Hybrid Modeling of Heterogeneous Human Teams for Collaborative Decision Processes

Amirhossein Ravari RAVARI.A@NORTHEASTERN.EDU
Department of Electrical and Computer Engineering, Northeastern University

Seyede Fatemeh Ghoreishi F.GHOREISHI@NORTHEASTERN.EDU
College of Engineering and Khoury College of Computer Sciences, Northeastern University

Tian Lan TLAN@GWU.EDU
Department of Electrical and Computer Engineering, George Washington University

Nathaniel D. Bastian NATHANIEL.BASTIAN@WESTPOINT.EDU
Department of Mathematical Sciences, United States Military Academy

Mahdi Imani M.IMANI@NORTHEASTERN.EDU
Department of Electrical and Computer Engineering, Northeastern University

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

The increasing integration of artificial intelligence (AI) enabled systems with human operators underscores the need for seamless collaboration across various domains. Accurate modeling of human behavior can enable AI-enabled systems to anticipate human decisions and align themselves to support humans in complex tasks. Unlike existing methods focusing primarily on individual human behavioral modeling, this paper models human behavior within heterogeneous teams working toward a cooperative objective. In such teams, members often have varying skills, knowledge, and levels of awareness, which influence their decision-making processes. This paper models team behavior as a sub-optimal, hybrid form of multi-agent reinforcement learning. By leveraging centralized training with hybrid centralized/decentralized execution, the model captures a spectrum of team behaviors, from fully centralized to fully decentralized and in between. This paper quantitatively models each team member's awareness and communication levels, enabling the inverse learning of these parameters from observed human team behavior data. Numerical experiments validate the robustness and accuracy of the framework across diverse scenarios and team compositions, underscoring its effectiveness in modeling complex human interactions.

Keywords: Inverse learning, Multi-agent reinforcement learning, Human behavior modeling.

1. Introduction

As artificial intelligence (AI) systems increasingly integrate into domains traditionally managed by humans, the need for effective collaboration between humans and AI agents becomes paramount. Examples include AI systems that help humans in cyberspace detect and counter adversarial actions (Sarker, 2023; Kazeminajafabadi and Imani, 2023), robots that work alongside humans in manufacturing to improve efficiency and precision (Matheson et al., 2019), AI-driven exploration missions where autonomous systems support human navigation and analysis in unknown environments (Wu et al., 2019; Shafti et al., 2020; Imbiriba et al., 2019), and safety applications in which AI systems aid critical decision making, such as in disaster response (O'Neill et al., 2022; Unhelkar et al., 2020). Consequently, modeling human behavior is essential to facilitate seamless human-AI collaboration, enabling AI agents to anticipate human actions and improve both safety and cooperation (Hong et al., 2020; Lin et al., 2024).

Several methods have emerged in recent years to model human behavior, including imitation learning (Le Mero et al., 2022), which learns policies that mimic human actions through supervised learning, and inverse reinforcement learning (IRL), which infers the underlying reward function driving human decisions (Arora and Doshi, 2021; Hoffman et al., 2024; Casper et al., 2023). Notable IRL methods such as Maximum Entropy Inverse Reinforcement Learning (Ziebart et al., 2008) handle uncertainty by modeling behaviors that maximize both reward and entropy. Furthermore, Generative Adversarial Imitation Learning (Ho and Ermon, 2016) uses adversarial training to mimic human behavior. However, these approaches are designed primarily to model individual human behaviors, relying on the trajectories of individuals performing similar tasks. Such techniques have been widely applied in autonomous navigation (Vasquez et al., 2014; Alali and Imani, 2024), human-robot interaction (Liu et al., 2022), and game-theoretic scenarios (Cao and Xie, 2022; Hosseini and Imani, 2024), where their strengths lie in modeling single agent behavior (Wilder et al., 2021). While these methods have been successful in modeling individuals, extending them to multi-agent settings presents challenges due to the complexities of team dynamics.

Recent extensions of IRL to multi-agent settings have aimed to extend single-agent models to the human team. However, these methods can capture homogeneous team behaviors where all members operate under fully centralized or fully decentralized settings. Centralized models assume that all team members have complete and continuous access to each other's information (Natarajan et al., 2010; Suresh et al., 2024). However, this assumption is not realistic in complex real-world environments, where individuals typically operate with only partial or role-specific knowledge of their teammates (Zarei and Shafai, 2024). On the other hand, fully decentralized models, such as those based on QMix (Rashid et al., 2020) or multi-agent proximal policy optimization (PPO) (Yu et al., 2022; Ahmad et al., 2024), model teams as isolated humans where members do not exchange information and have no knowledge of teammates' states. These decentralized models also fail to account for the heterogeneity of real human teams, where members may possess various communication skills, different levels of knowledge, and different situations of awareness. These attributes significantly influence the interactions and decision processes of individuals in the team, resulting in complex team dynamics that neither centralized nor decentralized models can adequately capture them (Tabrez et al., 2020; Zhang et al., 2025; Iftikhar et al., 2023).

This paper presents an innovative hybrid framework designed to address existing limitations in the modeling of team dynamics. Our framework offers a flexible and robust model that considers individual differences in awareness and communication, which are pivotal in collaborative decision making within heterogeneous human teams. For example, some individuals may exchange information with teammates, while others might operate with limited or no knowledge of the teammates. This proposed model captures two key characteristics that influence human decisions: awareness, reflecting the individual's information about the states of the teammates, and communication, indicating the extent to which individuals share information relevant to the task. We assume that communication is structured around a set of subtasks that need to be collaboratively performed by the team, enabling the modeling of the high-level shared mentality within the team. The model employs centralized training with hybrid centralized/decentralized execution, allowing learning policies that capture team behavior with diverse levels of awareness and communication skills. Using the available observed team behavior data in terms of state sequence, our approach formulates inverse learning to estimate the awareness and communication levels of all team members that best justify given the observed data. The analytical results show that fully centralized and fully decentralized models are two special cases of the proposed model, whereas the proposed model captures a wide spec-

trum of team behavior. Additionally, we demonstrate how human rationality levels, representing the degree of randomness of individual decisions, impact team modeling performance. Numerical experiments validate the framework's robustness and adaptability across diverse team compositions and scenarios, underscoring its effectiveness in accurately modeling complex, heterogeneous human interactions.

2. Background

The decentralized Markov decision process (Dec-MDP) is a powerful model for representing the behavior of multiple agents or humans working together (Amato et al., 2013). The Dec-MDP is expressed as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R \rangle$, where $\mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^n$ denotes the joint state space, with $\mathcal{S}^i$ representing the state space of the ith agent. Similarly, $\mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^n$ defines the joint action space, where $\mathcal{A}^i$ corresponds to the action space of the ith human. The transition probability function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ can be factorized as $\mathcal{P} = \mathcal{P}^1 \times \cdots \times \mathcal{P}^n$, where $\mathcal{P}^i(\mathbf{s}^i, a^i, \mathbf{s}'^i) = p(\mathbf{s}'^i \mid \mathbf{s}^i, a^i)$ represents the probability that the ith agent transitions to state $\mathbf{s}'^i$ after taking action a^i in state $\mathbf{s}^i$. The reward function, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, encodes the expected cooperative reward earned when agents take joint actions $\mathbf{a} \in \mathcal{A}$ in state $\mathbf{s} \in \mathcal{S}$.

3. Modeling Humans in a Team

Modeling individual human behavior has been extensively studied in recent years (Hoffman et al., 2024; Alali and Imani, 2025; Ravari et al., 2024; Casper et al., 2023). Humans are often modeled using a stochastic form of the optimal decision-maker, which accounts for the uncertain nature of human decisions within a Markov Decision Process (MDP) (Kazeminajafabadi and Imani, 2024). This stochasticity captures the inherent unpredictability of human decisions, which may arise from the complexity of human behavior, external unmodeled factors, or unmeasurable risks and preferences. Significant progress has been made in human modeling, particularly through IRL and imitation learning methods (Le Mero et al., 2022; Arora and Doshi, 2021), which aim to recover the underlying reward functions and policies from observed human behavioral data.

Extending human modeling and inverse learning to multi-agent settings has been an active area of research (Wu et al., 2023; Mosqueira-Rey et al., 2023; D'Avella et al., 2022; Maisto et al., 2023). While these approaches have advanced the field, they assume agents are homogeneous and operate within fully centralized or fully decentralized settings. This overlooks the inherent heterogeneity of real-world teams, where individuals possess diverse skills, varying levels of awareness, and different communication abilities that influence cooperative behavior. To address this gap, this paper presents a new model that captures the dynamics of heterogeneous human teams for more accurate behavior modeling in collaborative settings.

3.1. Human Awareness and Communication

In a team working toward a common goal, members may differ in the extent of their knowledge about the situation. Some individuals have comprehensive information, including the states and task completion of their teammates, while others have only partial knowledge of their surroundings and teammates. This variation in information access significantly influences individual behavior in the team. These influencing factors are defined using two key parameters:

- **Awareness:** Awareness reflects the extent to which an individual has access to information about teammates' current states. It captures an individual's situational understanding, influencing how they predict and respond to the actions of others. Higher awareness implies a more comprehensive understanding of the team's status, enabling more informed decision-making, while limited awareness confines individuals to act based on partial or local information.

- **Communication:** Communication measures how individuals share their knowledge and intentions with teammates. This parameter encompasses both the frequency and quality of information exchange within the team. Effective communication allows team members to align their actions toward shared goals by conveying subgoals and plans. High communication levels across the team enhance coordination, whereas low communication leads to less effective teamwork, resembling decentralized behavior.

Together, awareness and communication vary among team members, capturing a spectrum of team behaviors ranging from fully decentralized to fully centralized teamwork, with intermediate configurations in between. By parameterizing these factors, the proposed framework can analyze and infer individual behaviors from observed state sequences, yielding a more accurate and flexible representation of human interactions within teams.

Human awareness and communication for human i are represented using two parameters, r^i and c^i. Which r^i denotes the observation radius within which the i-th human can perceive others, and c^i represents the communication distance indicating the area within which the i-th human can exchange information with surrounding individuals. Figure 1 illustrates these ranges. In this example, the human in the center can communicate and share information with the purple agent, who is within the communication range and can observe the purple and red agents that are within the awareness range. It is important to note that the scenario in Figure 1 is merely illustrative; in practice, the communication range may be equal to or greater than the awareness range, depending on specific situational constraints or individual capabilities. It's also assumed that both awareness and communication parameters remain constant throughout the process, though these values may vary among different humans in the team.

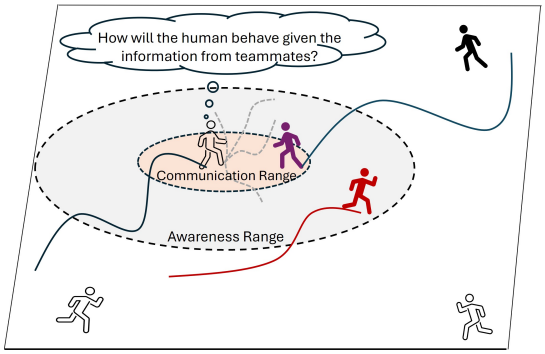

Figure 1: Illustration of a human's communication and awareness ranges in a team setting. The central individual can communicate with the purple agent (within communication range) and observe the red agent (within awareness range).

A special case where r_i and c_i are both zero for all humans corresponds to a fully decentralized team, where individuals act solely based on local information. Conversely, large values of r_i and c_i represent fully centralized decision-making, where each individual is aware of all teammates' states and information at all times. This centralized structure may only be feasible in teams with a single leader guiding all members. These two cases, as well as scenarios where all teammates have similar values of r and c, represent a homogeneous team structure. However, realistic teams are often heterogeneous, with individuals exhibiting diverse levels of awareness and communication. By allowing r_i and c_i to vary among team members, the model captures a wide spectrum of team behaviors, enabling a hybrid approach that reflects both centralized and decentralized dynamics within the team.

3.2. Problem Formulation

This section formulates the problem of inferring the awareness and communication ranges of multiple humans in a team based on their behavioral data. This paper considers a scenario where only human states are available in the collected data. Let $D_T = \{s_t^1, \ldots, s_t^n\}_{t=1}^T$ denote the state data from a team of n humans over T time steps. The objective is to inversely learn the human policy that led to the observed human data. It's assumed that the team's cooperative objective, or reward function, is fully known.

Let $\mathbf{r} = \{r^1, \ldots, r^n\}$ and $\mathbf{c} = \{c^1, \ldots, c^n\}$ represent a realization of the team's awareness and communication parameters. The best estimate of these parameters, given team data D_T, can be computed as:

$$\hat{\mathbf{r}}^*, \hat{\mathbf{c}}^* = \arg\max_{\mathbf{r},\mathbf{c}} p(D_T \mid \mathbf{r}, \mathbf{c}), \tag{1}$$

where $\hat{\mathbf{r}}^*$ and $\hat{\mathbf{c}}^*$ are the maximum likelihood estimates of the parameters, and the maximization is over the entire space of awareness and communication ranges. The likelihood in (1) can be further expanded as:

$$\begin{aligned} p(D_T \mid \mathbf{r}, \mathbf{c}) &= \prod_{t=1}^{T-1} P(\mathbf{s}_{t+1} \mid \mathbf{s}_{1:t}, \mathbf{r}, \mathbf{c}) = \prod_{t=1}^{T-1} \sum_{a^1 \in \mathcal{A}^1} \cdots \sum_{a^n \in \mathcal{A}^n} P(\mathbf{s}_{t+1}, a_t^1 = a^1, ..., a_t^n = a^n \mid \mathbf{s}_{1:t}, \mathbf{r}, \mathbf{c}) \\ &= \prod_{t=1}^{T-1} \sum_{a^1 \in \mathcal{A}^1} ... \sum_{a^n \in \mathcal{A}^n} P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, a_t^1 = a^1, ..., a_t^n = a^n, \mathbf{r}, \mathbf{c}) P(a_t^1 = a^1, ..., a_t^n = a^n \mid \mathbf{s}_{1:t}, \mathbf{r}, \mathbf{c}) \\ &= \prod_{t=1}^{T-1} \sum_{a^1 \in \mathcal{A}^1} ... \sum_{a^n \in \mathcal{A}^n} \underbrace{P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, a_t^1 = a^1, ..., a_t^n = a^n)}_{\text{Transition Term}} \underbrace{P(a_t^1 = a^1, ..., a_t^n = a^n \mid \mathbf{s}_{1:t}, \mathbf{r}, \mathbf{c})}_{\text{Policy Term}}, \end{aligned} \tag{2}$$

where $\mathbf{s}_t = (s_t^1, \ldots, s_t^n)$ represents the state of all team members at time step t. The likelihood is decomposed into two key terms: the transition term, which can be computed using the MDP definition, and the policy term, which indicates the probability of any joint actions by agents given the history of states and the team's awareness and communication parameters. This expression, further described in the next section, explicitly denotes that team behaviors and policy depend on awareness and communication parameters.

3.3. Individual Awareness and Communication in Team

This section describes how the awareness and communication levels impact the team's decisions.

Awareness Level: To model heterogeneous humans in the team, it's assumed that each human makes decisions using local information. Let $\mathbf{s}_{t+1} = (s_{t+1}^1, \ldots, s_{t+1}^n)$ represent the global state of all n team members at time step $t+1$. The ability of the i-th human to access the states of other teammates depends on their awareness level. If r_i represents the awareness radius of the i-th human, the local state accessible to them can be expressed as:

$$\tilde{s}_{t+1}^i(j) = \begin{cases} \mathbf{s}_{t+1}(j) & \text{if } d(\mathbf{s}_{t+1}(i), \mathbf{s}_{t+1}(j)) < r^i \\ \text{no access} & \text{otherwise} \end{cases}, \quad j = 1, \ldots, n, \tag{3}$$

where $d(\cdot, \cdot)$ denotes a distance measure, which can be specified based on the nature of the states (e.g., Euclidean distance). Here, $\mathbf{s}_{t+1}$ represents the global state, and $\tilde{s}_{t+1}^i$ denotes the local state accessible to the i-th human. If $r_i \to \infty$, the i-th human has full access to the states of all teammates

at all times, i.e., $\mathbf{s}_{t+1} = \tilde{s}^i_{t+1}$. Conversely, when $r_i \to 0$, the human has access only to their own local state.

Communication Level: To model communication within a team, a scenario where the team works collaboratively to complete m subtasks is considered, with $\mathcal{G}^j$ representing the terminal state of the j-th subtask. Subtasks provide a structured framework for communication among team members, where team members can exchange information about subtasks if they are within each other's communication range.

To represent subtask completion status within the team, a subtask tracker is defined, which records the status of each subtask for each individual. At time step t, the global subtask tracker is represented as $\eta_t = (\eta_t(1), \ldots, \eta_t(m))$, a binary vector where $\eta_t(j) \in \{0, 1\}$: $\eta_t(j) = 0$ indicates that the j-th subtask is incomplete, and $\eta_t(j) = 1$ indicates completion. Each human has access to its local subtask information, represented by $\tilde{\eta}^i_t$, a binary vector of size m that records the i-th human's knowledge of subtask status at time step t. Humans may have different, and often incomplete, information about the global subtask tracker η_t. Since η_t reflects the collective knowledge of all n humans, it can be expressed as $\eta_t = \bigvee_{l=1}^{n} \tilde{\eta}^l_t$, where $\bigvee$ denotes the element-wise logical OR operator.

Let c_i be the communication range parameter of the i-th human. The local subtask tracker is updated recursively based on the teammates' local subtask information within the communication range as:

$$\tilde{\eta}^i_{t+1}(j) = \tilde{\eta}^i_t(j) \vee 1_{\mathcal{G}^j \in \tilde{s}^i_{t+1}(i)} \vee \left(\bigvee_{l \in \mathcal{C}_i} \tilde{\eta}^l_t(j) \right), \mathcal{C}_i = \{ j \neq i \mid d(\mathbf{s}_{t+1}(i), \mathbf{s}_{t+1}(j)) \leq c^i \} \text{ for } j = 1, \ldots, m \tag{4}$$

where the indicator function $1_{\mathcal{G}^j \in \tilde{s}^i_{t+1}(i)}$ returns 1 if $\mathcal{G}^j \in \tilde{s}^i_{t+1}(i)$ and 0 otherwise. The first term contains the prior information of the ith human about the subtask tracker, the second term indicates if the $i$the human currently completes the jth subtask, and the third term reflects the collective communication from all humans within the communication range at the current time.

This subtask-level communication enables humans to become aware of teammates' knowledge regarding task progress, which significantly influences coordination and decision-making. Additionally, the local subtask tracker $\tilde{\eta}^i_t$ retains the collective history of all humans who have communicated with the i-th human, providing a richer information base that supports more informed decisions.

3.4. Modeling a Heterogeneous Human Team

To express the human team behavioral policy under given team awareness and communication parameters $(\mathbf{r}, \mathbf{c})$, the available human data D_T is mapped into individualized information states as follows:

$$D_T = \{\mathbf{s}_t(1), \ldots, \mathbf{s}_t(n)\}_{t=1}^{T} \to \tilde{D}_T^{\mathbf{r},\mathbf{c}} = \{(\tilde{s}^1_t(1), \tilde{\eta}^1_t), \ldots, (\tilde{s}^n_t(n), \tilde{\eta}^n_t)\}_{t=1}^{T}, \tag{5}$$

where $\tilde{s}^i_t$ and $\tilde{\eta}^i_t$ are computed recursively using (3) and (4), given $\mathbf{r}$ and $\mathbf{c}$. Specifically, when $\mathbf{r} = \mathbf{c} = [0, \ldots, 0]$, the mapping retains only the local state and subtask tracker for each human. In contrast, as $\mathbf{r} = \mathbf{c} \to \infty$, all humans have access to the same global state and subtask tracker.
Using the mapped data $\tilde{D}_T^{\mathbf{r},\mathbf{c}}$ in (5), the policy term in (2) can be expressed as:

$$\begin{aligned} p(a^1_t = a^1, ..., a^n_t = a^n \mid \mathbf{s}_{1:t}, \mathbf{r}, \mathbf{c}) &= p(a^1_t = a^1, ..., a^n_t = a^n \mid \tilde{s}^1_{1:t}, \tilde{\eta}^1_{1:t}, ..., \tilde{s}^n_{1:t}, \tilde{\eta}^n_{1:t}, \mathbf{r}, \mathbf{c}) \\ &= \prod_{i=1}^{n} p(a^i_t = a^i \mid \tilde{s}^i_t, \tilde{\eta}^i_t, \mathbf{r}, \mathbf{c}), \end{aligned} \tag{6}$$

where the history of states is replaced by each individualized information state, assuming that each human acts independently of the history, given the currently available information state. This formulation allows the policy term to be expressed as a product of probabilities, representing each human policy conditioned on an individualized information state.

For the computation of the last term in (6), the optimal policy for a team of humans with awareness and communication parameters $(\mathbf{r}, \mathbf{c})$ using a multi-agent reinforcement learning policy is represented. Let $\pi^i : \mathcal{S}^i \times \{0,1\}^m \to \mathcal{A}^i$ be a policy that maps the individualized information state of human i to its action space. The optimal multi-agent policy under parameters $(\mathbf{r}, \mathbf{c})$ can be expressed as:

$$\pi^{1*}_{\mathbf{r},\mathbf{c}}(\tilde{s}^1,\tilde{\eta}^1),...,\pi^{n*}_{\mathbf{r},\mathbf{c}}(\tilde{s}^n,\tilde{\eta}^n) = \underset{\pi^1,...,\pi^n}{\operatorname{argmax}} \mathbb{E}\left[\sum_{t=0}^{h}\gamma^t r_t \,|\, \mathbf{s}_0 = \bigvee_{l=1}^{n}\tilde{s}^l, \eta_0 = \bigvee_{l=1}^{n}\tilde{\eta}^l, a^1_{0:h}\sim\pi^1, ..., a^n_{0:h}\sim\pi^n, \mathbf{r}, \mathbf{c}\right], \quad (7)$$

for all $\tilde{s}^i$ and $\tilde{\eta}^i$, where γ is the discount factor in the horizon h, r_t is the reward at time t, $a^i_{0:h} \sim \pi^i$ denotes the sequence of actions for agent i is generated according to policy π^i and the expectation is taken with respect to the global state and subtask tracker transitions, and each agent policy is defined over a portion of the global information represented by individualized information states. For the special case of $\mathbf{r} = \mathbf{c} = [0, ..., 0]$, the optimization in (7) resembles centralized training with decentralized execution, whereas for $\mathbf{r} = \mathbf{c} \to \infty$, both training and execution are centralized.

The optimal multi-agent solution in (7) provides the best human policies for any given awareness and communication parameters $(\mathbf{r}, \mathbf{c})$. As a standard form of human modeling, the i-th human's model is considered as a stochastic approximation of the optimal policy, expressed through the following stochastic model (Arora and Doshi, 2021):

$$\mu^i_{\mathbf{r},\mathbf{c}}(a^i \mid \tilde{s}^i, \tilde{\eta}^i) = p(a^i \mid \tilde{s}^i, \tilde{\eta}^i, \mathbf{r}, \mathbf{c}) = \begin{cases} q^i + \frac{1-q^i}{|\mathcal{A}^i|} & \text{If } a^i = \pi^{i*}_{\mathbf{r},\mathbf{c}}(\tilde{s}^i, \tilde{\eta}^i) \\ \frac{1-q^i}{|\mathcal{A}^i|} & \text{If } a^i \neq \pi^{i*}_{\mathbf{r},\mathbf{c}}(\tilde{s}^i, \tilde{\eta}^i) \end{cases}, \quad (8)$$

where the parameter $q_i \in [0,1]$ specifies the rationality level of human i. Under this human model, the human i follows the optimal policy in (7) with probability $q^i + \frac{1-q^i}{|\mathcal{A}^i|}$ and a non-optimal action with $\frac{1-q^i}{|\mathcal{A}^i|}$ probability.

By substituting (6) and (8) into (2), and assuming that each human's transition function is independent of the others, the following expression for the likelihood function is obtained:

$$\begin{aligned} P(D_T \,|\mathbf{r}, \mathbf{c}) &= \prod_{t=1}^{T-1} \sum_{a^1 \in A^1} ... \sum_{a^n \in A^n} \prod_{i=1}^{n} p(s^i_{t+1} \mid s^i_t, a^i_t = a^i) p(a^i_t = a^i \mid \tilde{s}^i_t, \tilde{\eta}^i_t, \mathbf{r}, \mathbf{c}) \\ &= \prod_{t=1}^{T-1} \prod_{i=1}^{n} \sum_{a^i \in \mathcal{A}^i} \mathcal{P}^i(s^i_t, a^i, s^i_{t+1}) \left[\left(q^i + \frac{1-q^i}{|\mathcal{A}^i|} \right) 1_{a^i = \pi^{i*}_{\mathbf{r},\mathbf{c}}(\tilde{s}^i_t, \tilde{\eta}^i_t)} + \left(\frac{1-q^i}{|\mathcal{A}^i|} \right) 1_{a^i \neq \pi^{i*}_{\mathbf{r},\mathbf{c}}(\tilde{s}^i_t, \tilde{\eta}^i_t)} \right], \end{aligned} \quad (9)$$

where the transition term is factorized for each human. Maximizing this likelihood function provides the maximum likelihood estimate of the team's awareness and communication parameters in (1). Once this maximum likelihood estimate is obtained, the policy for the i-th human can be estimated by substituting the inferred $\mathbf{r}^*$ and $\mathbf{c}^*$ into (8).

The complexity of the proposed method depends on the number of agents, as each additional agent introduces new parameters and computational requirements. Evaluating the likelihood for

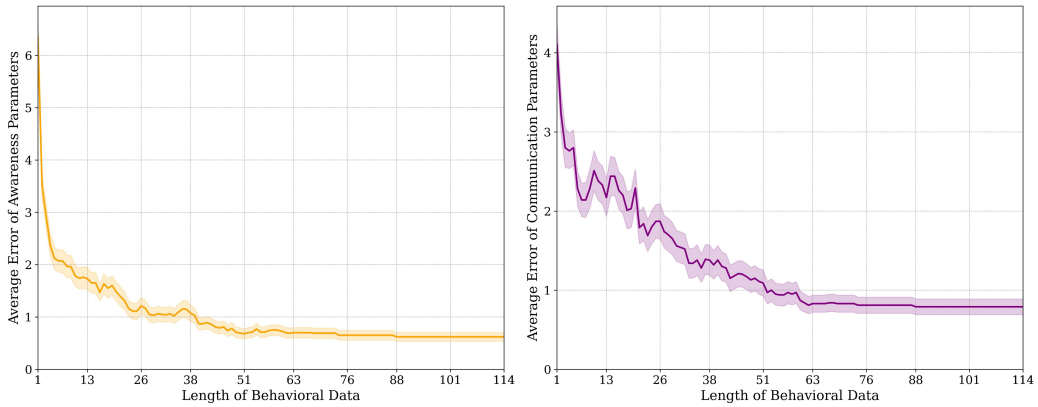

Figure 3: Average error of the inferred awareness (left) and communication (right) parameters with respect to the length of team behavioral data.

any given $(\mathbf{r}, \mathbf{c})$ requires performing a multi-agent reinforcement learning method to approximate the humans' optimal policies. Specifically, as team size increases, the number of awareness and communication parameters grows linearly, with $2n$ parameters for n humans. This scaling affects both the optimization process for parameter inference and the computational burden of evaluating the likelihood function in (9). However, due to the offline nature of optimization in this paper, sample-efficient methods like Bayesian optimization (Frazier, 2018; Imani and Ghoreishi, 2021) can be leveraged for efficient parameter estimation. Future research will explore time-varying human behavior, requiring real-time and recursive computation of parameters.

4. Numerical Experiments

In this section, the performance of the proposed method is evaluated in an environment shown in Figure 2, where four humans collaborate within a maze. Each human can take one of four actions: $\mathcal{A} = \{\text{Up}, \text{Left}, \text{Right}, \text{Down}\}$. The state transitions are stochastic; with probability $0 < \zeta \leq 1$, each human moves in the intended direction, while with probability $(1-\zeta)/2$, they may instead move to one of the perpendicular cells. If movement would result in hitting a wall, the human remains in place. The team's objective is to cooperatively reach five target cells, marked in orange, scattered throughout the maze. Arriving at a target cell rewards the team with +100 points, while each movement incurs a penalty of -1 point, encouraging the team to complete the objective in as few moves as possible. This setup balances individual action decisions and team-oriented goals, modeling realistic scenarios where efficient, coordinated decision-making is essential for optimal performance. Figure 2 illustrates a sample path of four humans with awareness radius $r^1 = ... = r^4 = 5$ and communication radius $c^1 = ... = c^4 = 5$. The human denoted by the black icon reaches $\mathcal{G}^2$ while the blue agent adjusts its trajectory to achieve $\mathcal{G}^1$. The behavior of the agents adapts based on their respective awareness and communication parameters, showcasing the model's ability to simulate coordination in heterogeneous teams.

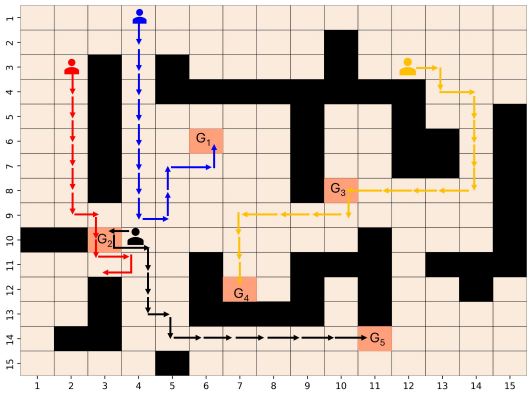

Figure 2: Environment setup in the numerical experiment with four agents and five goal cells.

The humans' varying levels of awareness and communication capabilities are represented by parameters $r^i, c^i \in [0, 10]$, where lower values indicate lower communication and awareness skills. For training, an Independent Deep Q-Network (IDQN) approach is used to perform centralized training of decentralized human policies based on local information. The following parameters are used in our experiments unless stated otherwise: $\gamma = 0.99$, $\zeta = 0.9$, and $q^i = 0.9$.

In the first experiment, the average error in the inferred awareness and communication parameters over 100 runs is evaluated, with each run using randomly generated true parameters, i.e., $\mathbf{r}^*, \mathbf{c}^*$. Figure 3 shows that the average error for both parameters decreases as more data are collected. Specifically, the final inferred parameters achieve a difference of approximately less than 1 level from the true parameters. This error arises from natural stochasticity in state transitions, variations in human rationality, and limited data availability, all of which obscure the visibility of the true latent parameters within human trajectories.

To evaluate the performance of the proposed team model in predicting team behavior, the predictive accuracy (i.e., the proportion of actions that the model selects identically to those of the ground-truth team) of our model is presented in comparison with existing centralized and decentralized team models. Figure 4 represents that the proposed method achieves significantly higher predictive accuracy, reaching approximately 77%, compared to 57% for the centralized model and 26% for the decentralized model. This improvement can be attributed to the model's ability to infer diverse human characteristics within the team, specifically in terms of awareness and communication, and to leverage these parameters to characterize future human behavior. In contrast, fully centralized and fully decentralized models fall short in scenarios involving varied team compositions, where some agents may operate with a centralized perspective while others adopt a more decentralized approach or something in between. By accounting for these intermediate configurations, the proposed model demonstrates enhanced flexibility and predictive capability in modeling complex team dynamics.

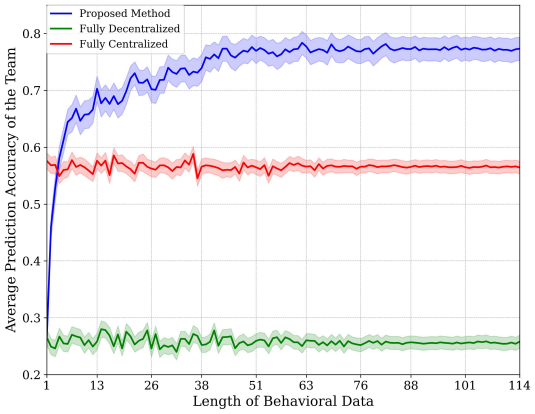

Figure 4: Average prediction accuracy with respect to the length of behavioral data for the proposed method, fully centralized and fully decentralized models.

Table 1 presents the impact of the number of agents on the prediction accuracy of the proposed method compared to centralized and decentralized models. The proposed method consistently outperforms the other approaches, achieving an accuracy of 0.64 ± 0.09 with 2 agents, increasing to 0.78 ± 0.08 with 3 agents, and maintaining high accuracy with 4 agents (0.77 ± 0.08). In contrast, the centralized and decentralized models show lower performance. When the team size is reduced to a single agent, all models perform similarly, as awareness and communication parameters have no effect. These results highlight the robustness and scalability of the proposed method across varying team sizes.

Table 1: Average prediction accuracy across different methods and varying numbers of humans.

Method	**2 Humans**	**3 Humans**	**4 Humans**
Proposed Method	0.64 ± 0.09	0.78 ± 0.08	0.77 ± 0.08
Centralized Method	0.46 ± 0.04	0.58 ± 0.02	0.5 ± 0.04
Decentralized Method	0.37 ± 0.09	0.22 ± 0.02	0.31 ± 0.03

Figure 5 illustrates the impact of human rationality levels on prediction accuracy across various methods. The proposed method (blue) consistently outperforms the centralized (brown) and decentralized (green) models, particularly at higher rationality levels ($q = 0.9$ and $q = 0.99$), where it achieves significantly higher accuracy. This demonstrates the method's ability to leverage human heterogeneity effectively, capturing agent actions with superior precision. As rationality decreases ($q = 0.1$), human actions increasingly resemble random behavior, reducing the performance gap between models. However, even under these conditions, the proposed method remains robust and adaptable, showcasing its effectiveness across diverse rationality scenarios.

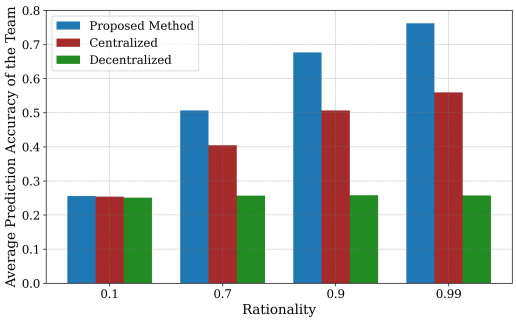

Figure 5: Average prediction accuracy of proposed, centralized, and decentralized methods as a function of human rationality levels.

Our final results evaluate how accurately the proposed method replicates the behavior of the original team, defined as the human team from which the observed data was collected and used to infer the parameters (r, c). We have set the awareness parameter to zero for all humans and examined three teams with uniform communication parameters ($c^i = 0, 3, 7$). The left plot in Figure 6 shows the average communication per step, with black bars representing the original team and blue bars showing the proposed method. Decentralized policies exhibit zero communication, while centralized policies have the highest communication. The right plot in Figure 6 displays the average reward error compared to the true team, where the proposed method achieves the smallest error. As the communication radius increases, the average reward error decreases for the centralized approach because it can leverage the added shared information to better match true centralized behavior. In contrast, the average reward error increases for the decentralized approach, as it moves away from its no-communication baseline.

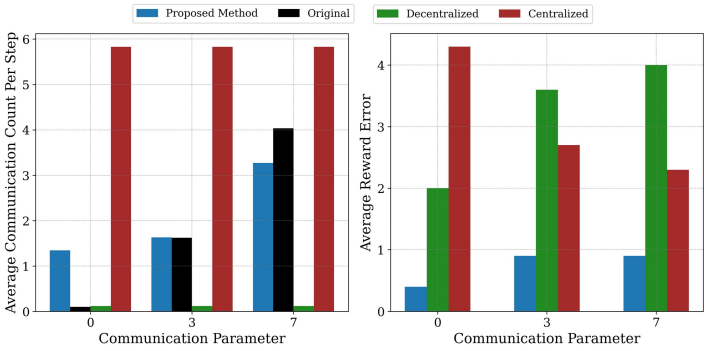

Figure 6: Average communication counts and reward errors for teams with varying communications: zero ($c^i = 0$), medium ($c^i = 3$), and high ($c^i = 7$), as inferred by different policies.

5. Conclusion

This paper presents a robust framework for modeling team behavior in human-AI collaborations, emphasizing the crucial roles of awareness and communication in decision-making processes. By combining centralized training with hybrid centralized/decentralized execution, the model captures a wide spectrum of team dynamics, from fully centralized to decentralized settings. Through inverse learning, the proposed approach effectively infers individual awareness and communication parameters, enabling AI systems to align closely with human behaviors. Analytical and numerical results demonstrate the framework's adaptability and predictive accuracy across diverse scenarios. Future work will explore scaling this framework to larger and more diverse team compositions, incorporating dynamic communication structures, real-time adjustments to awareness levels, and partial observability in complex environments.

Acknowledgments

The authors acknowledge the support of the U.S. Military Academy (USMA) under Cooperative Agreement No. W911NF-23-2-0175, Office of Naval Research award N00014-23-1-2850, Army Research Laboratory awards W911NF-24-1-0098 and W911NF-23-2-0207, and the National Science Foundation award IIS-2311969.

References

Ahmad Ahmad, Mehdi Kermanshah, Kevin Leahy, Zachary Serlin, Ho Chit Siu, Makai Mann, Cristian-Ioan Vasile, Roberto Tron, and Calin Belta. Accelerating proximal policy optimization learning using task prediction for solving games with delayed rewards. *arXiv preprint arXiv:2411.17861*, 2024.

Mohammad Alali and Mahdi Imani. Bayesian reinforcement learning for navigation planning in unknown environments. *Frontiers in Artificial Intelligence*, 7:1308031, 2024.

Mohammad Alali and Mahdi Imani. Deep Reinforcement Learning Data Collection for Bayesian Inference of Hidden Markov Models. *IEEE Transactions on Artificial Intelligence*, 2025.

Christopher Amato, Girish Chowdhary, Alborz Geramifard, N Kemal Üre, and Mykel J Kochenderfer. Decentralized control of partially observable Markov decision processes. In *52nd IEEE Conference on Decision and Control*, pages 2398–2405. IEEE, 2013.

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

Kun Cao and Lihua Xie. Game-theoretic inverse reinforcement learning: A differential pontryagin's maximum principle approach. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (11):9506–9513, 2022.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Salvatore D'Avella, Gerardo Camacho-Gonzalez, and Paolo Tripicchio. On multi-agent cognitive cooperation: Can virtual agents behave like humans? *Neurocomputing*, 480:27–38, 2022.

Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

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.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.

Seyed Hamid Hosseini and Mahdi Imani. Dynamic intervention in gene regulatory networks: A partially observed zero-sum markov game. In *2024 IEEE Conference on Control Technology and Applications (CCTA)*, pages 774–781. IEEE, 2024.

Rehan Iftikhar, Yi-Te Chiu, Mohammad Saud Khan, and Catherine Caudwell. Human–agent team dynamics: A review and future research opportunities. *IEEE Transactions on Engineering Management*, 2023.

Mahdi Imani and Seyede Fatemeh Ghoreishi. Scalable inverse reinforcement learning through multifidelity Bayesian optimization. *IEEE transactions on neural networks and learning systems*, 33 (8):4125–4132, 2021.

Tales Imbiriba, Gerald LaMountain, Peng Wu, Deniz Erdo**u**gmu**c**s, and Pau Closas. Change detection and Gaussian process inference in piecewise stationary environments under noisy inputs. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 530–534. IEEE, 2019.

Armita Kazeminajafabadi and Mahdi Imani. Optimal monitoring and attack detection of networks modeled by Bayesian attack graphs. *Cybersecurity*, 6(1):22, 2023.

Armita Kazeminajafabadi and Mahdi Imani. Optimal joint defense and monitoring for networks security under uncertainty: A pomdp-based approach. *IET Information Security*, 2024(1):7966713, 2024.

Luc Le Mero, Dewei Yi, Mehrdad Dianati, and Alexandros Mouzakitis. A survey on imitation learning techniques for end-to-end autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14128–14147, 2022.

Yuxin Lin, Seyede Fatemeh Ghoreishi, Tian Lan, and Mahdi Imani. High-level human intention learning for cooperative decision-making. In *2024 IEEE Conference on Control Technology and Applications (CCTA)*, pages 209–216. IEEE, 2024.

Wentao Liu, Junmin Zhong, Ruofan Wu, Bretta L Fylstra, Jennie Si, and He Helen Huang. Inferring human-robot performance objectives during locomotion using inverse reinforcement learning and inverse optimal control. *IEEE Robotics and Automation Letters*, 7(2):2549–2556, 2022.

Domenico Maisto, Francesco Donnarumma, and Giovanni Pezzulo. Interactive inference: a multiagent model of cooperative joint actions. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

Eloise Matheson, Riccardo Minto, Emanuele GG Zampieri, Maurizio Faccio, and Giulio Rosati. Human–robot collaboration in manufacturing applications: A review. *Robotics*, 8(4):100, 2019.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

Sriraam Natarajan, Gautam Kunapuli, Kshitij Judah, Prasad Tadepalli, Kristian Kersting, and Jude Shavlik. Multi-agent inverse reinforcement learning. In *2010 ninth international conference on machine learning and applications*, pages 395–400. IEEE, 2010.

Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, 64(5):904–938, 2022.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Amirhossein Ravari, Seyede Fatemeh Ghoreishi, and Mahdi Imani. Optimal inference of hidden Markov models through expert-acquired data. *IEEE Transactions on Artificial Intelligence*, 5(8): 3985–4000, 2024.

Iqbal H Sarker. Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. *Security and Privacy*, 6(5):e295, 2023.

Ali Shafti, Jonas Tjomsland, William Dudley, and A Aldo Faisal. Real-world human-robot collaborative reinforcement learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11161–11166. IEEE, 2020.

Prasanth Sengadu Suresh, Siddarth Jain, Prashant Doshi, and Diego Romeres. Open human-robot collaboration using decentralized inverse reinforcement learning. *arXiv preprint arXiv:2410.01790*, 2024.

Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, 1:259–267, 2020.

Vaibhav V Unhelkar, Shen Li, and Julie A Shah. Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 329–341, 2020.

Dizan Vasquez, Billy Okal, and Kai O Arras. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1341–1346. IEEE, 2014.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1526–1533, 2021.

Haochen Wu, Pedro Sequeira, and David V Pynadath. Multiagent inverse reinforcement learning via theory of mind reasoning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 708–716, 2023.

Peng Wu, Tales Imbiriba, Gerald LaMountain, Jordi Vila-Valls, and Pau Closas. Wifi fingerprinting and tracking using neural networks. In *Proceedings of the 32nd International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2019)*, pages 2314–2324, 2019.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

Fatemeh Zarei and Bahram Shafai. Consensus of multi-agent singular systems by using an algebraic transformation. In *2024 32nd Mediterranean conference on control and automation (MED)*, pages 682–687. IEEE, 2024.

Z. Zhang, H. Zhou, Mahdi Imani, T. Lee, and T. Lan. Learning to Collaborate with Unknown Agents in the Absence of Reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.