2024 IEEE Conference on Control Technology and Applications (CCTA)
August 21-23, 2024. Newcastle upon Tyne, UK

High-Level Human Intention Learning for Cooperative Decision-Making

Yuxin Lin, Seyede Fatemeh Ghoreishi, Tian Lan, and Mahdi Imani

2024 IEEE Conference on Control Technology and Applications (CCTA) | 979-8-3503-7094-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCTA60707.2024.10666622

Abstract— **Autonomous agents are increasingly popular in various practical domains, assisting humans in performing complex tasks. However, coordinating between agents and humans can be challenging, particularly when communication is limited or non-existent. This paper proposes a method for cooperative decision-making by enabling autonomous agents to infer high-level human intention through their behavioral data. Human is modeled as a sub-optimal reinforcement learning agent. A statistical learning method is developed for implicit probabilistic reasoning of human intentions by accounting for complex and unpredictable human behavior. The proposed method computes the exact likelihood and posterior of human intentions. The method is fully recursive and accounts for human priorities, making it applicable to various domains. The agents' decision-making is achieved using a combination of the active learning approach and quantified human intentions, which enables effective coordination of tasks and prevents duplication of efforts. The proposed method allows agents to adapt their strategy in real time based on partial knowledge of human intentions. Our numerical experiments demonstrate the efficacy of our proposed method in intention inference and task coordination.**

I. Introduction

Autonomous agents are increasingly prevalent in various practical domains. Solving most complex tasks requires close collaboration between humans and artificial intelligence (AI) agents. However, coordinating between agents and humans can be challenging, especially when communication between them is limited or non-existent [1]–[9]. For instance, consider a rescue team comprising humans and autonomous agents during or after a natural disaster, where a team of humans and AI agents are deployed to search for victims, evaluate areas, and provide medical aid. Human behavior in these tasks is often too complex to model or predict and is often influenced by several external factors. For example, humans may change their plans during operations if they perceive the task as dangerous or beyond their capabilities.

Several approaches have been developed for coordination between AI agents and humans [10]–[14], including imitation learning and inverse reinforcement learning (IRL) [15]–[19]. These techniques assume that human policy and reward are fixed and aim to learn them from a large number of available human demonstrations. Intention recognition algorithms have also been developed to learn the patterns of human behavior and infer the high-level human goals (i.e., intentions) from these patterns [20]–[24]. These methods, however, rely on the availability of low-level human actions and a large amount of human data. In practice, access to human actions requires direct communication, which is often impossible or becomes overwhelming for humans. Additionally, human intentions can change rapidly, making it difficult for agents to capture and adapt to changes. Furthermore, probabilistic methods have been developed for inferring human intention in robot manipulation tasks using human state or state-action sequences [25]–[28]. However, these methods do not rely on a human model but rather approximate the likelihoods with distance-based measures, limiting their applicability to domains where distance is not measurable or where human priorities may vary.

This paper presents a statistical learning approach to infer high-level human intentions through limited observed human data without direct communication. These high-level intentions are subtasks that agents or humans might perform at different times. To address the complexity and uncertainty in human behavior, humans are modeled as sub-optimal reinforcement learning agents. This model accounts for the potential priorities of humans in performing subtasks using the reward function. Given the availability of a human state sequence, this paper computes an exact probabilistic model of possible human intentions. A recursive and efficient method is introduced for updating the posterior of human intentions as new information from humans becomes available. Using the quantified human intention posterior, an active learning approach is developed to ensure the AI agent effectively cooperates with the human without duplicating efforts. We demonstrate that the active learning method becomes the exact optimal cooperative policy if the posterior distribution of human intentions peaks over a single subtask. Through numerical experiments, we have showcased the performance of the proposed policy in terms of learning human intents and enhancing the performance of human-AI collaboration.

II. Background - A Markov Decision Process

A Markov decision process (MDP) representing human and agent collaboration can be defined by a tuple $\langle \mathcal{S}^A, \mathcal{S}^H, \mathcal{A}^A, \mathcal{A}^H, \mathcal{P}, R \rangle$, where $\mathcal{S}^A$ and $\mathcal{S}^H$ are the agent and human state spaces; $\mathcal{A}^A$ and $\mathcal{A}^H$ are the agent and human action spaces; $\mathcal{P}$ is the state transition probability function where $\mathcal{P}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ represent the probability that the next agent or human state is $\mathbf{s}'$ if action $\mathbf{a}$ is taken at state $\mathbf{s}$; and R is a cooperative reward function with a real value outcome such that $R(\mathbf{s}^A, \mathbf{s}^H, \mathbf{a}^A, \mathbf{a}^H, \mathbf{s}^{A'}, \mathbf{s}^{H'})$ encode the reward earned when actions $(\mathbf{a}^H, \mathbf{a}^A)$ are taken

Yuxin Lin and Mahdi Imani are with the Department of Electrical and Computer Engineering at Northeastern University, S.F. Ghoreishi is with the Department of Civil and Environmental Engineering and Khoury College of Computer Sciences at Northeastern University, and T. Lan is with the Department of Electrical and Computer Engineering at George Washington University. Emails: lin.yuxi@northeastern.edu, f.ghoreishi@northeastern.edu, tlan@gwu.edu, m.imani@northeastern.edu.

in state $(\mathbf{s}^A, \mathbf{s}^H)$ and the human and agent states move to $(\mathbf{s}^{A'}, \mathbf{s}^{H'})$. The reward function depends on the actions and states of the human and agent, which models a cooperative setting where the human and agent should work together to ensure the overall coordination of tasks.

III. Probabilistic Reasoning of Human Intention

A. High-Level Human Intention

This paper focuses on domains in which a human and an AI agent work alongside each other without direct communication. In practice, human behavior can be too complex to model, and a full understanding of human intentions can be challenging or impossible. Thus, we aim to enable the AI agent to probabilistically reason about human intentions based on human data.

Let $T^1, T^2, ..., T^N$ be N subtasks that should be performed by the human or AI agent in a cooperative setting. We refer to these subtasks as high-level intentions, which the human and agent might undertake at different times. For example, consider a rescue operation in which humans and robots work together to achieve multiple subtasks, such as searching for victims in specific areas, providing medical aid to injured people, evacuating unsafe areas, and so on. Depending on the subtask being performed by the human, the agent should undertake other subtasks that prevent the duplication of efforts. For instance, if a human decides to perform an evacuation instead of providing medical aid, the agent should capture this and adapt itself to ensure the overall success of the rescue operation.

Let $p_0 = [p(T^1), ..., p(T^N)]$ be the prior probability for human intention, which represents the initial probability that the human might perform different subtasks. Given the sequence of human observed states $\mathbf{s}^H_{0:k} = (\mathbf{s}^H_0, ..., \mathbf{s}^H_k)$, the agent's understanding of the human's intention can be expressed using the following posterior distribution:

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

Note that the human actions that have led to the human state sequence $\mathbf{s}^H_{0:k}$ are often unknown to agents once no communication exists between them. This is in contrast with most inverse reinforcement learning and imitation learning approaches [15]–[19], which require human state-action pairs to quantify human policy or reward function.

To keep track of human intentions, a recursive computation of the posterior distribution of human intention is required. Given that $\mathbf{s}^H_{k+1}$ is the new observed human state at time step $k+1$, the new posterior distribution of human intention can be expressed as:

$$\begin{aligned} p_{k+1}(j) &= P(T^* = T^j \mid \mathbf{s}^H_{0:k+1}) \\ &= \frac{P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, T^j)\, P(T^j \mid \mathbf{s}^H_{0:k})}{\sum_{l=1}^N P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, T^l)\, P(T^l \mid \mathbf{s}^H_{0:k})} \\ &= \frac{P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, T^j)\, p_k(j)}{\sum_{l=1}^N P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, T^l)\, p_k(l)}, \end{aligned} \tag{2}$$

for $j = 1, ..., N$. In the last line of (2), the first term in the numerator can be further simplified as:

$$\begin{aligned} &P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, T^j) \\ &= \sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1}, \mathbf{a}^H_k = \mathbf{a}^H \mid \mathbf{s}^H_{0:k}, T^j) \\ &= \sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, \mathbf{a}^H_k = \mathbf{a}^H, T^j) P(\mathbf{a}^H_k = \mathbf{a}^H \mid \mathbf{s}^H_{0:k}, T^j) \\ &= \sum_{\mathbf{a}^H \in \mathcal{A}^H} \underbrace{P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_k, \mathbf{a}^H_k = \mathbf{a}^H)}_{\text{Human Transition}} \underbrace{P(\mathbf{a}^H_k = \mathbf{a}^H \mid \mathbf{s}^H_{0:k}, T^j)}_{\text{Human Intention}}, \end{aligned} \tag{3}$$

where the human transition term depends on the stochasticity in the environment and the human intention term quantifies the probability that the human takes action $\mathbf{a}^H \in \mathcal{A}^H$ given the sequence of states $\mathbf{s}^H_{0:k}$ and a known intention T^j. One can see the likelihood function in (3) as the weighted sum of the human transition terms, where weights representing human intentions specify the extent to which each action could be taken by the human.

In domains with multiple subtasks, a binary auxiliary vector can be used to keep track of performed and unperformed subtasks. The status of human and AI agent states and subtasks at time step k can be represented using a vector $[\mathbf{s}^A_k, \mathbf{s}^H_k, \eta_k]$, where $\eta_k \in \{0,1\}^N$ expresses the status of subtasks with the value $\eta_k(j) = 1$ and $\eta_k(j) = 0$ corresponding to cases where the jth subtask is performed and not performed, respectively. The element of the subtask tracker for any subtask that is performed (terminated) turns to 1 and stays 1, while for unperformed subtasks stays 0. Thus, one can represent the deterministic transition of the subtask tracker when the state moves from $(\mathbf{s}^A_k, \mathbf{s}^H_k, \eta_k)$ to $(\mathbf{s}^A_{k+1}, \mathbf{s}^H_{k+1})$ as:

$$\eta_{k+1}(j) = \begin{cases} 1 & \text{if } \mathbf{s}^A_{k+1} \text{ or } \mathbf{s}^H_{k+1} = \mathcal{G}^j \\ \eta_k(j) & \text{otherwise} \end{cases}, \text{ for } j = 1, ..., N, \tag{4}$$

where $\mathcal{G}^j$ is the terminal state for the jth subtask. Note that $\eta_k(j) = 0$ implies that the jth subtask is not performed by human and agent up to time step k, represented by $\mathcal{G}^j \notin \mathbf{s}^A_{0:k}$ and $\mathcal{G}^j \notin \mathbf{s}^H_{0:k}$. Meanwhile, as a performed subtask does not require the human or agent to handle it a second time, the posterior of the human intention towards subtask T^j becomes 0 as soon as it is terminated, which can be expressed by:

$$p_{k+1}(j) \propto \begin{cases} 0 & \text{if } \mathbf{s}^A_{k+1} \text{ or } \mathbf{s}^H_{k+1} = \mathcal{G}^j \\ p_k(j) & \text{otherwise} \end{cases}, \text{ for } j = 1, ..., N, \tag{5}$$

where the elements of p_{k+1} is sum to 1.

The cooperative reward function can be expressed individually for the human and AI agent using the subtask tracker as: $R(\mathbf{s}^A, \eta, \mathbf{a}^A, \mathbf{s}^{A'}, \eta')$ and $R(\mathbf{s}^H, \eta, \mathbf{a}^H, \mathbf{s}^{H'}, \eta')$. The auxiliary variables ensure that the performed subtasks by either humans or agents and denoted with non-zero η values do not contribute multiple times to the accumulated cooperative rewards.

B. Quantification of Human Intention

The human objective for taking the subtask T^j can be expressed through the subtask tracker with all elements 1 except for the jth element set to 0, denoted by $\mathbf{e}^j$. The human reward function $R(\mathbf{s}^H, \eta = \mathbf{e}^j, \mathbf{a}^H, \mathbf{s}^{H'}, \eta')$ is the reward value gained by the human if ends up in $(\mathbf{s}^{H'}, \eta')$ after taking action $\mathbf{a}^H$ in state $(\mathbf{s}^H, \eta = \mathbf{e}^j)$. This reward is only defined for the subtask T^j, as the current subtask tracker is set as $\mathbf{e}^j$ (i.e., T^j is the only remaining subtask).

We define a policy $\pi^H : \mathcal{S}^H \to \mathcal{A}^H$, which associates an action to each human state. The optimal policy for human undertaking subtask T^j without considering the agent can be expressed as:

$$\pi^{*,H}_{T^j}(\mathbf{s}^H, \eta = \mathbf{e}^j) = \underset{\pi^H}{\operatorname{argmax}}\, \mathbb{E}\Bigg[\sum_{t=0}^{h} \gamma^t R(\mathbf{s}^H_t, \eta_t, \mathbf{a}^H_t, \mathbf{s}^H_{t+1}, \eta_{t+1}) \mid \mathbf{s}^H_0 = \mathbf{s}^H, \eta_0 = \mathbf{e}^j, \mathbf{a}^H_{0:h} \sim \pi^H \Bigg], \tag{6}$$

for all $\mathbf{s}^H \in \mathcal{S}^H$; where the maximization is over all deterministic human policies, γ is a discount factor, and $\eta_0 = \mathbf{e}^j$ ensures that the human will receive a positive reward only if it achieves the subtask T^j. In the finite-horizon case, the discount factor is typically set to 1, whereas in the infinite horizon case ($h = \infty$), the discount factor $\gamma \in [0, 1)$ is included to obtain a finite sum. The optimal solution for maximization in (6) can be achieved using dynamic programming approaches for finite state and action spaces and reinforcement learning techniques for large and complex state and action spaces.

We model human behavior as a sub-optimal reinforcement learning (RL) agent, whose imperfectness (i.e., deviation from the optimal policy) is modeled through the well-known ϵ-greedy or Boltzmann policy [29]. Let $\pi^{*,H}_{T^j}$ be the optimal human policy for subtask T^j obtained in (6). We model the human decisions according to the following ϵ-greedy policy [15]:

$$\pi^H_{T^j}(\mathbf{a}^H|\mathbf{s}^H) := P(\mathbf{a}^H \mid \mathbf{s}^H, T^j) = \begin{cases} q + \frac{1-q}{|\mathcal{A}^H|} & \text{If } \mathbf{a}^H = \pi^{*,H}_{T^j}(\mathbf{s}^H, \mathbf{e}^j) \\ \frac{1-q}{|\mathcal{A}^H|} & \text{If } \mathbf{a}^H \neq \pi^{*,H}_{T^j}(\mathbf{s}^H, \mathbf{e}^j) \end{cases}, \tag{7}$$

for $\mathbf{a}^H \in \mathcal{A}^H, \mathbf{s}^H \in \mathcal{S}^H$, where $0 \leq q \leq 1$ represents the confidence of the human, and $\pi^{*,H}_{T^j}$ is computed in (6). Values of q close to 1 represent confident (i.e., rational) humans, whereas the values close to 0 model humans with random decision-making behavior.

The Boltzmann policy representing the human policy can be expressed as:

$$\pi^H_{T^j}(\mathbf{a}^H|\mathbf{s}^H) := P(\mathbf{a}^H \mid \mathbf{s}^H, T^j) = \frac{\exp\left(\beta\, q^{*,H}_{T^j}(\mathbf{s}^H, \mathbf{e}^j, \mathbf{a}^H)\right)}{\sum_{\mathbf{a}' \in \mathcal{A}^H} \exp\left(\beta\, q^{*,H}_{T^j}(\mathbf{s}^H, \mathbf{e}^j, \mathbf{a}')\right)}, \tag{8}$$

for $\mathbf{a}^H \in \mathcal{A}^H, \mathbf{s}^H \in \mathcal{S}^H$; where

$$q^{*,H}_{T^j}(\mathbf{s}^H, \eta = \mathbf{e}^j, \mathbf{a}^H) = \mathbb{E}\Bigg[\sum_{t=0}^{h} \gamma^t R(\mathbf{s}^H_t, \eta_t, \mathbf{a}^H_t, \mathbf{s}^H_{t+1}, \eta_{t+1}) \mid \mathbf{s}^H_0 = \mathbf{s}^H, \eta_0 = \mathbf{e}^j, \mathbf{a}^H_0 = \mathbf{a}^H, \mathbf{a}^H_{1:h} \sim \pi^{*,H}_{T^j} \Bigg], \tag{9}$$

is the state-action value function under the optimal policy for subtask T^j, and $\beta > 0$ represents the confidence (rationality level) of the human. Large values of β model confident humans, whereas smaller values model humans close to random decision-makers. It should be noted that both human models in (7) and (8) are widely used in the inverse reinforcement learning context for modeling the imperfect behavior of humans [15].

Considering that our problem is within a MDP framework, we believe that human behavior is determined solely by their current state and intent, rather than by previous trajectories. Therefore, according to (7) and (8), the human intention term in (3) can be expressed as:

$$P(\mathbf{a}^H_k = \mathbf{a}^H \mid \mathbf{s}^H_{0:k}, T^j) = P(\mathbf{a}^H_k = \mathbf{a}^H \mid \mathbf{s}^H_k, T^j) = \pi^H_{T^j}(\mathbf{a}^H|\mathbf{s}^H_k). \tag{10}$$

If a subtask is performed by the human or agent at the current time, we assume that the information is shared with them, and the human will not follow that subtask in the next steps. Therefore, the intention probability for the performed subtask should be set as zero. Replacing (10) into (2) and (3) leads to the following recursive posterior update of intentions:

$$p_{k+1}(j) = \frac{\sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_k, \mathbf{a}^H_k = \mathbf{a}^H) \pi^H_{T^j}(\mathbf{a}^H|\mathbf{s}^H_k) 1_{\eta_{k+1}(j)=0} p_k(j)}{\sum_{l=1}^{N} \sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_k, \mathbf{a}^H_k = \mathbf{a}^H) \pi^H_{T^l}(\mathbf{a}^H|\mathbf{s}^H_k) 1_{\eta_{k+1}(l)=0} p_k(l)}, \tag{11}$$

for $j = 1, ..., N$; where the indicator $1_{\eta_{k+1}(j)=0}$ returns 1 if the jth subtask has not been performed (i.e., $\eta_{k+1}(j) = 0$), and 0 otherwise.

IV. Incorporating Human Intentions into Agent Policy

A. Agent Optimal Policy under Known Human Intention

According to the intention formulation in the previous section, a known human intention at time step k can be expressed using p_k with a single element corresponding to the true human intention 1 and others 0. This represents the case where the agent is fully aware of the human intention. Given that human is performing subtask T^j (i.e., $p_k(j) = 1$), we define a deterministic policy $\pi^A_{H=T^j} : \mathcal{S}^A \times \{0, 1\}^{N-1} \to \mathcal{A}^A$ as a mapping from the agent state and subtask tracker to agent action space. The optimal policy for the agent under human intention T^j depends on the unperformed subtasks, denoted by the subtask tracker η. Therefore, the current state of agent $\mathbf{s}^A$ and the status of subtasks η can be used for defining the optimal policy for the agent under a known

human intention T^j as:

$$\pi^{*,A}_{H=T^j}(\mathbf{s}^A,\eta)=\underset{\pi^A}{\operatorname{argmax}}\,\mathbb{E}\Bigg[\sum_{t=0}^{h}\gamma^t R(\mathbf{s}^A_t,\eta_t,\mathbf{a}^A_t,\mathbf{s}^A_{t+1},\eta_{t+1})\mid \mathbf{s}^A_0=\mathbf{s}^A,\eta_0=\eta,\eta_0(j)=1,\mathbf{a}^A_{0:h}\sim\pi^A\Bigg], \tag{12}$$

for all $\mathbf{s}^A\in\mathcal{S}^A$ and $\eta\in\{0,1\}^{N-1}$; where $\pi^{*,A}_{H=T^j}(\mathbf{s}^A,\eta)$ is the optimal agent policy at state $(\mathbf{s}^A,\eta)$ given that the human is performing the subtask T^j, and the expectation is with respect to the stochasticity in the agent transition probability. Note that the expectation is conditioned on $\eta(j)=1$, which ensures the agent does not have the incentive to duplicate subtask T^j that the human is performing, rather focusing on unperformed subtasks indicated by zero elements of η.

B. Agent Policy under Unknown Human Intention

While the solution for optimization in (12) provides the optimal agent policy when the agent is fully aware of the true human intention, in practice, under no direct communication, the agent may not have complete knowledge of the human intention. The proposed approach described in the previous section allows for keeping the probabilistic knowledge of human intention over time. In the following, we describe how this probabilistic knowledge can be incorporated into agent policy.

We start with the *active learning* policy, which, along with its variations, is a widely used class of techniques for decision-making under uncertainty [30], [31]. We define the optimal state-action value function under the agent policy $\pi^{*,A}_{H=T^j}$ defined in (12), when the human is performing the subtask T^j, as:

$$q^{*,A}_{H=T^j}(\mathbf{s}^A,\eta,\mathbf{a}^A)=\mathbb{E}\Bigg[\sum_{t=0}^{h}\gamma^t R(\mathbf{s}^A_t,\eta_t,\mathbf{a}^A_t,\mathbf{s}^A_{t+1},\eta_{t+1})\mid \mathbf{s}^A_0=\mathbf{s}^A,\eta_0=\eta,\eta_0(j)=1,\mathbf{a}^A_0=\mathbf{a}^A,\mathbf{a}^A_{1:h}\sim\pi^{*,A}_{H=T^j}\Bigg], \tag{13}$$

where $q^{*,A}_{H=T^j}(\mathbf{s}^A,\eta,\mathbf{a}^A)$ is the expected return if the agent takes action $\mathbf{a}^A$ at state $(\mathbf{s}^A,\eta)$, then follows policy $\pi^{*,A}_{H=T^j}$. Depending on the size of the state space, the Q-values can be obtained offline by employing N parallel dynamic programming or reinforcement learning algorithms, each tuned to a specific subtask $T\in\{T^1,...,T^N\}$ that the human is undertaking. Let $\mathbf{s}^A_k$ be the current agent state, η_k be the current status of performed subtasks, and p_k be the posterior distribution of the human intention given the information up to time step k. The agent can take action at time step k according to the following:

$$\begin{aligned}\mathbf{a}^A_k&=\underset{\mathbf{a}^A\in\mathcal{A}^A}{\operatorname{argmax}}\,\mathbb{E}[q^{*,A}_{H=T}(\mathbf{s}^A_k,\eta_k,\mathbf{a}^A)\mid\mathbf{s}^H_{0:k}]\\&=\underset{\mathbf{a}^A\in\mathcal{A}^A}{\operatorname{argmax}}\,\mathbb{E}_{p_k}[q^{*,A}_{H=T}(\mathbf{s}^A_k,\eta_k,\mathbf{a}^A)]\\&=\underset{\mathbf{a}^A\in\mathcal{A}^A}{\operatorname{argmax}}\sum_{j=1}^{N}p_k(j)\,q^{*,A}_{H=T^j}(\mathbf{s}^A_k,\eta_k,\mathbf{a}^A),\end{aligned} \tag{14}$$

where the expectation is with respect to the posterior distribution of human intention. If the uncertainty in human intention approaches zero, then the active learning policy in (14) becomes the optimal policy as it yields the exact Bellman optimality. A large uncertainty in human intention corresponds to the case where the Q-values associated with different human subtasks contribute to the active learning policy. Note that the quantified human intention can be incorporated into other policies, such as those relying on a single most probable human intention [32], [33] or those relying on samples of possible human intentions (e.g., rollout) [34]–[36].

The detailed steps of the proposed method are presented in Algorithm 1. The process of the proposed method in (14) consists of offline and online steps. In the offline step, two parallel sets of dynamic programming or reinforcement learning approaches should be employed; one set learns N human policies defined over the human state space, each corresponding to a single human subtask; the second set learns N agent policies defined over the agent state space and subtask tracker space, each corresponding to a fixed human intention. The computed policy can then be used for the recursive update of human intention as well as adaptive decision-making for the agent.

V. Discussions and Analyses

To investigate the impact of human confidence (or rationality level) on the posterior of human intention in (11), we represent the posterior distribution of human intentions using the prior intention probabilities and the likelihood of observed human data as:

$$p_{k+1}(j)=\frac{P(\mathbf{s}^H_{0:k+1}\mid T^j)\,p_0(j)}{\sum_{l=1}^{N}P(\mathbf{s}^H_{0:k+1}\mid T^l)\,p_0(l)}, \tag{15}$$

where

$$\begin{aligned}P(\mathbf{s}^H_{0:k+1}\mid T^j)&=\prod_{r=0}^{k}P(\mathbf{s}^H_{r+1}\mid\mathbf{s}^H_{0:r},T^j)\\&=\prod_{r=0}^{k}\sum_{\mathbf{a}^H\in\mathcal{A}^H}P(\mathbf{s}^H_{r+1},\mathbf{a}^H_r=\mathbf{a}^H\mid\mathbf{s}^H_{0:r},T^j)\\&=\prod_{r=0}^{k}\Bigg[\sum_{\mathbf{a}^H\in\mathcal{A}^H}P(\mathbf{s}^H_{r+1}\mid\mathbf{s}^H_r,\mathbf{a}^H_r=\mathbf{a}^H)\\&\quad\times\Bigg[\left(q+\frac{1-q}{|\mathcal{A}^H|}\right)1_{\mathbf{a}^H=\pi^{*,H}_{T^j}(\mathbf{s}^H_r,\mathbf{e}^j)}+\frac{1-q}{|\mathcal{A}^H|}1_{\mathbf{a}^H\neq\pi^{*,H}_{T^j}(\mathbf{s}^H_r,\mathbf{e}^j)}\Bigg]\Bigg],\end{aligned} \tag{16}$$

where the ϵ-greedy model in (7) is used for representing human behavior. A confident/rational human can be modeled with a large confidence parameter, i.e., $q=1$. In this case, the posterior update in (15) can be expressed as:

$$\begin{aligned}p_{k+1}(j)&\propto\prod_{r=0}^{k}\Bigg[\sum_{\mathbf{a}^H\in\mathcal{A}^H}P(\mathbf{s}^H_{r+1}\mid\mathbf{s}^H_r,\mathbf{a}^H_r=\mathbf{a}^H)\times 1_{\mathbf{a}^H=\pi^{*,H}_{T^j}(\mathbf{s}^H_r,\mathbf{e}^j)}\Bigg]p_0(j)\\&=\prod_{r=0}^{k}\Big[P(\mathbf{s}^H_{r+1}\mid\mathbf{s}^H_r,\mathbf{a}^H_r=\pi^{*,H}_{T^j}(\mathbf{s}^H_r,\mathbf{e}^j))\Big]p_0(j),\end{aligned} \tag{17}$$

Algorithm 1 The proposed high-level human intention learning for cooperative decision-making.

High-level subtasks $\{T^1, T^2, ..., T^N\}$; terminal state of subtasks $\mathcal{G}^1, ..., \mathcal{G}^N$; human confidence q.

Offline Step

Run N reinforcement learning algorithms corresponding to all subtasks: $\pi^{*,H}_{T^j}(\mathbf{s}^H, \eta = \mathbf{e}^j)$, for $\mathbf{s}^H \in \mathbf{S}^H$ and $j = 1, ..., N$.

Run N reinforcement learning algorithms corresponding to $\pi^{*,A}_{H=T^j}(\mathbf{s}^A, \eta)$, for $\mathbf{s}^A \in \mathbf{S}^A$, $\eta \in \{0,1\}^N$ and $j = 1, ..., N$.

Online Step

Set the initial human intention as p_0 (e.g., $[\frac{1}{N}, ..., \frac{1}{N}]^N$), and the initial status of subtasks as $\eta_0 = [0, ..., 0]^T$.

Set the initial human and agent state as $\mathbf{s}^H_0, \mathbf{s}^A_0$.

for $k = 0, 1, 2, ...$ **do** /* Until all subtasks are performed */

Active learning according to the posterior of human intention:

$$\mathbf{a}^A_k = \underset{\mathbf{a}^A \in \mathcal{A}^A}{\operatorname{argmax}} \sum_{j=1}^{N} p_k(j)\, q^{*,A}_{H=T^j}(\mathbf{s}^A_k, \eta_k, \mathbf{a}^A).$$

Agent takes action $\mathbf{a}^A_k$, and new states of agent and human are observed, i.e., $\mathbf{s}^A_{k+1}$, s^H_{k+1}.

Update the subtask tracker: $\eta_{k+1}(j) = \begin{cases} 1 & \text{if } \mathbf{s}^A_{k+1} \text{ or } \mathbf{s}^H_{k+1} = \mathcal{G}^j \\ \eta_k(j) & \text{otherwise} \end{cases}$, for $j = 1, ..., N$.

Human intention posterior update:

$$p_{k+1}(j) = \frac{\sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, \mathbf{a}^H_k = \mathbf{a}^H) \pi^H_{T^j}(\mathbf{a}^H \mid \mathbf{s}^H_k)\, 1_{\eta_{k+1}(j)=0}\, p_k(j)}{\sum_{l=1}^{N} \sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{k+1} \mid \mathbf{s}^H_{0:k}, \mathbf{a}^H_k = \mathbf{a}^H) \pi^H_{T^l}(\mathbf{a}^H \mid \mathbf{s}^H_k)\, 1_{\eta_{k+1}(l)=0}\, p_k(l)}.$$

end for

for $j = 1, ..., N$. The posterior of human intention depends on the likelihood of state transitions under the actions reflecting the optimal human policy for any given subtask. In this case, the true intention is expected to be better recognizable in the posterior of intentions since the movements of confident human provide rich information about their intentions.

In contrast, if the human is not confident and acts like a random decision-maker, its behavior can be modeled using the confidence parameter $q = 0$. Replacing this into (16) leads to:

$$p_{k+1}(j) \propto \frac{1}{|\mathcal{A}^H|} \prod_{r=0}^{k} \left[\sum_{\mathbf{a}^H \in \mathcal{A}^H} P(\mathbf{s}^H_{r+1} \mid \mathbf{s}^H_r, \mathbf{a}^H_r = \mathbf{a}^H) \right], \tag{18}$$

for $j = 1, ..., N$; where the intention posterior is independent of subtask T^j, meaning that the posterior distribution of the intention becomes completely indistinguishable in this condition. In fact, the posterior of human intention remains unchanged and equal to the prior, no matter how much human data is observed. Intuitively, this can be interpreted as a lack of measurable intentions in data from a human with random behavior, which is also captured by the proposed method in (18).

VI. Numerical Experiments

In this section, we analyze the performance of the proposed policy through a grid problem consisting of a single human and an agent. The comparison is made with the following approaches: 1) *Baseline policy*, which demonstrates the case where the agent is fully aware of the human intention, and the agent's decisions are made according to this knowledge. The baseline policy represents the best results that could be achieved with no knowledge about human intention and limited communication. 2) *Distance-based probabilistic policy* [25], [28], which approximates the likelihood of observing any data using a distance-based measure (e.g., Euclidean) without accounting for the human trajectory. Here, we measure the distance by employing the closest human path from any given state to all subtasks' terminal states. 3) *Non-communicative policy*, which

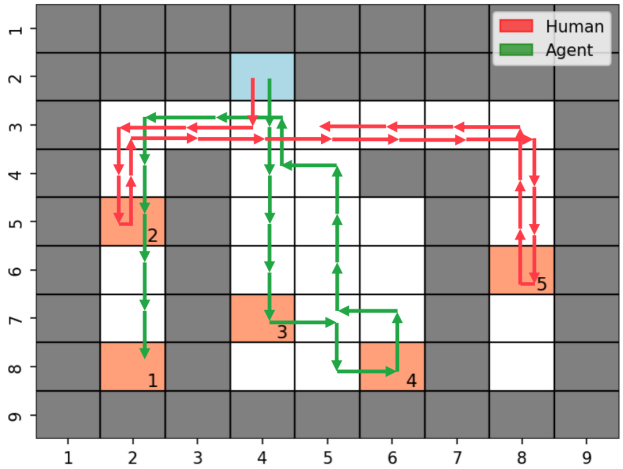

Fig. 1: The maze environment containing a human and an agent with 5 subtasks. The blue cell indicates the initial agent and human state, and the orange cells indicate the terminal states for five subtasks. The human's and the agent's movements are shown by red and green arrows, respectively.

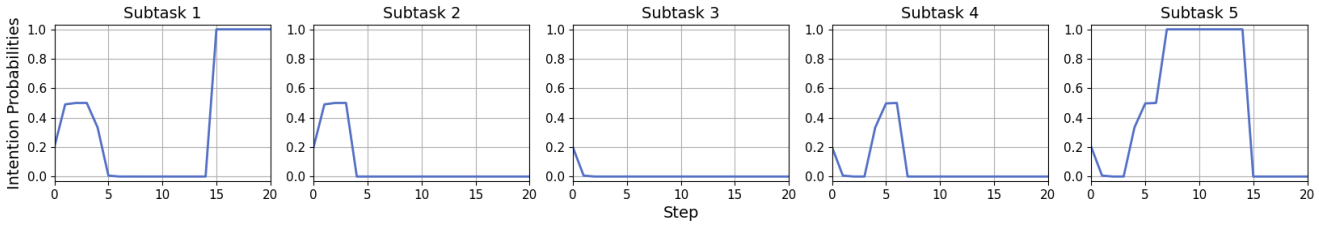

Fig. 2: The posterior distribution of human intentions for five subtasks computed by the proposed method.

assumes the agent makes isolated actions without accounting for human intentions and decisions. Note that the non-communicative policy can be obtained using a reinforcement learning approach defined over the agent state and subtask trackers.

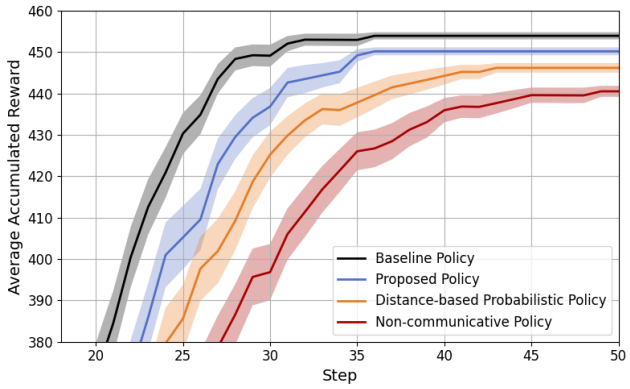

Fig. 3: The average accumulated reward obtained by different policies.

Fig. 1 represents the maze environment containing a single agent and a human with 5 subtasks, indicated by 1 to 5. The human and agent can be in one of the 32 possible cells in the maze and can select one of the four possible actions $\mathcal{A}^A = \mathcal{A}^H = \{\text{Up, Down, Left, Right}\}$ at any given state and at each step. The human and agent movements are stochastic; they move to the state at the direction indicated by the selected action with probability 0.95, and end up in one of the cross perpendicular cells with 0.025 probability. The cooperative reward function, according to the subtask tracker, can be expressed as:

$$R(\mathbf{s}, \eta, \mathbf{a}, \mathbf{s}', \eta') = 100\,\|\eta' - \eta\|_1 - 2, \tag{19}$$

where performing each subtask has a reward of 100, and each step delay has a reward of -2. Thus, the reward encourages the agent and human to perform all subtasks as quickly as possible. It should be noted that the proposed method is capable of considering the priority of different subtasks by using distinct rewards for each subtask.

The human is modeled using ϵ-greedy policy in (7), with the confidence rate $q = 0.8$. For generating the optimal human policy, the discount factor is set to $\gamma = 0.95$. All average results are obtained using 100 independent runs with the standard error mean represented by shadows.

In the first experiment, both the human and the agent start at the same location on top of the maze, shown in Fig. 1. The human moves in the environment with the intention

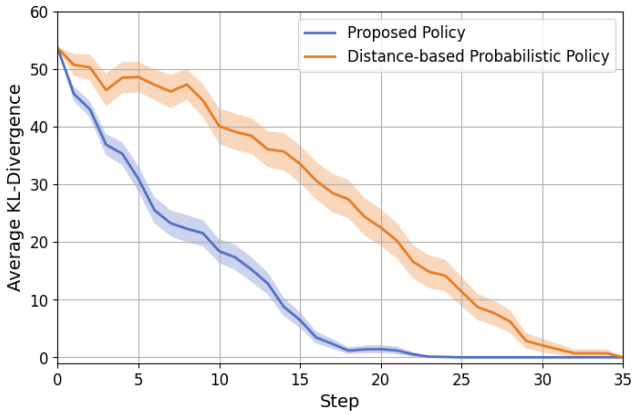

Fig. 4: The average KL-divergence of the true human intention and the inferred posterior of human intentions.

of performing subtasks 2, 5, 3, and 4, respectively. The agent movements under the proposed policy are shown in Fig. 1. Fig. 2 represents the intention probabilities obtained by the proposed policy, which is obtained using only human movements. Initially, the human moves towards subtask 2, which is evident from the higher intention probability for subtasks 1 and 2 in the early steps. Upon completion of subtask 2, the human shifts towards subtask 5, rather than subtask 1, even though it is nearby. This behavior illustrates the complexity of human behavior, influenced by factors such as risk, capability, or other external information. One can observe that the proposed method quickly adapts its policy and moves toward subtask 1, as soon as it assigns high probability to the human intention to perform subtask 5. This adaptability in intention learning and decision-making helps to prevent task duplication and facilitate cooperation, as it is evident in the agents' movements in Fig. 1 and the intention probabilities in Fig. 2.

The average accumulated reward obtained by different policies is presented in Fig. 3. The human and the agent start from the same location, which is randomly selected in different runs from all the states in the maze except the terminal states. The order of the subtasks that the human wants to perform is also generated randomly in each run. As expected, the baseline yields the highest average accumulated reward, as the agent can navigate optimally according to the full knowledge of human intentions (i.e., perfect communication). The proposed policy yields the best results among methods that do not rely on knowledge of human intention. This demonstrates the capability of the proposed policy to implicitly reason about human intention and incorporate it

for effective and adaptable decision-making. A worse performance can be seen under the non-communicative policy, which is due to the inability of this policy to account for human intention. Finally, the distance-based probabilistic policy has achieved better performance than the non-communicative policy but worse than the proposed method. This is due to the likelihood approximation and distance-based measure employed by this policy, which makes it oblivious to the relationship between intention and human trajectory.

Fig. 4 represents the average Kullback-Leibler (KL) divergence between the true human intention and the inferred posterior of human intention by the proposed policy and distance-based probabilistic policy. Both methods exhibit a large initial KL value in the early steps, due to the uniform initial intention probabilities and the limited observed human data. As more human data is observed, the KL divergence under the proposed policy quickly approaches zero, indicating a close match between the inferred and true human intentions. However, the reduction in the KL value is slower under the distance-based policy, indicating the challenge of capturing the true human intention by this method.

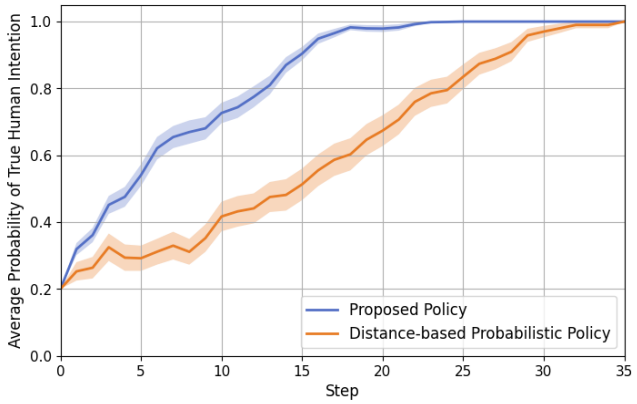

Fig. 5: The average probability of the true human intention.

Fig. 5 represents the average probability of the true human intention under the proposed policy and the distance-based probabilistic policy, i.e., $p_k(i_k^*)$, where i_k^* is the true human intention at time step k. A uniform initial intention probability has led to a small initial probability of the true human intention by both methods at early steps. However, as more data are observed, the proposed method has outperformed the distance-based policy, and the true human intention has quickly converged to one. This demonstrates the capability of the proposed policy to better infer the true human intention implicitly, without any communication or access to human actions.

The impact of human confidence/rationality on the performance of the proposed policy is also investigated. Fig. 6 shows the average accumulated reward upon termination of all subtasks when the agent is guided by different policies alongside the human with the following four confidence rates: $q = 1, 0.8, 0.6, 0.4$. The average accumulated reward decreases for all policies as human confidence decreases. It can be observed that the proposed policy performs similar to the baseline in cases with large human confidence rates. This can be justified given the fact that the intentions of the confident human are more recognizable; as a human becomes less confident, actions become less predictable, and implicit reasoning about the human intention becomes more challenging. In a particular case of $q = 0.4$, which represents a low-confident human, the results of the proposed policy decline sharply and approach those of the non-communicative policy. The proposed policy outperforms the distance-based probabilistic policy in all cases, demonstrating the superiority of the proposed method in accurately capturing the human intentions.

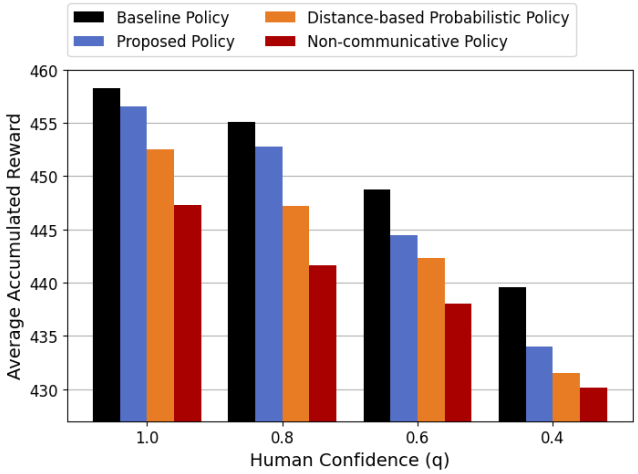

Fig. 6: The impact of human confidence/rationality on the performance of the proposed policy.

VII. Conclusion

This paper presents a method for inferring high-level human intentions within a human-AI team. The objective is to coordinate tasks among humans and AI agents, especially in scenarios where limited or no communication exists among them. Human is modeled as a sub-optimal reinforcement learning agent, and a recursive statistical learning method is introduced for implicit probabilistic reasoning of human intentions. The quantified human knowledge is incorporated into an active learning approach to allow the agent to make effective decisions and coordinate tasks without duplicating efforts. Numerical experiments demonstrate the efficacy of the proposed method in coordinating tasks. Our future work will focus on the scalability of the proposed framework to multiple humans and multiple agents, as well as the extension to partially observable environments.

Acknowledgment

This work has been supported in part by the Office of Naval Research award N00014-23-1-2850, the National Science Foundation awards IIS-2311969 and IIS-2202395, ARMY Research Laboratory award W911NF2320179, and ARMY Research Office award W911NF2110299.

REFERENCES

[1] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. De Freitas, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International conference on machine learning*, pp. 3040–3049, PMLR, 2019.

[2] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–autonomy teaming: A review and analysis of the empirical literature," *Human factors*, vol. 64, no. 5, pp. 904–938, 2022.

[3] E. Seraj and M. Gombolay, "Coordinated control of UAVs for human-centered active sensing of wildfires," in *2020 American control conference (ACC)*, pp. 1845–1852, IEEE, 2020.

[4] T. B. Sheridan, "Human–robot interaction: status and challenges," *Human factors*, vol. 58, no. 4, pp. 525–532, 2016.

[5] J. P. Vasconez, G. A. Kantor, and F. A. A. Cheein, "Human–robot interaction in agriculture: A survey and current challenges," *Biosystems engineering*, vol. 179, pp. 35–48, 2019.

[6] A. K. Inkulu, M. R. Bahubalendruni, and A. Dara, "Challenges and opportunities in human robot collaboration context of industry 4.0-a state of the art review," *Industrial Robot: the international journal of robotics research and application*, vol. 49, no. 2, pp. 226–239, 2022.

[7] A. Kazeminajafabadi and M. Imani, "Optimal joint defense and monitoring for networks security under uncertainty: A POMDP-based approach," *IET Information Security*, vol. 2024, 2024.

[8] N. Asadi, S. H. Hosseini, M. Imani, D. P. Aldrich, and S. F. Ghoreishi, "Privacy-preserved federated reinforcement learning for autonomy in signalized intersections," in *ASCE International Conference on Transportation and Development (ICTD), American Society of Civil Engineers*, 2024.

[9] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *International Journal of Social Robotics*, vol. 11, pp. 575–608, 2019.

[10] J. Lee, "A survey of robot learning from demonstrations for human-robot collaboration," *arXiv preprint arXiv:1710.08789*, 2017.

[11] L. Roveda, J. Maskani, P. Franceschi, A. Abdi, F. Braghin, L. Molinari Tosatti, and N. Pedrocchi, "Model-based reinforcement learning variable impedance control for human-robot collaboration," *Journal of Intelligent & Robotic Systems*, vol. 100, no. 2, pp. 417–433, 2020.

[12] Z. Zhang, H. Zhou, M. Imani, T. Lee, and T. Lan, "Collaborative AI teaming in unknown environments via active goal deduction," *arXiv preprint arXiv:2403.15341*, 2024.

[13] A. Ravari, S. F. Ghoreishi, and M. Imani, "Implicit human perception learning in complex and unknown environments," in *American Control Conference (ACC)*, IEEE, 2024.

[14] S. W. Abeyruwan, L. Graesser, D. B. D'Ambrosio, A. Singh, A. Shankar, A. Bewley, D. Jain, K. M. Choromanski, and P. 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.

[15] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[16] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," *arXiv preprint arXiv:1507.04888*, 2015.

[17] N. Jaques, A. Ghandeharioun, J. H. Shen, C. Ferguson, A. Lapedriza, N. Jones, S. Gu, and R. Picard, "Way off-policy batch deep reinforcement learning of implicit human preferences in dialog," *arXiv preprint arXiv:1907.00456*, 2019.

[18] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[19] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[20] J.-H. Han, S.-J. Lee, and J.-H. Kim, "Behavior hierarchy-based affordance map for recognition of human intention and its application to human–robot interaction," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 5, pp. 708–722, 2016.

[21] A. Ravari, S. F. Ghoreishi, and M. Imani, "Optimal inference of hidden Markov models through expert-acquired data," *IEEE Transactions on Artificial Intelligence*, 2024.

[22] C. Gomez Cubero and M. Rehm, "Intention recognition in human robot interaction based on eye tracking," in *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*, pp. 428–437, Springer, 2021.

[23] A. Buerkle, W. Eaton, N. Lohse, T. Bamber, and P. Ferreira, "Eeg based arm movement intention recognition towards enhanced safety in symbiotic human-robot collaboration," *Robotics and Computer-Integrated Manufacturing*, vol. 70, p. 102137, 2021.

[24] S. Ni, L. Zhao, A. Li, D. Wu, and L. Zhou, "Cross-view human intention recognition for human-robot collaboration," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 189–195, 2023.

[25] Y. Cheng, L. Sun, C. Liu, and M. Tomizuka, "Towards efficient human-robot collaboration with robust plan recognition and trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2602–2609, 2020.

[26] M. Imani and S. F. Ghoreishi, "Two-stage Bayesian optimization for scalable inference in state-space models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5138–5149, 2021.

[27] M. Imani and S. F. Ghoreishi, "Graph-based Bayesian optimization for large-scale objective-based experimental design," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5913–5925, 2021.

[28] S. Jain and B. Argall, "Recursive Bayesian human intent recognition in shared-control robotics," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3905–3912, IEEE, 2018.

[29] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[30] O. Rudovic, M. Zhang, B. Schuller, and R. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," in *2019 International Conference on Multimodal Interaction*, pp. 6–15, 2019.

[31] M. Imani, S. F. Ghoreishi, and U. M. Braga-Neto, "Bayesian control of large MDPs with unknown dynamics in data-poor environments," in *Advances in neural information processing systems*, pp. 8156–8166, 2018.

[32] J. P. L. S. de Almeida, R. T. Nakashima, F. Neves-Jr, and L. V. R. de Arruda, "Bio-inspired on-line path planner for cooperative exploration of unknown environment by a multi-robot system," *Robotics and Autonomous Systems*, vol. 112, pp. 32–48, 2019.

[33] L. Chang, L. Shan, C. Jiang, and Y. Dai, "Reinforcement based mobile robot path planning with improved dynamic window approach in unknown environment," *Autonomous Robots*, vol. 45, no. 1, pp. 51–76, 2021.

[34] M. Jafarnia-Jahromi, L. Chen, R. Jain, and H. Luo, "Posterior sampling-based online learning for the stochastic shortest path model," in *Uncertainty in Artificial Intelligence*, pp. 922–931, PMLR, 2023.

[35] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson, "VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning," in *International Conference on Learning Representations*, 2019.

[36] M. Budd, P. Duckworth, N. Hawes, and B. Lacerda, "Bayesian reinforcement learning for single-episode missions in partially unknown environments," in *Conference on Robot Learning*, pp. 1189–1198, PMLR, 2023.