Belief-Driven Value Alignment for Human-Robot Collaboration

Saisai Li, Bing Shi*, Yiming Xia, Xiao Su

School of Computer Science and Artificial Intelligence, Wuhan University of Technology
Wuhan, 430070, China
{saisai.li, bingshi, 284660, 349012}@whut.edu.cn

Abstract

As intelligent systems advance rapidly, human-robot collaboration is becoming increasingly important. Ensuring that the intelligent agent's behaviors match human intentions and value preferences is crucial for effective collaboration, which is termed the value alignment problem. Within the Reinforcement Learning (RL) paradigm, value alignment typically relies on pre-designed reward functions, and Cooperative Inverse Reinforcement Learning (CIRL) is often used to model value alignment as a human-robot game. However, existing works often assume that human is perfectly rational, and can fully obtain robot's belief on human's preference. To address this limitation, we propose a Particle Filter-based Hierarchical Dynamic Programming algorithm (PFHDP). By modeling the robot's belief state, this algorithm ensures the correct updates of human's estimate of the robot's belief. This allows human to adopt more targeted pedagogical behaviors to guide the robot based on her understanding of the robot's current belief, achieving belief alignment between human and robot and thereby promoting value alignment more effectively. Furthermore, we run experiments to evaluate the proposed method in two cooperative scenarios against some typical benchmark approaches. The experimental results show that our method can strengthen the alignment of belief states between human and robot, leading to enhanced value alignment.

Introduction

With the rapid development of artificial intelligence and the gradual rise of applications such as autonomous driving, service robotics, and virtual reality, human demand for intelligent machine assistants has increased significantly, which poses significant challenges for safe and effective human-robot interaction(Semeraro, Griffiths, and Cangelosi 2023; Inkulu et al. 2022). Since more advanced agents tend to have greater undesirable effects (e.g., manipulation (Perez et al. 2023) and deception (Park et al. 2024b)), this has triggered research efforts to ensure that these agents pursue the right goals (Ord 2020; Bucknall and Dori-Hacohen 2022). Thus, the value alignment problem has been proposed to reach a common understanding and awareness of task goals between humans and AI agents, i.e., to ensure that AI agents behave in accordance with human intentions and preferences (Gabriel 2020; Ji et al. 2023). In real-world scenarios, value alignment is generally manifested in the ability of an AI agent to infer user preferences based on user conversations or behaviors, and to perform behaviors expected by humans (Zhang et al. 2025; Gabriel and Ghazavi 2022) (Subsequently, "robot" refers generically to an AI agent).

Specifically, in Reinforcement Learning (RL) scenarios, this process is realized through a human pre-designed reward function that the robot continuously interacts with the environment based on this reward function, continuously adjusts its strategy, and ultimately accomplishes the human-specified task goals. Currently, such reinforcement learning methods have been widely used in several fields, including games (Vinyals et al. 2019), autonomous driving (Zhou et al. 2020), robot control (Chen et al. 2022), and so on. In these scenarios, robots are able to formulate effective strategies by maximizing cumulative rewards when a suitable reward structure exists. However, in many scenarios, humans are often unable to explicitly formalize or express their true goals.

Therefore, instead of optimizing a pre-specified reward function, existing research focuses on robots attempting to infer human preferences, i.e., robots can learn reward functions over time using human behavior. Common approaches include Reinforcement Learning based on Human Feedback (RLHF) (Yu et al. 2024), Imitation Learning (IL) (Wu et al. 2019), Inverse Reinforcement Learning (IRL) (Abbeel and Ng 2004), and Collaborative Inverse Reinforcement Learning (CIRL) (Hadfield-Menell et al. 2016). CIRL formulates value alignment as a real-time two-player game where human and robot share an identical reward function. More importantly, the reward function and signals are invisible to the robot, which can only be inferred from the human. There is some work that models human in CIRL as perfectly rational (Shah et al. 2020), which is clearly unrealistic; therefore, some researchers have been trying to relax the assumption of rationality in CIRL (Malik et al. 2018; Fisac et al. 2020).Meanwhile, scholars have highlighted the significance of guidance from human to robot. They argue that people tend to act in a pedagogical manner, actively choosing actions that provide information about preferences (Fisac et al. 2020). Existing work on pedagogical CIRL, takes into account the random and pedagogical behavior of human, which also models the process——robot needs to

*Corresponding author.

make decisions by sustaining Bayesian beliefs regarding human preferences, leveraging the game's information asymmetry. However, the work assumes that human can infer robot's belief directly as an element of their own decision-making process. Evidently, this assumption is unrealistic in real-world scenarios. It fails to incorporate human's understanding of the robot's belief, leading to a critical breakdown in belief alignment. Specifically, human may lack targeted behaviors to align her belief with the robot's during decision-making, hindering value alignment effectiveness.

To address this limitation, we propose a Particle Filter-based Hierarchical Dynamic Programming algorithm (PFHDP) for solving the value alignment problem in CIRL. The algorithm adopts particle filtering to model the robot's belief state based on historical interaction trajectories. The method maintains the human's approximate estimate of the robot's current belief distribution, enabling correct updates to the human's estimation of the robot's belief. Ultimately, it realizes the human's estimation of the robot's belief aligning with the robot's true belief distribution, while the robot's belief can be aligned to the human's true value preference. Thus, human can adopt more targeted pedagogical behaviors to guide the robot based on her estimation of the robot's current belief, realizing human-robot belief alignment, and promoting value alignment more effectively.

We validate the effectiveness of the PFHDP algorithm in two experimental scenarios. The results show that, compared with three typical value alignment methods, our algorithm achieves human-robot belief alignment based on estimating the robot's current beliefs. Furthermore, it yields a more effective value alignment strategy. In addition, ablation experiments on different rationality coefficients demonstrate the following: when the rationality of human decision-making is maintained within a certain range, human actions can provide signals conducive to the update of the robot's belief state. This enables the realization of value alignment.

Related Work

RLHF RLHF dynamically aligns robots to human preferences through human feedback. Some work focuses on accurately modeling human preferences, constructing various reward functions based on different preference categories and granularity (Christiano et al. 2017; Park et al. 2024a; Chakraborty et al. 2024). Other strands of research focus on modeling the feedback approach to optimize the alignment effect, where the robots try to fit human values through behaviors such as labeling, rewarding, and demonstrating in different scenarios (Yu et al. 2024). However, these methods often suffer from feedback lag, making them difficult to adapt to real-time scenarios. CIRL solves this problem through instantaneous interaction in human-robot collaborative scenarios.

IL and IRL IL approach bypasses the design of the reward function and learns human behaviors directly with goal-oriented, implicitly realizing value alignment (Silver et al. 2016; Wu et al. 2019; Stiennon et al. 2020; Lightman et al. 2023). However, the learning process of IL focuses more on the current local state characteristics without sufficiently considering the effects of future states. This may lead to poor performance of the learned strategies in long-term or complex environments. To address this limitation, IRL aligns human preference values by learning reward functions optimized by experts from provided demonstrations.

IRL is an important branch of IL. Unlike general IL methods, IRL considers the recovery of intrinsic reward functions through expert trajectories to achieve value alignment (Abbeel and Ng 2004; Ziebart et al. 2008). Some of the work considers a different reward design model, namely Reward Modeling (Leike et al. 2018; Wu et al. 2021). However, IRL usually assumes that the expert strategy is the optimal one, e.g., robot passively observes isolated human expert. If some of the expert's actions are not optimal, the targeted optimization becomes difficult to implement. CIRL relaxes this restriction. Instead of simply assuming that the human policy is optimal and mimicking human behavior, the robot assists the human as much as possible in completing the task, which is therefore robust to suboptimal expert policies.

CIRL CIRL focuses on achieving value alignment in human-robot interactions , which allows robot to maintain uncertainty about a goal, rather than trying to optimize a potentially suboptimal goal (Hadfield-Menell et al. 2016). Since human cannot define a perfect goal all at once, human rewards are parameterized in the model, and the reward function is modified through constant observation and interaction. There is some work modeling human in CIRL as a perfectly rational agent (Chan et al. 2019; Shah et al. 2020). Furthermore, some researchers have explored relaxing the rationality assumption in CIRL. (Malik et al. 2018) relaxes the assumption of perfect human rationality by modeling human with random exploratory behavior and using value iteration to derive optimal agent strategies. (Büning, George, and Dimitrakakis 2022) considers suboptimal human decision model, frames human-robot collaborative decision-making as a Stackelberg game, and approximates optimal responses via extended Bayesian Inverse Reinforcement Learning.

Meanwhile, researchers have noted the significance of guidance from human to robot, arguing that people will tend to act in a pedagogical manner, actively choosing actions that provide information about preferences (Fisac et al. 2020). In contrast, our work adds human understanding of robot's belief, incorporating the critical belief alignment process. This helps human make pedagogical behaviors that lead to alignment with robot's belief during decision-making process, facilitating the value alignment effectiveness.

Preliminaries

Cooperative Inverse Reinforcement Learning Game

Cooperative Inverse Reinforcement Learning (CIRL) (Hadfield-Menell et al. 2016) formalizes value alignment as a two-player game between a human H and a robot R. Formally, a CIRL game can be described by a tuple $\langle S, \{A_H, A_R\}, T, \{\Theta, r\}, P_0, \gamma \rangle$ where S denotes the set of possible states of the world; A_H and A_R are the sets of actions available to H and R respectively;

$T : S \times S \times A_H \times A_R \to [0,1]$ represents the discrete transition probabilities to the next state, conditioned on the previous state and the actions of H and R, i.e., $T(s' \mid s, a_H, a_R)$; Θ is the set of static value preference spaces; $r : S \times A_H \times A_R \times \Theta \to \mathbb{R}$ is the parameterized reward function, that is, $r(s, a_H, a_R; \theta)$; $P_0 : S \times \Theta \to [0,1]$ is the probability measure of the initial state and value preference; $\gamma \in [0,1]$ is the discount factor.

The robot cannot directly access the human's true value preference $\theta \in \Theta$, but instead maintains a belief over θ, denoted as b_R. Simultaneously, we explicitly model the human's belief about the robot's belief, defining it as $b_H(b_R)$. We assume that belief and the estimation over belief can be parameterized (that is always feasible if Θ is a finite set). We define Δ_Θ as the corresponding finite-dimensional parameter space, expressed by $b_R \in \Delta_\Theta, b_H(b_R) \in \Delta_\Theta$. Subsequently, let $Q : S \times \Delta_\Theta \times A_H \times A_R \times \Theta \to \mathbb{R}$ denotes the joint action value function of the CIRL game for a given value preference θ.

Human Decision Behavior Modeling

To solve for Q, we first should have an appropriate human decision-making model. As people can often predict a robot's next action if they see the start of it, we assume H can observe a_R before committing to a_H each round. A well-established model in psychology and econometrics is Luce choice rule, which simulates human decisions probabilistically, favoring higher-utility options. In particular, we use a usual case of Luce choice rule, the Boltzmann model of noisy rationality, where selection probability declines exponentially with an option's utility relative to competitors. Here, the utility is Q, reflecting H's expected best outcomes for each a_H. Thus, the probability of H choosing a_H is:

$$\pi_H(a_H \mid s, b_H(b_R), a_R, \theta) \propto \exp\left(\beta Q(s, b_H(b_R), a_H, a_R; \theta)\right) \tag{1}$$

where $\beta > 0$ is H's rationality coefficient. A higher β means H's choices are more concentrated around the optimum. As $\beta \to \infty$, H becomes perfectly rational; as $\beta \to 0$, H becomes indifferent to Q. H exhibits pedagogical behavior because she acts according to Equation (1), which considers how her actions will influence the robot's belief about the goal.

Methodology

In this paper, we propose a Particle Filter-based Hierarchical Dynamic Programming algorithm (PFHDP) to address the value alignment problem in CIRL game. This algorithm constructs a bi-directional alignment framework where the robot aligns with human value preference and human aligns with the robot's belief state. It solves for the optimal strategy of the human-robot system through our hierarchical dynamic programming method. The PFHDP algorithm is primarily divided into two modules: the forward strategy optimization module, which corresponds to R's decision-making process, and the inverse teaching optimization module, corresponding to human decision-making. The overall framework of the algorithm is illustrated in Figure 1.

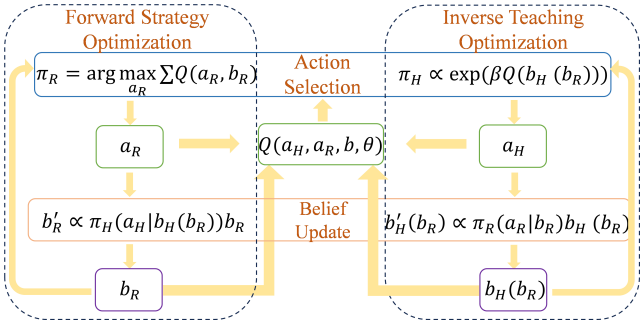


Figure 1: The overall framework of the PFHDP algorithm

Forward Strategy Optimization

This module models R's belief $b_R(\theta)$ about H's value preferences and the strategy $\pi_R(s, b_R)$. Through interaction with H, the belief value of b_R on the true value preference gradually increases. When the algorithm converges, the final strategy π_R is generated. Unlike H, R exhibits no random behavior and is thus described as a perfectly rational agent:

$$\pi_R(s, b_R) = \underset{a_R}{\operatorname{argmax}} \sum_{a_H, \theta} Q(s, b_R, a_H, a_R, \theta) \cdot \pi_H(a_H \mid s, b_H(b_R), a_R, \theta) \tag{2}$$

Since H makes decisions probabilistically based on the joint action value, and at each step a_H follows a_R, making it necessary to consider H's upcoming action, the joint state value is expressed as a probabilistic sum with respect to H's strategies. For any given $\theta \in \Theta$, we have:

$$Q(s, b_R, a_H, a_R, \theta) = r(s, a_H, a_R, \theta) + \mathbb{E}[\gamma Q'(s', b'_R, a'_H, a'_R, \theta)] \tag{3}$$

Here, $\mathbb{E}[\gamma Q'(s', b'_R, a'_H, a'_R, \theta)]$ represents the expected joint action value of the next state s' under the joint action (a'_H, a'_R). However, a_R cannot be directly computed yet because $b_H(b_R)$ is unknown to R. Considering that $b_R(\theta) = b_H(b_R)(\theta)$ when convergence is finally reached, $b_H(b_R)$ is directly substituted with b_R here. Assuming eventual convergence, the results calculated using b_R and $b_H(b_R)$ will be completely consistent. Consequently, R's decision model is transformed as:

$$\pi_R(s, b_R) = \underset{a_R}{\operatorname{argmax}} \sum_{a_H, \theta} Q(s, b_R, a_H, a_R, \theta) \cdot \pi_H(a_H \mid s, b_R, a_R, \theta) \tag{4}$$

Following the decision-making process with the strategy model π_R, we will update the belief b_R. Using the Bayesian update formula, the updated belief b'_R is given by:

$$b'_R(\theta \mid s, b_R, a_R, a_H) \propto \pi_H(a_H \mid s, b_H(b_R), a_R, \theta)\, b_R(\theta) \tag{5}$$

Similarly, substituting $b_H(b_R)$ with b_R, the belief update formula for R simplifies to:

$$b'_R(\theta \mid s, b_R, a_R, a_H) \propto \pi_H(a_H \mid s, b_R, a_R, \theta)\, b_R(\theta) \tag{6}$$

Through this module, R can make decisions and update its belief about the value preference. Upon completion of the iteration, R will derive the corresponding strategy to cooperate with H in collaborative scenarios.

Inverse Teaching Optimization

We now introduce H's decision module. Since π_H has already been presented in the section on human decision mod-

eling, this module will focus primarily on detailing the update method for $b_H(b_R)(\theta)$. This constitutes the paper's core innovative contribution: explicit modeling of $b_H(b_R)(\theta)$ facilitates more realistic implementation of instructional behaviors, guiding R toward belief alignment.

First, we define the Bellman equation for H under the boundedly rational strategy π_H. For any given $\theta \in \Theta$:

$$Q(s, b_H(b_R), a_H, a_R, \theta) = r\left(s, a_H, a_R, \theta\right) + \mathbb{E}\left[\gamma Q'\left(s', b'_H(b_R), a'_H, a'_R, \theta\right)\right] \tag{7}$$

Analogous to Equation (5), the update formula for H's belief estimation $b_H(b_R)(\theta)$ is derived using the Bayesian update formula:

$$b'_H(b_R)\left(\theta \mid s, b_H\left(b_R\right), a_H, a_R\right) \propto \pi_R\left(a_R \mid s, b_R, a_R, \theta\right) b_H(b_R)(\theta) \tag{8}$$

If we follow the approach of the previous module replacing b_R with $b_H(b_R)$ both modules would use biased belief estimates, making convergence hardly achievable. Thus, alternative methods are required to compute $\pi_R\left(a_R \mid s, b_R, a_R, \theta\right)$.

Without compromising the actual results, $\pi_R\left(a_R \mid s, b_R, a_R, \theta\right)$ can be substituted with $P(a_R \mid s, \pi_R)$. Through marginalization of the belief distribution, H's uncertainty regarding b_R can be addressed:

$$P\left(a_R \mid s, \pi_R\right) = \sum_{b_R} P\left(a_R \mid s, b_R, \pi_R\right) \cdot P(b_R \mid \mathcal{H}_t) \tag{9}$$

where $\mathcal{H}_t = \{s^{(0)}, a_R^{(0)}, a_H^{(0)}, \ldots, s^{(t)}\}$ denotes the historical interaction trajectory, and $P(b_R \mid \mathcal{H}_t)$ represents H's estimation of R's possible beliefs.

At this stage, the problem reduces to approximating R's belief distribution based on the historical interaction trajectory. Here, we employ particle filtering to estimate R's belief distribution, leveraging its effectiveness in addressing state estimation within nonlinear and non-Gaussian systems.

First, N_b belief particles $\{b_R^{(0,i)}\}_{i=1}^{N_b}$ are sampled from the prior distribution $P(b_R^{(0)})$. Then, for each time step t, predicted particles are generated via R's belief update formula, conditioned on the historical actions $a_R^{(t-1)}$ and $a_H^{(t-1)}$:

$$\widetilde{b}_R^{(t,i)} \sim P\left(b_R^{(t)} \mid b_R^{(t-1,i)}, a_R^{(t-1)}, a_H^{(t-1)}, s^{(t-1)}\right) \tag{10}$$

Particle weights are computed based on the observed R's action $a_R^{(t)}$:

$$\omega^{(t,i)} = \pi_R\left(a_R^{(t)} \mid s^{(t)}, \widetilde{b}_R^{(t,i)}\right) \omega^{(t-1,i)} \tag{11}$$

Following weight update, resampling of the particle set is necessitated to prevent weight degeneracy, where weights concentrate on a small subset of particles; this process preserves a fixed total number of particles. To begin, the cumulative distribution function is computed:

$$C_i = \sum_{k=1}^{i} \omega^{(t,k)} \tag{12}$$

Given that resampling may fail to cover critical regions of the belief space, leading to sample impoverishment, systematic resampling is employed to perform uniformly spaced sampling. This method enforces replication proportional to weights, ensuring adequate coverage of the belief space. A sample $u \sim \mathcal{U}[0, 1/N_b)$ is drawn from the uniform distribution, and a sequence of sampling points is constructed:

$$u_i = u + \frac{i-1}{N_b} \tag{13}$$

For each u_i, the particle index k satisfying $C_{k-1} \leq u_i < C_k$ is identified, and $b_R^{(t,k)}$ is replicated to the new set, resulting in a new particle set $\{b_R^{(t,i)}\}_{i=1}^{N_b}$. Finally, $P(b_R \mid \mathcal{H}_t)$ is approximated by the particle set as:

$$P\left(b_R \mid \mathcal{H}_t\right) \approx \frac{1}{N_b} \sum_{i=1}^{N_b} \delta\left(b_R - b_R^{(t,i)}\right) \tag{14}$$

where $\delta(\cdot)$ is the Dirac delta function. Furthermore, H's estimation of R's belief is essentially a belief state tracking problem in POMDPs, which can be formally defined by the joint probability:

$$P\left(b_R^{(t)} \mid \mathcal{H}_t\right) = \eta \pi_R\left(a_R^{(t)} \mid s^{(t)}, b_R^{(t)}\right) \int P\left\{b_R^{(t)} \mid b_R^{(t-1)}, a_H^{(t-1)} P\left(b_R^{(t-1)} \mid \mathcal{H}_{t-1}\right)\right\} db_R^{(t-1)} \tag{15}$$

where η is the normalization constant. The update of $b_H(b_R)(\theta)$ can be computed using Equations (8), (9), and (15).

Upon completion of the algorithmic iterations, $b_H(b_R)(\theta) = b_R(\theta)$ indicates that H has successfully aligned with R's belief state. As indicated by Equation (7), H and R's Q-values are also equal at this point, implying indirect convergence in terms of value. Thus, R's optimal strategy can be derived from the joint action value:

$$\pi_R^*\left(s, b_R\right) = \arg\max_{a_R} \textstyle\sum_{a_H, \theta} Q\left(s, b_R, a_H, a_R, \theta\right) \cdot \pi_H\left(a_H \mid s, b_R, a_R, \theta\right) \tag{16}$$

Furthermore, a larger value of $b(\theta)$ indicates that R is more confident that θ is H's true value preference. Thus, when the algorithm converges successfully, the θ that maximizes $b(\theta)$ can be taken as H's true value preference:

$$\theta_{\max} = \arg\max_{\theta \in \Theta} b(\theta) \tag{17}$$

At this point, Equation (16) can be simplified to obtain R's optimal strategy:

$$\pi_R^*\left(s, b_R, \theta_{\max}\right) = \arg\max_{a_R} \textstyle\sum_{a_H} Q\left(s, b_R, a_H, a_R, \theta_{\max}\right) \cdot \pi_H\left(a_H \mid s, b_R, a_R, \theta_{\max}\right) \tag{18}$$

This presupposes that when $b_H(b_R)(\theta) = b_R(\theta)$, R can align with H's true preference $\hat{\theta}$, in which case:

$$\lim_{t \to \infty} b_R(\hat{\theta}) \xrightarrow{p} 1 \tag{19}$$

Thus, $\theta_{\max}$ may serve as a proxy for $\hat{\theta}$. In practice, however, $b_H(b_R)(\theta) = b_R(\theta)$ is merely a necessary condition for Equation (18). To achieve value alignment through belief state alignment, H's pedagogical action sequence $\{a_H^{(t)}\}$ must satisfy information completeness—i.e., incorporate adequate discriminatory signals to distinguish between distinct θ, thereby enabling R to update its belief state accurately. In this case:

$$\forall \theta \neq \hat{\theta}, \mathbb{E}\left[\pi_H\left(a_H \mid \hat{\theta}\right) \cdot \pi_R\left(a_R \mid \theta\right)\right] < \mathbb{E}\left[\pi_H\left(a_H \mid \hat{\theta}\right) \cdot \pi_R\left(a_R \mid \hat{\theta}\right)\right] \tag{20}$$

Algorithm 1: Particle Filter-based Hierarchical Dynamic Programming (PFHDP)

Input: Human H, Robot R, Number of belief particles N_b, Maximum iterations T

Output: Optimal policy π_R^*

Initialize R's belief b_R, H's belief $b_H(b_R)$, and environment state s
Sample N_b belief particles $\left\{b_R^{(0,i)}\right\}_{i=1}^{N_b}$
for $t = 1$ to T **do**
  R selects action a_R according to Equation (4)
  H selects action a_H according to Equation (1)
  H and R receive reward $r(s, a_H, a_R, \theta)$, environment transitions to next state s'
  Generate predicted particles $\left\{\widetilde{b}_R^{(t,i)}\right\}_{i=1}^{N_b}$ using Equation (10)
  Compute particle weights $\left\{\omega^{(t,i)}\right\}_{i=1}^{N_b}$ using Equation (11)
  Perform systematic resampling to obtain updated particle set $\left\{b_R^{(t,i)}\right\}_{i=1}^{N_b}$ using Equations (12) and (13)
  Update R's belief b'_R using Equation (6)
  Update H's belief $b'_H(b_R)$ using Equations (8), (9), and (15)
end for
Estimate maximum value preference $\theta_{\max}$ via Equation (17)
Derive optimal robot policy π_R^* via Equation (18)

As a result, the expected utility of incorrect targets is suppressed, and the system stabilizes at $\hat{\theta}$.

However, due to H's characteristic of bounded rationality, when the rational coefficient β is too small, H gradually degenerates into random decision-making and fails to provide actions that sufficiently distinguish different θ, thereby causing the algorithm to fail to converge and ultimately resulting in alignment failure.

Thus, to ensure that calculations can proceed according to Equation (17), the following condition should be satisfied as much as possible:

$$\forall \theta \neq \hat{\theta}, \sum_{t=1}^{\infty} \log \frac{\pi_H\left(a_H^{(t)} \mid \hat{\theta}\right)}{\pi_H\left(a_H^{(t)} \mid \theta\right)} \rightarrow \infty \tag{21}$$

In other words, humans must continuously provide discriminative actions related to $\hat{\theta}$.

Experiments

Environments

This study adopts the test environments designed by (Hadfield-Menell et al. 2016), namely ChefWorld and GridWorld. These environments differ in the design of value preferences and reward functions, enabling the evaluation of the generalization capability of value alignment algorithms across diverse scenarios.

ChefWorld. ChefWorld is a one-dimensional collaborative decision-making scenario. We assume N possible recipes, M types of ingredients, and a storage limit $\tau = 5$ for raw items. H selects ingredients based on recipes and preferences, while R infers H's target dish by observing H's behaviors and assists in ingredient selection. At each step, both H and R choose an ingredient. A reward of 1 is granted if the selected ingredients match H's target dish.

GridWorld. GridWorld is a two-dimensional collaborative decision-making scenario. We define the side length of the grid as $Size$ and the number of feature points as $Features$. H and R move within this $Size \times Size$ grid to collect rewards. At each step, both agents select a direction of movement (left, right, up, down, or stay). H holds distinct value preferences over this set of feature points. The reward associated with a feature point is +1 or -1, depending on H's preferences. For all other points, the reward is computed as a weighted sum of the squared Euclidean distances to all feature points.

Hyperparameters

In all the experiments, we set the general parameters of the PFHDP algorithm as follows: the discount factor γ is set to 0.9 to balance immediate and future rewards; the number of iteration steps T is 200, providing sufficient time for learning and convergence; the number of belief particles N_b is 100, ensuring an efficient approximation of the belief state distribution while maintaining computational efficiency; and the rational coefficient β is 1, moderately regulating the rationality of human decisions, balancing rationality and randomness in human behavior.

Baseline Algorithms and Evaluation Metrics

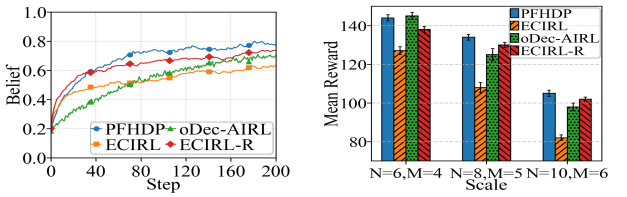
To evaluate the effectiveness of the PFHDP algorithm, it is compared against three baseline algorithms:

ECIRL(Malik et al. 2018). A classical cooperative inverse reinforcement learning algorithm that models humans as perfectly rational decision-makers. It formulates CIRL as a Coordinator-POMDP and solves the problem based on Monte Carlo Tree Search.

oDec-AIRL(Suresh et al. 2024). oDec-AIRL is a multi-agent inverse reinforcement learning algorithm. It addresses multi-agent coordination problems using decentralized control methods in Dec-MDP environments.

ECIRL-R(Malik et al. 2018). A variant of ECIRL that models human as boundedly rational yet neglects human's estimation of the robot's value preference beliefs.

In ChefWorld, the evaluation metrics are the belief in the true value preference and the mean cumulative reward. The belief in the true value preference is reflected by $b_R(\hat{\theta})$, which represents the confidence of R's estimated value preference in aligning with H's true value preference. This metric serves as a standard to measure whether R successfully aligns with H. The mean cumulative reward can be interpreted as the number of dishes successfully completed. A higher cumulative reward indicates a better alignment between R and H's value preferences.



(a) Evolution of belief in true value preference over interactions

(b) Experimental results of cumulative rewards across different environment scales

Figure 2: ChefWorld experimental results

In GridWorld, the evaluation metrics are the difference between the true and estimated values and the mean cumulative reward. The difference between the true and estimated values is calculated as the Euclidean distance between θ and $\hat{\theta}$, weighted by the belief distribution $b_R(\theta)$. This is expressed as $\text{Difference}(\theta) = \sum_{\theta \in \Theta} b_R(\theta)|\theta - \hat{\theta}|^2$. A smaller difference indicates that R's value preferences are closer to those of H, making it a criterion for assessing the alignment success of R with H. The mean cumulative reward varies based on H's different value preferences, with different locations offering different rewards. A higher cumulative reward indicates that R has accurately grasped H's value preferences for different locations, resulting in a better alignment outcome.

Experimental Results

This section analyzes the performance of the PFHDP algorithm and the aforementioned baseline algorithms across two experimental environments, concluding with an analysis of the results from ablation experiments on the rationality coefficient β within the algorithm's inverse teaching optimization module. All experiments were repeated 10 times, with the mean value computed to ensure the stability and reliability of the results. Furthermore, since the human strategy employed is boundedly rational, humans make random decisions with a certain probability based on the value function. When the algorithm is designed with a joint action value function, such a function serves as the value function; otherwise, the value function is by default H's own, with R's actions not factored in.

ChefWorld In interactions, R updates its beliefs over all possible value preferences. We thus analyze the evolution of R's belief in H's true value preference $\hat{\theta}$ as interactions proceed. Figure 2(a) shows the evolution of R's belief $b(\hat{\theta})$.

When humans make irrational decisions, the alignment of $b(\hat{\theta})$ exhibits a clear upper bound. This is because H may choose actions that do not maximize the value function, thereby misleading R's inference of H's true preferences. In ECIRL, if H's action does not maximize the value, the algorithm directly discards the current value estimate, potentially missing the correct update and leading to a marked decline in how well $\hat{\theta}$ is aligned. While oDec-AIRL achieves inferior alignment compared to PFHDP and ECIRL-R, it outperforms ECIRL. This can probably be attributed to oDec-AIRL's lack of explicit modeling of H's strategy (which avoids errors from mis-specification) and its implicit inference of H's policy through multi-agent reinforcement learning. PFHDP and ECIRL-R, through accurate modeling of bounded rational human decisions, yield higher $b(\hat{\theta})$ alignment. Notably, PFHDP's additional consideration of H's perception of R's beliefs enables more precise modeling of interactive decision-making and belief updates, resulting in superior alignment.

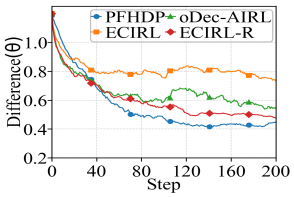
We use mean cumulative reward to further evaluate value alignment across scaled scenarios. Results are shown in Figure 2(b). Evidently, ECIRL performs poorly in aligning $b(\hat{\theta})$ and consequently yields lower cumulative rewards, as it neglects the stochasticity and pedagogical nature of human decisions. oDec-AIRL, which implicitly models human strategies via multi-agent learning, is less sensitive to variations in H's policy. In contrast, PFHDP and ECIRL-R adopt more reasonable human decision models. Specifically, PFHDP incorporates $b_H(b_R)$ into the joint action value calculation, enabling H to guide R's alignment more effectively by estimating R's current belief. This leads to R's faster convergence to H's true value. The superiority of PFHDP is validated in the more complex scenarios($N = 8, M = 5$ and $N = 10, M = 6$), where it achieves the highest mean cumulative reward.

GridWorld This section presents experiments in the more complex GridWorld environment. As shown in Figure 3(a), we analyze how the difference between R's estimated and true values evolves over interaction steps.

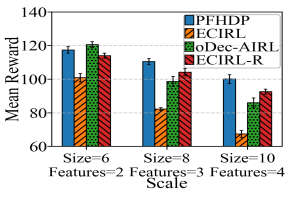
Notably, oscillations during the interaction process were significantly amplified compared to the ChefWorld experiments, indicating that achieving value alignment becomes more challenging in complex scenarios. ECIRL, which models human as a perfectly rational agent, struggles when H exhibits stochastic behavior, it erroneously updates its value estimates, thereby leading to resulting in degraded alignment performance.

In contrast, oDec-AIRL, ECIRL-R, and PFHDP demonstrate better performance. This can be attributed to their more refined human modeling frameworks that adeptly incorporate bounded rationality considerations. Specifically, oDec-AIRL implicitly models human strategies through multi-agent interactions, while ECIRL-R accurately captures H's boundedly rational decision-making but overlooks H's estimation of R's beliefs. PFHDP, however, explicitly incorporates $b_H(b_R)$ into its framework, enabling more precise modeling of interactive belief updates and thus achieving superior alignment.

We further investigated the impact of environmental scale on value alignment performance, with mean cumulative rewards presented in Figure 3(b). As the figure illustrates, all algorithms exhibit a slight decline in mean cumulative reward as the environment scale increases. This trend is likely attributable to the expanded state and action spaces, which induce greater instability in belief updates. Furthermore, the reward rankings of the algorithms are broadly consistent with the belief estimation discrepancies shown in Figure 3(a); specifically, smaller discrepancies between the true and

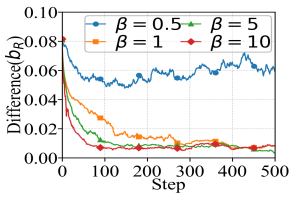


(a) Difference between estimated and true values over interaction steps

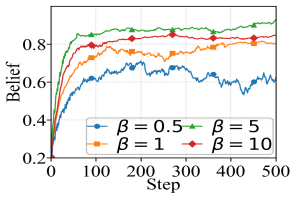


(b) Cumulative rewards across varying environment scales

Figure 3: GridWorld experimental results



(a) Belief alignment performance under varied rationality coefficients



(b) True value belief over interaction steps under different rationality coefficients

Figure 4: Ablation results on rationality coefficient β

estimated values correlate with higher rewards.

Ablation Studies

In the inverse teaching optimization module, this paper analytically shows that the rationality coefficient β in the human decision model influences a_H, thereby altering R's belief distribution updates and ultimately impacting alignment performance. However, this finding lacks experimental validation. Thus, this section investigates the effect of varying rationality coefficients β on actual alignment.

Four distinct values of β are tested: 0.5, 1, 5, and 10. When $\beta = 0.5$, H tends to make random decisions with high probability. For $\beta = 1$, H follows the PFHDP algorithm. At $\beta = 5$, H exhibits a stronger tendency toward rational decisions. When $\beta = 10$, H approaches perfectly rational decision-making, with a high likelihood of selecting the action maximizing joint action value.

The experiment, designed to validate the hypothesis in the inverse teaching optimization module (that $b_H(b_R) = b_R$ does not guarantee R's alignment with H's values), is conducted in the ChefWorld environment with two evaluation metrics: the first is the human-robot belief discrepancy, defined as $\text{Difference}(b_R) = \|\theta_{b_H(b_R)} - \theta_{b_R}\|_2$, which quantifies the belief divergence between H and R; The second is $b_R(\hat{\theta})$, R's belief in the true value, which should approach 1 with successful alignment to H's preference values. We use a 500-step interaction horizon to ensure algorithmic convergence. Concurrently, to avoid initial $b_H(b_R) = b_R$, R's belief is initialized uniformly, while H's belief is randomly sampled with a unit-sum constraint.

Figure 4(a) presents experimental results on how β influences belief alignment. For $\beta = 0.5$, the algorithm oscillates persistently: a sufficiently small rationality coefficient causes H to act randomly, making it difficult for R to extract useful information for belief updates. Erroneous updates, in turn, further prevent R from signaling its current beliefs to H, resulting in failed alignment.

For other β values, algorithms converge to comparable levels but differ in convergence speed. $\beta = 10$ converges approximately at Step 80, $\beta = 5$ at Step 140, and $\beta = 1$ at Step 260. This highlights the critical role of H's rationality in enabling convergence. Meanwhile, since the algorithms with three distinct rationality coefficients ultimately converge to the same level, we infer belief alignment will eventually be achieved, if H can continuously provide valid behavioral information relevant to its own values.

To further validate the hypothesis that $b_H(b_R) = b_R$ does not guarantee value alignment, and to compare the influence of β on value alignment, we analyze R's belief regarding the true value across interaction steps for varying β. The results are shown in Figure 4(b). At $\beta = 0.5$, the necessary condition $b_H(b_R) = b_R$ for value alignment is not met, causing $b_R(\hat{\theta})$ to exhibit persistent oscillations without convergence. Conversely, although the algorithms corresponding to other rationality coefficients have achieved alignment in belief states (i.e., $b_H(b_R) = b_R$), the degree to which R aligns with H's true value preference $\hat{\theta}$ differs. This validates the hypothesis put forward in the inverse teaching optimization module: $b_H(b_R) = b_R$ is not a sufficient condition for value alignment. Notably, while $\beta = 10$ yields a higher degree of belief alignment, its value alignment degree is inferior to that of $\beta = 5$. This suggests that highly rational humans, while more likely to achieve belief alignment, are less adept at producing pedagogical actions, leading to reduced value alignment. Thus, we infer that the rationality coefficient must be constrained within a certain range.

Discussion and Conclusion

This paper proposes a Particle Filter-based Hierarchical Dynamic Programming algorithm (PFHDP) to address the value alignment problem in human-robot collaboration. By incorporating the processes of human understanding of the robot's belief and belief alignment, the algorithm breaks away from the idealized assumption in traditional CIRL that human can directly infer the robot's belief. Thus, we provide a more practical solution for human-robot collaboration in complex scenarios. Experimental results in two typical collaborative scenarios demonstrate that, PFHDP, by incorporating human understanding of the robot's belief, achieves superior enhancement of human-robot belief consistency compared to baselines. This superior belief consistency translates to a improvement in value alignment performance. Ablation studies further confirm that when human rationality coefficients are within a reasonable range, human behaviors can provide more effective signals for updating the robot's belief, promoting value alignment. Future work will involve integrating deep networks to enhance our algorithm's scalability for larger-scale problems, as well as extending it to multi-agent systems and partially observable environments. Furthermore, we will investigate how to fuse other communication modalities to achieve value alignment.

Acknowledgments

This paper was supported by the Fundamental Research Funds for the Central Universities (Grant No. 104972025KFYrs0055, Grant No. 104972025KFYjc0033). The authors also acknowledge Beijing PARATERA Technology Co., LTD for providing high-performance and AI computing resources for contributing to the research results reported within this paper.

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1.

Bucknall, B. S.; and Dori-Hacohen, S. 2022. Current and near-term AI as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 119–129.

Büning, T. K.; George, A.-M.; and Dimitrakakis, C. 2022. Interactive inverse reinforcement learning for cooperative games. In *International Conference on Machine Learning*, 2393–2413. PMLR.

Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Huang, F.; Manocha, D.; Bedi, A.; and Wang, M. 2024. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Chan, L.; Hadfield-Menell, D.; Srinivasa, S.; and Dragan, A. 2019. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 354–363. IEEE.

Chen, Y.; Wu, T.; Wang, S.; Feng, X.; Jiang, J.; Lu, Z.; McAleer, S.; Dong, H.; Zhu, S.-C.; and Yang, Y. 2022. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 5150–5163.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Fisac, J. F.; Gates, M. A.; Hamrick, J. B.; Liu, C.; Hadfield-Menell, D.; Palaniappan, M.; Malik, D.; Sastry, S. S.; Griffiths, T. L.; and Dragan, A. D. 2020. Pragmatic-pedagogic value alignment. In *Robotics research: the 18th international symposium Isrr*, 49–57. Springer.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Gabriel, I.; and Ghazavi, V. 2022. The challenge of value alignment. *The Oxford handbook of digital ethics*, 336–355.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2016. Cooperative inverse reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3916–3924. Curran Associates Inc.

Inkulu, A. K.; Bahubalendruni, M. R.; Dara, A.; and K, S. 2022. 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*, 49(2): 226–239.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Malik, D.; Palaniappan, M.; Fisac, J.; Hadfield-Menell, D.; Russell, S.; and Dragan, A. 2018. An efficient, generalized bellman update for cooperative inverse reinforcement learning. In *International Conference on Machine Learning*, 3394–3402. PMLR.

Ord, T. 2020. *The precipice: Existential risk and the future of humanity*. Hachette UK.

Park, C.; Liu, M.; Zhang, K.; and Ozdaglar, A. 2024a. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2.

Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2024b. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, 13387–13434.

Semeraro, F.; Griffiths, A.; and Cangelosi, A. 2023. Human–robot collaboration and machine learning: A systematic review of recent research. *Robotics and Computer-Integrated Manufacturing*, 79: 102432.

Shah, R.; Freire, P.; Alex, N.; Freedman, R.; Krasheninnikov, D.; Chan, L.; Dennis, M. D.; Abbeel, P.; Dragan, A.; and Russell, S. 2020. Benefits of assistance over reward learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587): 484–489.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33: 3008–3021.

Suresh, P. S.; Jain, S.; Doshi, P.; and Romeres, D. 2024. Open human-robot collaboration using decentralized inverse reinforcement learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7092–7098. IEEE.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *nature*, 575(7782): 350–354.

Wu, J.; Ouyang, L.; Ziegler, D. M.; Stiennon, N.; Lowe, R.; Leike, J.; and Christiano, P. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Wu, Y.-H.; Charoenphakdee, N.; Bao, H.; Tangkaratt, V.; and Sugiyama, M. 2019. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, 6818–6827. PMLR.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 1–45.

Zhou, M.; Luo, J.; Villella, J.; Yang, Y.; Rusu, D.; Miao, J.; Zhang, W.; Alban, M.; Fadakar, I.; Chen, Z.; et al. 2020. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. *arXiv preprint arXiv:2010.09776*.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; Dey, A. K.; et al. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, 1433–1438. Chicago, IL, USA.