides the maximum expected rewards for any given belief, enabling optimal cooperation through inferred human intents. By leveraging probabilistic knowledge of intent in real time, the AI agent selects actions that maximize cooperative outcomes without requiring explicit human feedback.

## 5 Deep Reinforcement Learning for Bayesian Planning

The large size of the belief space makes the computation of the exact solution for optimization in (12) infeasible. This paper introduces a deep reinforcement learning approach that approximates the optimal Bayesian policy over the entire belief space. By leveraging known belief state transitions and reward structures, our method allows for pre-training the policy without real agent or human data, instead using simulated belief samples to represent a wide range of potential behaviors and task uncertainties that might be faced in practice. This approach enables efficient policy learning that generalizes across scenarios prior to observing any real data from the human or AI agent. Depending on the size of the action space, either value-based or policy-based deep reinforcement learning methods can be utilized. In this paper, we focus on value-based approaches, specifically for finite action spaces.

For an arbitrary policy $\mu^A$ defined over the belief space, we define the expected discounted reward function at a belief state $\mathbf{b} \in \mathcal{B}$ after taking action $\mathbf{a}^A \in \mathcal{A}^A$ and then following the policy $\mu^A$ as:

$$Q_{\mu^A}(\mathbf{b}, \mathbf{a}^A) = \mathbb{E}\left[\sum_{t=0}^{h} \gamma^t R^B(\mathbf{b}_t, \mathbf{a}_t^A) \mid \mathbf{b}_0 = \mathbf{b}, \mathbf{a}_0^A = \mathbf{a}^A, \mathbf{a}_{1:h} \sim \mu^A\right], \tag{13}$$

where the expectation is taken over belief transitions. The optimal Q-function, $Q^*$, provides the maximum expected return under the optimal policy $\mu^{*,A}$, such that:

$$\mu^{*,A}(\mathbf{b}) = \underset{\mathbf{a}^A \in \mathcal{A}^A}{\operatorname{argmax}}\, Q^*(\mathbf{b}, \mathbf{a}^A), \text{ for any } \mathbf{b} \in \mathcal{B}. \tag{14}$$

We employ the Double Deep Q-network (DDQN) technique (Van Hasselt et al., 2016), a well-known value-based deep reinforcement learning method. The optimal Q-function $Q^*$ is approximated using two feed-forward neural networks: the Q-network $Q_{\mathbf{w}}$ and the target network $Q_{\mathbf{w}^-}$. Both networks share the same architecture and are initialized with identical, randomly assigned weights. The input to the Q-network is the belief state $\mathbf{b}$, and its outputs are the Q-values $Q_{\mathbf{w}}(\mathbf{b}, \mathbf{a}^1), \ldots, Q_{\mathbf{w}}(\mathbf{b}, \mathbf{a}^{|\mathcal{A}^A|})$, each corresponding to an agent action.

Training involves a fixed size replay memory $\mathcal{D}$ to store the transition experiences of beliefs. The episode starts with an initial belief state sampled from the belief space, denoted as $\mathbf{b}_0 = [\mathbf{s}_0, \eta_0, p_0]^T \in \mathcal{B}$, and ends when all subtasks have been performed. At each time step $t$ in an episode, an action $\mathbf{a}_t^A$ is selected using the epsilon-greedy policy based on $Q_{\mathbf{w}}$:

$$\mathbf{a}_t^A \sim \begin{cases} \text{greedy: } \operatorname{argmax}_{\mathbf{a}^A \in \mathcal{A}^A} Q_{\mathbf{w}}(\mathbf{b}_t, \mathbf{a}^A) & \text{with probability } 1-\epsilon, \\ \text{random: } \mathbf{a}^A \in \mathcal{A}^A & \text{with probability } \epsilon, \end{cases} \tag{15}$$

where $\epsilon \in [0, 1]$ is the exploration rate. After selecting the agent's action $\mathbf{a}_t^A$, the next belief state $\mathbf{b}_{t+1}$ is sampled according to the belief transition $P(\cdot \mid \mathbf{b}_t, \mathbf{a}_t^A)$ from Equation (8). The reward $r_{t+1}$ is then calculated, and the experience tuple $(\mathbf{b}_t, \mathbf{a}_t^A, \mathbf{b}_{t+1}, r_{t+1})$ replaces the oldest entry in $\mathcal{D}$ if the memory limit is reached.

The Q-network $Q_{\mathbf{w}}$ is iteratively updated after collecting a sufficient number of experiences using a minibatch sampled from $\mathcal{D}$: $Z = \{(\tilde{\mathbf{b}}_n, \tilde{\mathbf{a}}_n, \tilde{\mathbf{b}}_{n+1}, \tilde{r}_{n+1})\}_{n=1}^{N_{\text{batch}}} \sim \mathcal{D}$, where each tuple $(\tilde{\mathbf{b}}_n, \tilde{\mathbf{a}}_n, \tilde{\mathbf{b}}_{n+1}, \tilde{r}_{n+1})$ is generated randomly and the samples do not necessarily represent consecutive tuples in replay memory. For each sample in the minibatch, the target values are computed as:

$$y_n = \tilde{r}_{n+1} + Q_{\mathbf{w}^-}(\tilde{\mathbf{b}}_{n+1}, \underset{\mathbf{a}^A \in \mathcal{A}^A}{\operatorname{argmax}}\, Q_{\mathbf{w}}(\tilde{\mathbf{b}}_{n+1}, \mathbf{a}^A)), \tag{16}$$

where $Q_{\mathbf{w}^-}$ (the target network) provides the target values at the action holding the maximum value of the policy network. The Q-network weights $\mathbf{w}$ are then updated by minimizing the mean squared error:

$$\mathbf{w} \leftarrow \mathbf{w} - \beta \nabla_{\mathbf{w}} \frac{1}{N_{\text{batch}}} \sum_{n=1}^{N_{\text{batch}}} \Big( y_n - Q_{\mathbf{w}}(\tilde{\mathbf{b}}_n, \tilde{\mathbf{a}}_n) \Big)^2, \tag{17}$$

where $\beta$ is the learning rate. The optimization can be performed using stochastic gradient descent, such as with the Adam optimizer (Kingma & Ba, 2015). The target network $Q_{\mathbf{w}^-}$ gets updated by iteratively becoming closer to the Q-Network through the following soft update:

$$\mathbf{w}^- \leftarrow \tau \mathbf{w} + (1 - \tau)\mathbf{w}^-, \tag{18}$$

where $\tau$ is the soft update rate, controlling the rate of change of $\mathbf{w}^-$ to ensure the training stability.

The training process is considered complete when the loss function converges and the policy performance meets the desired criteria. Afterward, the learned Q-network approximates the optimal agent policy, such that $\mu^{*,A}(\mathbf{b}) \approx \operatorname{argmax}_{\mathbf{a}^A} Q_{\mathbf{w}}(\mathbf{b}, \mathbf{a}^A)$. Figure 1 depicts the human-agent collaboration where the pre-trained $Q_{\mathbf{w}}$ generates the agent's actions based on the belief state computed in real time. Policy generation involves two steps: (1) computing Q-values using $Q_{\mathbf{w}}$ for all possible actions and (2) selecting the action with the maximum Q-value. The overall inference complexity is $O(L \times M + |\mathcal{A}^A|)$, where $L$ is the number of layers and $M$ is the number of neurons per layer.

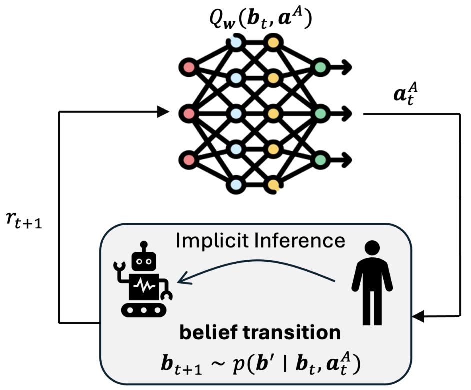


Figure 1: Bayesian planning for decision-making in belief space under uncertain human intent.

## 6 Numerical Experiments

In this paper, we analyze the performance of our proposed method in a 2D grid-world environment, as illustrated in Figure 2. It is a maze consisting of 64 states and 9 subtasks, where the orange cells represent terminal states of the subtasks, and gray cells indicate impenetrable obstacles. An AI agent and a human collaborate to complete all subtasks. Both share a common action space: $\mathcal{A}^A = \mathcal{A}^H = \{\text{Up}, \text{Down}, \text{Left}, \text{Right}\}$. Movement is stochastic: with a probability of 0.95, they move to the intended direction, and with a probability of 0.025, they move to one of the two perpendicular directions. If a move leads to an obstacle, the agent or human remains in the current state. Each step incurs a penalty of -2, while completing a subtask (entering an orange cell) yields a reward of +100, encouraging efficient collaboration and task completion in the stochastic environment.

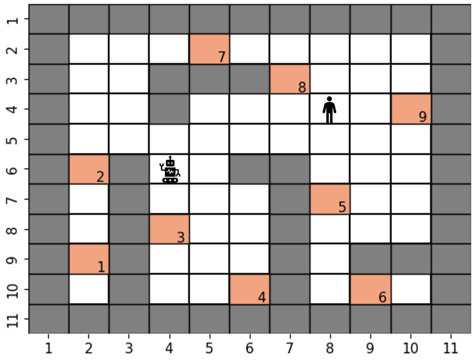
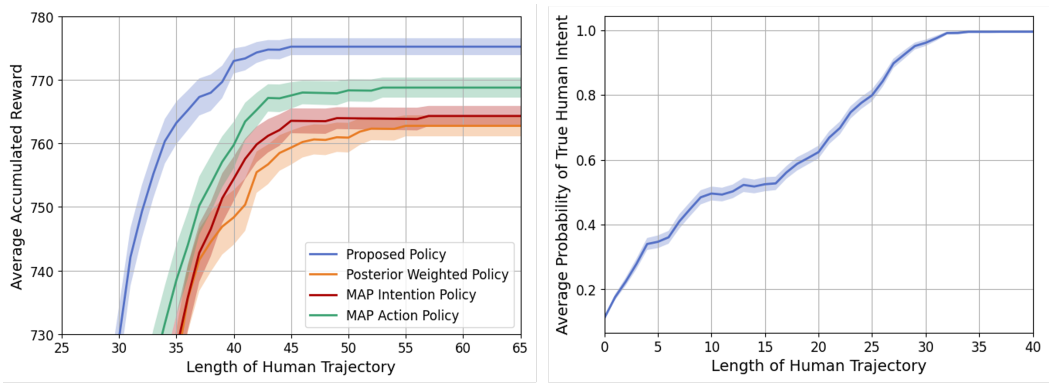


Figure 2: 2D grid-world environment for human-AI collaboration, with 9 subtasks (orange terminal states) and gray cells as obstacles.

In our experiment, the proposed Bayesian policy's double DQN model uses both policy and target networks, $Q_{\mathbf{w}}$ and $Q_{\mathbf{w}^-}$, each with three hidden layers containing 1024 neurons per layer. The belief vector, which serves as the input to these networks, has a size of 22. It includes 4 bits for the agent and human coordinates, along with 18 digits representing the subtask tracker and intention probabilities. Key parameters for training include a learning rate of $\beta = 10^{-5}$, a replay memory size of $|\mathcal{D}| = 50000$, a minibatch size of $N_{\text{batch}} = 256$, a discount factor $\gamma = 0.99$, and an exploration rate that decays from $\epsilon = 0.99$ to $\epsilon = 0.1$. The target network $Q_{\mathbf{w}^-}$ is updated every 4 steps, with a soft update parameter of $\tau = 10^{-3}$ for synchronizing the policy network $Q_{\mathbf{w}}$. Human parameters are set with low-level rationality $q = 0.9$ and high-level rationality $\alpha = 0.5$. After the training is finalized, all experiments for evaluation are conducted using the pre-trained Q-network. The grid-world environment provides a controlled testbed for evaluating the proposed approach, capturing essential aspects of human-AI collaboration, including stochastic transitions, cooperative task execution, and uncertain human behavior. While simplified, this environment establishes a foundation for extending our method to more complex, real-world settings.

Existing approaches to human-AI collaboration focus on identifying human intent (either probabilistically or deterministically) and acting to complement it for coordination (Hoffman et al., 2024; Ni et al., 2023; Jain & Argall, 2019). We compare our proposed method with three baseline approaches: (1) the maximum a posteriori (MAP) intention policy, which treats the best estimate of human intent as ground truth and enables the agent to perform optimally based on this assumed intent (Jain & Argall, 2019; Hoffman et al., 2024); (2) the posterior-weighted policy, which estimates the posterior distribution of human intent and assigns weights to each possible subtask, allowing the agent to act in a way complementary to the human's likely actions, akin to the QMDP approach in partially observable environments (Littman et al., 1995; Karkus et al., 2017); and (3) the MAP action policy, commonly used in robust planning, where the AI agent selects the most probable action based on the posterior of human intent (Kiran et al., 2021; Zhang et al., 2024).

In the first experiment, the starting points of the human and the agent are randomly selected, and results are averaged over 200 independent trials. Figure 3(a) shows the average accumulated reward across different methods as a function of the human trajectory length, where the shaded area represents the standard error of the mean. The proposed policy significantly outperforms all baseline approaches, achieving higher rewards due to its ability to effectively integrate human intent uncertainty into decision-making. Among the baselines, the MAP action policy achieves the second-best performance, but it still lags behind the proposed policy. This is attributed to the proposed policy's dynamic use of belief distributions for long-term planning, while traditional methods rely on fixed, less flexible logic. Figure 3(b) illustrates the evolution of the average probability of true human intent, starting from a uniform prior and converging to 1 as the trajectory length increases. Initially, the inferred human intent is uncertain, but as more subtasks are completed, the probability improves due to the probabilistic inference framework. This steady convergence reflects the accuracy of the intention estimation approach and highlights its role in achieving the high rewards observed in Figure 3(a).



(a) (b)

Figure 3: Average accumulated reward across different methods and true inferred human intent probability.

Figures 4 illustrate the impact of human rationality on the average accumulated rewards across different policies. As human behavior becomes less rational at both low and high levels, the complexity of human-AI collaboration increases, resulting in lower overall rewards. Despite these challenges,

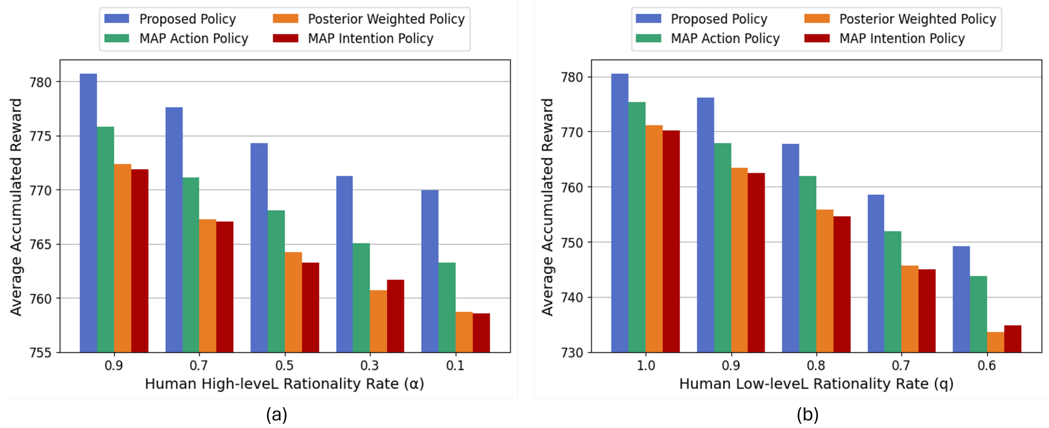


Figure 4: Impact of the high-level and low-level human rationality on the performance of different methods.

the proposed policy consistently outperforms all baseline policies in all scenarios. Among the baselines, the MAP action policy exhibits better performance, while the MAP intention policy and the posterior-weighted policy show similar and relatively poor performance.

More specifically, Figure 4(a) shows the effect of high-level human rationality ($\alpha$) on accumulated rewards. Lower $\alpha$ values model humans with less rational decision-making at the subtask level, resulting in more challenging collaboration dynamics. Notably, the gap between the proposed policy and the baselines widens under these conditions, highlighting the ability of our approach to effectively account for uncertainty. Additionally, the influence of $\alpha$ on the proposed policy's performance diminishes as rationality decreases further, underscoring its robustness in handling complex and unpredictable human behavior.

Figure 4(b) examines the impact of human rationality ($q$) at action level, representing the consistency of human actions toward performing a subtask. A fully rational human ($q = 1$) follows an optimal policy, while a less rational human ($q = 0.6$) exhibits highly stochastic behavior. Lower values of $q$ ($q < 0.6$) are not considered, as these situations effectively reduce collaboration to the AI agent completing all subtasks independently. Since $q$ directly influences every human action, its impact on collaboration is more pronounced than that of $\alpha$. As $q$ decreases, all policies experience reduced performance due to increased randomness in human actions, which complicates intent inference and lengthens task completion trajectories. Despite this, the proposed policy remains more effective than the baselines, demonstrating its ability to maintain superior collaboration even under significant uncertainty in human behavior.

Table 1: Impact of state transition stochasticity on the performance of different methods.

| Transition stochasticity $\zeta$ | 0.95 | 0.85 | 0.75 | 0.65 | 0.55 |
|---|---|---|---|---|---|
| Proposed | 775.21 | 754.34 | 728.36 | 694.51 | 639.03 |
| MAP action | 768.54 | 747.65 | 715.99 | 680.25 | 629.11 |
| MAP intention | 763.13 | 742.52 | 709.78 | 668.64 | 613.21 |
| Posterior-weighted | 764.52 | 742.72 | 711.67 | 676.68 | 609.83 |

This section analyzes the impact of stochasticity in state transitions on the performance of the proposed method, which is characterized by the parameter $\zeta \in [0, 1]$. Larger values of $\zeta$ represent more deterministic movements, and smaller values indicate increased uncertainty in transitions. Table 1 presents the average rewards achieved by various policies under five different levels of stochasticity. As stochasticity increases, the performance of all policies declines, highlighting the challenges posed by uncertainty in accurately capturing human intent and enabling effective human-agent collaboration. Despite the significant performance reduction of the proposed method under higher stochasticity, it consistently achieves the best results compared to all competing methods in every scenario. This superior performance stems from effectively accounting for uncertainty in the long-term decision-making process.

Table 2: Impact of the explicit feedback, implicit inference and lack of information on the performance of different methods.

| | Direct Feedback<br>(True Intent) | Implicit Inference<br>(Estimated Intent) | No Information<br>(Uniform Distribution) |
|---|---|---|---|
| Proposed | 776.07 | 775.21 | 773.42 |
| MAP action | 768.98 | 768.54 | 763.38 |
| MAP intention | 768.74 | 763.13 | 754.47 |
| Posterior-weighted | 769.21 | 764.52 | 754.79 |

Finally, beyond performance analyses through implicit learning, we investigate two extreme scenarios: (1) constant direct feedback is provided by the human, causing the intent probability to peak sharply over the true human intent; and (2) there is an absence of implicit data or feedback, where the agent has no knowledge of its human teammate's intentions, represented by a uniform distribution throughout the process. Table 2 shows the average rewards achieved by various methods under these conditions. As expected, the best results are obtained when continuous human feedback is available, performing similarly to the implicit inference scenario. This demonstrates that implicit inference of human intentions can be highly effective in scenarios where explicit feedback is unavailable or nonexistent. Conversely, the absence of both implicit inference and explicit feedback results in the worst performance. The proposed policy achieves the best results across all scenarios. This success is attributed to training over the belief space, where the policy is optimized across a wide spectrum of information about human intent, from full knowledge in explicit feedback scenarios to incomplete knowledge in implicit inference and even to no knowledge in cases of absent data or interactions.

## 7 Conclusion

This paper presented a Bayesian reinforcement learning framework to enhance human-AI collaboration by probabilistically modeling human intent and incorporating this information into adaptive decision-making. By defining a belief state that integrates both low- and high-level representations of human behavior and demonstrating its Markov property, we derived an optimal Bayesian policy. Using a pre-trained deep reinforcement learning model, the framework enables efficient real-time decision-making without explicit human interaction. Experimental results validated its robustness and adaptability, highlighting its potential for scalable and reliable human-AI collaboration. Future work will explore its application to larger, unstructured domains and multi-agent settings, as well as the integration of implicit and explicit interactions to maximize collaboration efficiency.

## References

Saminda Wishwajith Abeyruwan, Laura Graesser, David B D'Ambrosio, Avi Singh, Anish Shankar, Alex Bewley, Deepali Jain, Krzysztof Marcin Choromanski, and Pannag R Sanketi. i-Sim2Real: Reinforcement learning of robotic policies in tight human-robot interaction loops. In *Conference on Robot Learning*, pp. 212–224. PMLR, 2023.

Arsha Ali, Hebert Azevedo-Sa, Dawn M Tilbury, and Lionel P Robert Jr. Heterogeneous human–robot task allocation based on artificial trust. *Scientific reports*, 12(1):15304, 2022.

Sushilkumar Ambhore. A comprehensive study on robot learning from demonstration. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 291–299. IEEE, 2020.

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*, 2017.

Sophie Berretta, Alina Tausch, Greta Ontrup, Björn Gilles, Corinna Peifer, and Annette Kluge. Defining human-AI teaming the human-centered way: a scoping review and network analysis. *Frontiers in Artificial Intelligence*, 6:1250725, 2023.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Konstantinos Charalampous, Ioannis Kostavelis, and Antonios Gasteratos. Recent trends in social aware robot navigation: A survey. *Robotics and Autonomous Systems*, 93:85–104, 2017.

Jiayu Chen, Tian Lan, and Vaneet Aggarwal. Hierarchical adversarial inverse reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Allyson I Hauptman, Beau G Schelble, Nathan J McNeese, and Kapil Chalil Madathil. Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior*, 138:107451, 2023.

Mark K Ho and Thomas L Griffiths. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):33–53, 2022.

Guy Hoffman, Tapomayukh Bhattacharjee, and Stefanos Nikolaidis. Inferring human intent and predicting human action in human–robot collaboration. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2024.

Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33:15931–15941, 2020.

Siddarth Jain and Brenna Argall. Probabilistic human intent recognition for shared autonomy in assistive robotics. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(1):1–23, 2019.

Tharindu Kaluarachchi, Andrew Reis, and Suranga Nanayakkara. A review of recent deep learning approaches in human-centered machine learning. *Sensors*, 21(7):2514, 2021.

Peter Karkus, David Hsu, and Wee Sun Lee. QMDP-Net: Deep learning for planning under partial observability. *Advances in neural information processing systems*, 30, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Meng-Lun Lee, Sara Behdad, Xiao Liang, and Minghui Zheng. Task allocation and planning for product disassembly with human–robot collaboration. *Robotics and Computer-Integrated Manufacturing*, 76:102306, 2022.

Michael L Littman, Anthony R Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*, pp. 362–370. Elsevier, 1995.

Yang Lu. Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics*, 6(1):1–29, 2019.

Tien Mai, Thanh Nguyen, et al. Inverse factorized soft Q-learning for cooperative multi-agent imitation learning. *Advances in Neural Information Processing Systems*, 37:27178–27206, 2025.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6292–6299. IEEE, 2018.

Payam Nasernejad, Tarek Sayed, and Rushdi Alsaleh. Modeling pedestrian behavior in pedestrian-vehicle near misses: A continuous gaussian process inverse reinforcement learning (GP-IRL) approach. *Accident Analysis & Prevention*, 161:106355, 2021.

Shouxiang Ni, Lindong Zhao, Ang Li, Dan Wu, and Liang Zhou. Cross-view human intention recognition for human-robot collaboration. *IEEE Wireless Communications*, 30(3):189–195, 2023.

Alexander Obaigbena, Oluwaseun Augustine Lottu, Ejike David Ugwuanyi, Boma Sonimitiem Jacks, Enoch Oluwademilade Sodiya, and Obinna Donald Daraojimba. AI and human-robot interaction: A review of recent advances and challenges. *GSC Advanced Research and Reviews*, 18(2):321–330, 2024.

Carl Orge Retzlaff, Srijita Das, Christabel Wayllace, Payam Mousavi, Mohammad Afshari, Tianpei Yang, Anna Saranti, Alessa Angerschmid, Matthew E Taylor, and Andreas Holzinger. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *Journal of Artificial Intelligence Research*, 79:359–415, 2024.

Ronal Singh, Tim Miller, Joshua Newn, Eduardo Velloso, Frank Vetere, and Liz Sonenberg. Combining gaze and AI planning for online human intention recognition. *Artificial Intelligence*, 284: 103275, 2020.

Liting Sun, Wei Zhan, and Masayoshi Tomizuka. Probabilistic prediction of interactive driving behavior via hierarchical inverse reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2111–2117. IEEE, 2018.

Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6):3692–3711, 2023.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 2024.

Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–25, 2021.

Zuyuan Zhang, Hanhan Zhou, Mahdi Imani, Taeyoung Lee, and Tian Lan. Collaborative AI teaming in unknown environments via active goal deduction. *arXiv preprint arXiv:2403.15341*, 2024.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.