Learning Human Preferences Over a Human-Robot Collaboration Based on Explicit and Implicit Human Feedback

Kate Candon
Yale University
New Haven, CT, USA
kate.candon@yale.edu

Qiping Zhang
Yale University
New Haven, CT, USA
qiping.zhang@yale.edu

Alexander Lew
Yale University
New Haven, CT, USA
a.lew@yale.edu

Houston Claure
Yale University
New Haven, CT, USA
houston.claure@yale.edu

Lena Qian
Yale University
New Haven, CT, USA
lena.qian@yale.edu

Alyssa Quarles
Yale University
New Haven, CT, USA
alyssa.quarles@yale.edu

Chayan Sarkar
Robotics & Autonomous Systems, TCS Research
New Dehli, India
sarkar.chayan@tcs.com

Marynel Vázquez
Yale University
New Haven, CT, USA
marynel.vazquez@yale.edu

Abstract

There is significant interest in enabling robots to learn to perform tasks directly from interactions with non-expert users. Typically, a human serves as a teacher whose only task is to provide feedback to a robot learner. However, in real-world human-robot collaborations, the human often assists with the task while also offering feedback. Our key insight is that we can extract additional, implicit feedback from the human's actions in the collaboration to augment the robot learning process. Under the assumption of fixed-role assignments, we first propose to formalize human preferences over a human-robot collaboration as a shared set of parameters encoding alignment between two reward functions: one that drives human behavior, and another that should direct robot behavior. This allows us to extract implicit feedback from an interaction by reasoning about the human's actions in the task as actions that reveal the human's preferences. Then, we combine this implicit feedback with traditional explicit human feedback to facilitate estimating the human's preferences. We evaluated our proposed approach for Preference learning from Implicit and Explicit feedback (PIE) in simulations and with real users in a cooking scenario. Our simulation results indicate that combining multiple modalities of human feedback improves a robot's ability to estimate human preferences over the collaboration, with a similar trend observed in real-world evaluations. Practically, these findings highlight a promising direction for more seamless interactions by enabling robots to adapt to a user's preference model more quickly, thereby reducing the amount of time a person must spend teaching a robot collaborator.

CCS Concepts

• **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**.

Keywords

human-robot collaboration, implicit feedback, preference learning

ACM Reference Format:
Kate Candon, Qiping Zhang, Alexander Lew, Houston Claure, Lena Qian, Alyssa Quarles, Chayan Sarkar, and Marynel Vázquez. 2026. Learning Human Preferences Over a Human-Robot Collaboration Based on Explicit and Implicit Human Feedback. In *Proceedings of HRI (HRI '26)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Advances in robotics hardware and physical manipulation capabilities are fueling a growing interest in enabling robots to adapt to human users, so that robots and humans can solve tasks collaboratively. For example, robots can collaborate with humans to place and seal screws [29], assemble objects from a collection of parts [20, 46], and cook together [18], etc. For these collaborations to be effective, robots must be able to learn how humans want them to collaborate, in a way that is natural and intuitive for the humans.

In interactive robot learning, it is common for the human to serve as a dedicated teacher, whose only role is to provide explicit feedback to the robot (e.g., [17, 27, 42]). While effective in controlled interaction setups, this paradigm falls short in real-world human-robot collaborations, where the human is typically engaged in the task itself and does not focus solely on teaching. Expecting people to provide continuous, explicit feedback to a robot is not only impractical but can also lead to frustration and disengagement [49].

In contrast, when humans collaborate with one another, an important amount of learning happens implicitly. People observe each other's actions and infer preferences, all without formal teaching. Consequently, we propose that robots should similarly learn from the implicit signals embedded in a collaborative interaction in addition to learning from explicit feedback.

We contribute an approach for **Preference learning from Implicit and Explicit feedback (PIE)** in collaborative human-robot interactions with fixed role assignments. The key novelty of our work lies in framing human preferences over the human-robot collaboration that a person experiences – instead of over the robot's behavior only, as is typical in prior interactive robot learning work. We formalize the human preferences as a shared set of parameters encoding alignment between the human's behavior and the desired robot's behavior. Then, by modeling human actions in a task as

HRI '26, Edinburgh, Scotland, UK

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

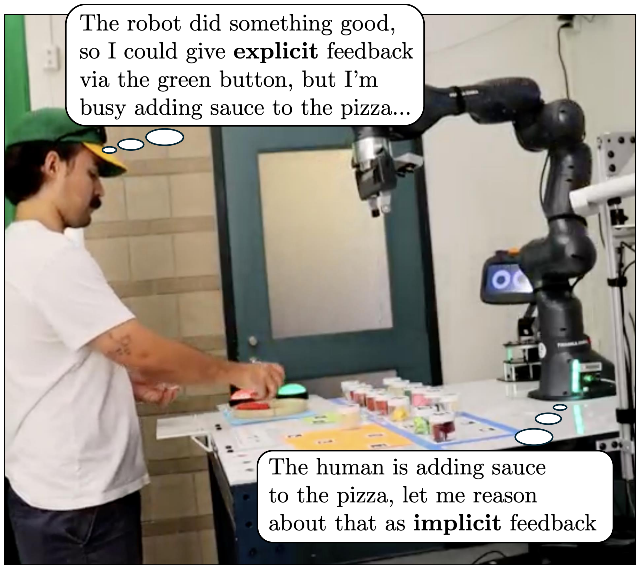

Figure 1: We study the problem of learning human preferences over a human-robot collaboration. Preference learning commonly assumes that the human's only task is to teach a robot via explicit feedback. In addition to explicit feedback, we propose to leverage the human's actions as implicit feedback that can help the robot learn preferences over the collaboration. We evaluate this idea in a laboratory, where participants assemble pizzas with a robot and can give it explicit feedback by pressing physical buttons on the workspace.

being (approximately) optimal with respect to the human's preferences, we show how a robot can treat the human's behavior as implicit feedback during the collaboration, effectively "listening" to what the human's actions reveal about the human's underlying preferences. Finally, we combine this implicit feedback with more traditional explicit human feedback to enable a robot to quickly estimate human preferences during an ongoing collaboration. We evaluate our proposed approach in simulations and with real users in a cooking collaboration. Our simulation results confirm that combining multiple modalities of human feedback improves a robot's ability to estimate human preferences, with a similar trend observed in real-world evaluations.

In summary, this paper has three main contributions. First, we propose a novel formulation for preference learning over human-robot collaborations, which leverages *both* implicit and explicit human feedback. Second, we systematically investigate the effectiveness of our approach for preference learning (PIE) in simulations and in the real-world. Our experiments consider varying cooking tasks, assumptions about the rationality of human behavior, and different preferences. Lastly, we open-source our implementation to facilitate future replication and benchmarking efforts.[1]

[1]Link to be added upon acceptance.

2 Related Work

Physical Human-Robot Collaboration (HRC). Collaboration involves work on a shared space towards common goals. HRC often leverages the complementary strengths of both humans (e.g., dexterity, judgment, adaptability) and robots (e.g., strength, precision, repeatability) [24, 43], typically resulting in specialized roles for the interactants. Common applications include manufacturing [35] and assembly [59], although recent advancements are enabling collaborations in less structured environments where user preferences can have an important impact on robot adoption, like hospitals [51], homes [21, 28], and hospitality environments [31, 45]. In our work, we study human-robot interactions in a cooking scenario, a popular setup for studying collaboration (e.g., see [7, 40, 47]) due to its relevance to the service and assistive robotics domains.

Human Preference Learning in HRI. Enabling robots to understand and align their behavior with human preferences can result in enhanced efficiency and safety [36] as well as higher user satisfaction [1] during human-robot interactions. Humans can teach robots via a variety of explicit feedback modalities, including demonstrations [2], corrections [33], rankings [38] and preferences [54]. Learning can be framed under unified learning frameworks that allow extracting information from different feedback modalities [17, 26], potentially resulting in better and faster preference learning. Specifically, we leverage the INQUIRE framework [17] in our work, which frames human feedback in terms of inferences about accepted robot behavior and rejected behavior. However, instead of investigating interactions where a human serves as a dedicated teacher, whose only task is to teach a robot, we investigate collaborations where a human teaches a robot while also taking task-relevant actions.

Implicit Human Feedback in HRI. Robots traditionally learn from explicit human feedback, like demonstrations[2, 13] or evaluative feedback[27], where the implications of the feedback for robot learning are clear. But relying solely on these methods can be burdensome for the human, particularly within a collaboration where the person is also focused on task execution [8]. Thus, researchers have increasingly explored implicit feedback – signals that require interpretation because they are not necessarily intended for teaching the robot, but which nonetheless convey information about the human's state, intentions, or preferences [16]. Prior work has utilized various implicit channels, including non-verbal cues like gaze for inferring attention or target objects [37], facial expressions for gauging affect or understanding [15, 22, 56], and physiological signals (e.g., EEG, GSR) for emotion estimation in robot learning [25, 53]. Others have inferred user states or assessed robot performance based on interaction dynamics, timing, or hesitations [52, 57]. Relatedly, Learning from Observation (LfO) focuses on robots learning tasks by watching human actions [22], but typically aims for skill acquisition or goal inference rather than understanding preferences about the interaction itself. Our work is inspired by this body of research, but takes a distinct approach by treating the human's task-oriented actions as a rich source of implicit feedback, which conveys the human's preferences over the collaboration.

3 Problem Setup: Learning Preferences Over a Human-Robot Collaboration

We consider collaborative interactions in which a human H and a robot R work together to complete a physical task and where each has a specific role in the collaboration. For example, in our pizza-making scenario of Figure 1, the robot may pass ingredients to the human from a storage area, and the human may use the ingredients to assemble a desired pizza.

At a time-step t, the human and robot observe a given state $\mathbf{s}$ and take simultaneous high-level actions a_R and a_H. We model these high-level actions as parameterized actions. For example, two high-level actions for the robot may be `pick(<ingredient>, <location>)` and `place(<ingredient>, <location>)` in Figure 1.

Collaborations are characterized by teammates trying to maximize a shared reward [3, 5, 48]. It is typical for the reward to be based on task success only [5, 11] or human preferences that encapsulate the desired task outcome [17, 32, 34]. However, in many interactions, it can be helpful to explicitly model both the task goal and human preferences, e.g., because not taking action toward the task goal is worse than violating preferences [39], or the task goal is public information while the human preferences are not [58]. We assume that both situations are true in our work, so we propose to define the shared reward as a combined reward:

$$R(\mathbf{s}, a_R, a_H) = R^{\text{goal}}(\mathbf{s}, a_R, a_H) + \gamma R^{\text{pref}}(\mathbf{s}, a_R, a_H) \tag{1}$$

where γ is a parameter that controls the relative importance of the two rewards. This reward formulation is similar to Zhao et al. [58] preference learning setup, but we assume that human preferences are over the human-robot collaboration (not just the human's contribution to the task) and the robot must follow these preferences (rather than having its own individual reward). As we show in this work, our framing enables the robot to leverage observed human actions as implicit feedback for preference learning.

Goal Reward: Motivated by the fixed-role assignments, we propose to decompose the goal reward in eq. (1) into two components, one for the robot and one for the human:

$$R^{\text{goal}}(\mathbf{s}, a_R, a_H) = R_R^{\text{goal}}(\mathbf{s}, a_R) + R_H^{\text{goal}}(\mathbf{s}, a_H) \tag{2}$$

Preference Reward: We assume the that preference reward, R^{pref} in eq. (1), does not conflict with R^{goal}. In addition, we assume that the human preferences for the team members are aligned with each other, such that the preference reward can also be decomposed into two terms parameterized by the same weights $\mathbf{w}$:

$$\begin{aligned} R^{\text{pref}}(\mathbf{s}, a_R, a_H) &= R_R^{\text{pref}}(\mathbf{s}, a_R) + R_H^{\text{pref}}(\mathbf{s}, a_H) \\ &= \mathbf{w}^\intercal \phi_R(\mathbf{s}, a_R) + \mathbf{w}^\intercal \phi_H(\mathbf{s}, a_H) \\ &= \mathbf{w}^\intercal (\phi_R(\mathbf{s}, a_R) + \phi_H(\mathbf{s}, a_H)) \end{aligned} \tag{3}$$

We implement the preference reward as a linear function of features of the state (encoded via ϕ_R and ϕ_H) to keep the reward interpretable in this work. While this setup is common in the preference learning literature (e.g., [17, 26]), future work could investigate ways to relax this assumption (e.g., via more complex reward models implemented as neural networks [13, 23]).

Taken together, the above assumptions mean that if the robot and human act rationally, they maximize their individual rewards:

$$R_R(\mathbf{s}, a_R) = R_R^{\text{goal}}(\mathbf{s}, a_R) + \gamma R_R^{\text{pref}}(\mathbf{s}, a_R) \tag{4}$$

$$R_H(\mathbf{s}, a_H) = R_H^{\text{goal}}(\mathbf{s}, a_H) + \gamma R_H^{\text{pref}}(\mathbf{s}, a_H) \tag{5}$$

The Robot's Learning Objective: In this work, we assume that the robot knows the goal reward and the features of the state that may matter for the human preferences over the collaboration (ϕ_R and ϕ_H in eq. (3)); however, the robot does not know the weights $\mathbf{w}$ that parameterize the preference reward functions R_H^{pref} and R_R^{pref}. Thus, the goal of the robot is to estimate the weights $\mathbf{w}$ based on human feedback gathered *during* the human-robot collaboration.

4 Preference Learning from Implicit and Explicit Feedback (PIE)

We propose that robots estimate a human's preferences for their collaboration per eq. (3) based on both *explicit* feedback provided by the human about the robot's behavior as well as *implicit* feedback provided by the human's own actions in the task. Explicit feedback corresponds to binary evaluative feedback in our work, which is implemented in our real-world evaluation via physical button presses (e.g., similar to [41]). The main assumption for the implicit human feedback is that the human's actions are driven by their reward, as in eq. (5).

Algorithm 1 describes PIE, our proposed approach for preference learning, considering a given interaction step in a human-robot collaboration. First, the robot observes the current state of the world, takes action according to some policy π_R, and sees how the human behaves (lines 1-3 in Alg. 1). Every action that the human takes is then interpreted as implicit feedback for preference learning and the implications of the human's behavior are stored in a feedback set (lines 4-6). Each element in this feedback set includes a pair of accepted behavior (f^+) and rejected behavior (f^-) along with their associated state. When the human chooses to give explicit feedback, the implications of this feedback are also stored in the feedback set (lines 11-14). Finally, a non-parametric belief over preferences $\mathbf{W}$, implemented via M weight samples, is computed using the feedback set (lines 15-17). Our approach to preference learning builds on the INQUIRE formalism [17] for combining various types of feedback during interactive robot learning.

One key difference between INQUIRE [17] and our proposed approach, PIE, is that we consider implicit human feedback, not just explicit feedback. This results in different implications for preference learning. For explicit feedback directed intentionally from the human to the robot, the implications of the feedback are set in relation to the robot's behavior. However, for the implicit feedback, we instead define the implications in relation to human behavior because this feedback is a direct consequence of the human's own actions. Next, we describe in more detail how we interpret the feedback for preference learning and estimate the preference belief.

4.1 Implication of Explicit Human Feedback

We specifically consider explicit feedback in the form of binary feedback.When a robot takes a high-level action a_R in a state $\mathbf{s}$ (line 2 in Alg. 1), the human may choose to indicate whether the

Algorithm 1: Learning from Explicit & Implicit Feedback (PIE)

Input: Prior belief over pref. weights $\mathbf{W} = \{\mathbf{w}^i\}_{i=1}^M$, and prior feedback set $\mathbf{F}$ with pairs of acceptable (f^+) and rejected (f^-) behaviors in prior states
Output: Updated belief $\mathbf{W}$, and updated feedback set $\mathbf{F}$

// Interact with the environment
1 Observe current state $\mathbf{s}$;
2 Robot takes action $a_R \leftarrow \pi_R(\mathbf{s})$;
3 Observe current human action a_H;
// Store implicit human feedback
4 $f^+_{imp} \leftarrow a_H$; /* Observed a_H aligns with pref. */
5 $f^-_{imp} \leftarrow \mathcal{A}_H(\mathbf{s}) \setminus \{a_H\}$;
6 $\mathbf{F} \leftarrow \mathbf{F} \cup \{(f^+_{imp}, f^-_{imp}, \mathbf{s})\}$;
// Store explicit feedback on button press
7 **if** *human indicates acceptable robot behavior* **then**
8   $f^+_{exp} \leftarrow a_R$; /* Robot action a_R aligns with pref. */
9   $f^-_{exp} \leftarrow \mathcal{A}_R(\mathbf{s}) \setminus \{a_R\}$;
10   $\mathbf{F} \leftarrow \mathbf{F} \cup \{(f^+_{exp}, f^-_{exp}, \mathbf{s})\}$;
11 **else if** *human indicates unacceptable robot behavior* **then**
12   $f^+_{exp} \leftarrow \mathcal{A}_R(\mathbf{s}) \setminus \{a_R\}$; /* Other robot actions are better aligned with pref. than a_R */
13   $f^-_{exp} \leftarrow a_R$;
14   $\mathbf{F} \leftarrow \mathbf{F} \cup \{(f^+_{exp}, f^-_{exp}, \mathbf{s})\}$;
// Update non-parametric belief over preference weights
15 **foreach** $\mathbf{w}^i \in \mathbf{W}$ **do**
  // MLE via gradient ascent, starting from prior $\mathbf{w}^i$
16   $\mathbf{w}^i \leftarrow$ `optimize_w_to_maximize_likelihood`$(\mathbf{F}, \mathbf{w}^i)$;
17 **end**
18 **return** $\mathbf{W}$, $\mathbf{F}$;

behavior is acceptable or not. We define the implication of binary feedback in two ways, based on the specific human choice. If the human indicates that the robot behavior is acceptable (lines 7-10 in Alg. 1), the robot's action is added to the accepted behavior set $f^+_{exp} = \{a_R\}$, and all other viable robot actions in the state are added to the rejected behavior set $f^-_{exp} = \mathcal{A}_R(\mathbf{s}) \setminus \{a_R\}$. However, if the human indicates unacceptable robot behavior (lines 11-14 in Alg. 1), the implication is the opposite. In this case, the accepted behavior set includes viable robot actions in the state that the robot did not take ($f^+_{exp} = \mathcal{A}_R(\mathbf{s}) \setminus \{a_R\}$), and the rejected behavior set includes only the action that was taken by the robot ($f^-_{exp} = \{a_R\}$).

Our formulation for the implication of binary feedback follows INQUIRE [17], with the exception that we do not study active preference learning in this work, so the robot does not pose questions to the human for which feedback is received in return. Rather, the human may choose to give or not give explicit feedback at any point during the collaboration. Because explicit feedback is potentially sparse and prior findings show that people may reduce the amount of feedback that they give to a robot during collaborations [9], we propose to also consider implicit feedback.

4.2 Implication of Implicit Human Feedback

A key novelty of our work is framing the human's actions as a source for implicit feedback about the human preferences over the collaboration. These preferences are encoded in the weight vector $\mathbf{w}$ shared by $R_H^{\text{pref}}(\cdot)$ and $R_R^{\text{pref}}(\cdot)$, per eq. (3). If during the collaboration, the human takes actions that are approximately optimal with respect to the human's reward $R_H(\cdot)$ in eq. (5), then the human's behavior will leak information about $\mathbf{w}$ through $R_H^{\text{pref}}(\cdot)$.

Formally, we assume that the human takes action on a given state $\mathbf{s}$ following a Boltzmann rational policy:

$$P(a_H|\mathbf{s}) \propto \exp(\beta_H R_H(\mathbf{s}, a_H)) \tag{6}$$

where β_H controls how rational the human's actions are. This model of human behavior is common in economics [44], psychology [4], and preference learning [26]. Importantly, eq. (6) results in myopic high-level decision-making because the human is said to take actions based on the reward of the current state. We find that this myopic assumption is reasonable for preference learning over high-level actions that span multiple time-steps during the collaboration and when the human's reward is not sparse. However, for sparse rewards, this formulation would need to be adapted to a Boltzmann policy based on expected future rewards (e.g., via Q-values [6]). We discuss this future work in Section 7.

Equipped with a model of human actions per eq. (6), we can formulate feedback based on these actions for preference learning. At a given state $\mathbf{s}$, we consider the set of valid actions of the human, $\mathcal{A}_H(\mathbf{s})$, as the set of possible choices that the human has for implicit feedback in that state. Hence, when the human takes action $a_H \in \mathcal{A}_H(\mathbf{s})$ at a given point in the collaboration, the implication of that choice is that that action is accepted behavior ($f^+_{imp} = \{a_H\}$) and that other viable human actions are rejected behavior ($f^-_{imp} = \mathcal{A}(\mathbf{s}) \setminus \{a_H\}$). The implication of implicit feedback from human actions is outlined in lines 4-6 of Algorithm 1.

By reframing human actions as implicit feedback with the implication described previously, a robot can gain information about the preference weights $\mathbf{w}$ potentially *all throughout* a collaboration, without having to continuously query the human for feedback. Mathematically, the implication that we propose for implicit human feedback is equivalent to the implication for human demonstrations of robot behavior used in INQUIRE [17]. However, our implication defines sets of accepted and rejected *human* behavior, rather than sets of robot behavior, so the implications are conceptually different.

4.3 Estimating Belief Over Preference Weights

Whenever the robot receives human feedback, it stores the implications of the feedback ($f^+_{\mathrm{m}}, f^-_{\mathrm{m}}$) in a cumulative feedback set $\mathbf{F}$, where m is the modality (explicit or implicit) of the feedback, alongside the current state $\mathbf{s}$ of the interaction when the feedback was received (see lines 6, 10, and 14 in Algorithm 1). In PIE, the modality m of the feedback is critical because it dictates the perspective from which the robot should reason about the implications of the feedback.

The robot's objective consists of estimating the preference weights that maximize the likelihood of the accepted behavior implied by the feedback in $\mathbf{F}$. Specifically, the likelihood is:

$$\mathcal{L}(\mathbf{w}) = \prod_{(f_m^+, f_m^-, \mathbf{s}) \in \mathbf{F}} P(f_m^+|\mathbf{w}) = \prod_{(f_m^+, f_m^-, \mathbf{s}) \in \mathbf{F}} \frac{\sum_{a \in f_m^+} B_m(\mathbf{s}, a)}{\sum_{a \in f_m^+ \cup f_m^-} B_m(\mathbf{s}, a)} \tag{7}$$

where:

$$B_m(\mathbf{s}, a) = e^{\hat{\beta}_m \overbrace{(R_{\text{agent}}^{\text{goal}}(\mathbf{s}, a) + \gamma \mathbf{w}^\top \phi_{\text{agent}}(\mathbf{s}, a))}^{\text{agent's reward}}} \tag{8}$$

is the exponential component of the Boltzmann rationality model. The agent subscript in eq. (8) denotes the interactant associated with the modality m: when m is explicit, the agent is the robot R; when m is implicit, the agent is the human H. Thus, the agent's reward in eq. (8) is implemented per eq. (4) or eq. (5), respectively. Finally, the parameter $\hat{\beta}_m$ in eq. (8) models the robot's assumptions about how rational the human is at providing explicit or implicit feedback (depending on m) as a function of the agent's reward.

The goal of preference learning can then be expressed as: $\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathcal{L}(\mathbf{w})$. We solve this Maximum Likelihood Estimation (MLE) problem using a belief distribution for the preference weights, which is implemented via a sample set $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^M$, as indicated in lines 15-17 of Algorithm 1. Specifically, we use gradient ascent on the log-likelihood $LL(\mathbf{w}) = \log \mathcal{L}(\mathbf{w})$ to find suitable preference weights. When the update takes place, gradient ascent is applied on each weight $\mathbf{w}_i \in \mathbf{W}$ using the gradient:

$$\nabla LL(\mathbf{w}) = \sum_{(f_m^+, f_m^-, \mathbf{s}) \in \mathbf{F}} \left(\frac{\sum_{a \in f_m^+} \hat{\beta}_m \gamma \phi_{\text{agent}}(\mathbf{s}, a) B_m(\mathbf{s}, a)}{\sum_{a \in f_m^+} B_m(\mathbf{s}, a)} - \frac{\sum_{a \in f_m^+ \cup f_m^-} \hat{\beta}_m \gamma \phi_{\text{agent}}(\mathbf{s}, a) B_m(\mathbf{s}, a)}{\sum_{a \in f_m^+ \cup f_m^-} B_m(\mathbf{s}, a)} \right) \tag{9}$$

The belief $\mathbf{W}$ is randomly initialized when learning begins but, in subsequent time-steps, we start gradient descent using the belief estimated from the prior time-step. Reusing prior estimates of $\mathbf{W}$ is essential for gradient ascent to converge quickly because the feedback set $\mathbf{F}$ grows over time, making the objective more complex.

5 Evaluation in Simulation

We first evaluate our proposed approach for learning human preferences over a human-robot collaboration in simulation. We focus the evaluation on understanding the impact of the feedback modalities and key parameters of PIE. In particular, we consider different values for $\hat{\beta}_{imp}$ and $\hat{\beta}_{exp}$ in eq. (8). These parameters are used to find the weights $\mathbf{w}$ that maximize the likelihood of the accepted behavior implied by the human's feedback (see eq. (7)). For implicit feedback, $\hat{\beta}_{imp}$ indicates how rational the robot considers the human to be in choosing its own actions during the collaboration. For explicit feedback, $\hat{\beta}_{exp}$ indicates how rational the robot considers the human to be at deciding whether the robot's behavior is acceptable or not. More specifically, our research questions are:

(RQ1) *How does the type of feedback considered by the robot affect preference learning?* A motivating hypothesis for this work is that combining explicit and implicit feedback will facilitate learning preferences over the collaboration. Thus, we compare three experimental conditions: 1) *explicit-only* feedback, where the robot infers preferences by only reasoning about binary feedback provided by the human; 2) *implicit-only* feedback, where the robot infers preferences by only reasoning about the human's high-level actions during the collaboration; and 3) *combined* feedback, where the robot learns from both explicit and implicit feedback with PIE.

We know that people can deviate from optimal decision making in varied ways [12, 30], so RQ1 considers different levels of rationality for the human's actions in the collaboration (β_H in eq. (6)). Also, because RQ1 is focused on the effect of different feedback modalities, we assumed that the robot knows the level of rationality of the human's actions, so $\hat{\beta}_{imp} = \beta_H$. Lastly, we set $\hat{\beta}_{exp} = \hat{\beta}_{imp}$ for simplicity, as prior work often considers a single rationality coefficient β for integrating various types of feedback [17].

(RQ2) *How does the correctness of the robot's assumptions about the rationality of the human's feedback affect preference learning?* Our second experiment evaluates the performance of our PIE approach when we introduce the complication that the robot does not know how rational the human truly is. We systematically study in simulation how preference learning performance with PIE is affected by the alignment, or misalignment, between β_H and $\hat{\beta}_{imp}$. Specifically, we consider the actions taken by the human to be more ($\beta_H = 10$) or less rational ($\beta_H = 1$). Then, we consider two situations per β_H: the assumptions on the human's rationality are aligned with the simulated human (e.g., $\hat{\beta}_{imp} = \beta_H = 1$), or they are misaligned (e.g., $\hat{\beta}_{imp} = 1$ but $\beta_H = 10$). Also, the robot reasons about human button presses in two ways. It assumes that the human's explicit feedback is more rational with $\hat{\beta}_{exp} = 10$, or less rational with $\hat{\beta}_{exp} = 1$.

5.1 The Pizza-Making Task

We consider collaborations where the human and robot prepare pizzas together. The robot passes ingredients to the human, while the human is responsible for more complicated manipulation tasks involving assembling the pizza. This results in different action spaces for the human and the robot. The robot's high level actions include: `pick(broccoli, storage)` or `place(broccoli, workstation)`, whereas the human's high level actions include `add(pepperoni)` or `return(pepperoni)`. Both the robot and the human know the goal of the task, which is defined by the ingredients that comprise a given pizza. For example, they may work towards making a pizza with sauce, cheese, pepperoni, mushrooms, and olives. However, the robot does not know the true human preferences $\mathbf{w}^*$ for how the team should reach the goal. Thus, the robot selects actions during a collaboration according to only the goal reward: $\pi_R(a_R|\mathbf{s}) = P(a_R|\mathbf{s}) \propto \exp(\beta_R R_R^{\text{goal}}(\mathbf{s}, a_R))$, where β_R controls how rational the robot's actions are. We set $\beta_R = 1$ in our experiments.

We consider two types of preferences. First, the human can have *ordering preferences* for the ingredients, e.g., cheese should be placed on the pizza before the sauce. We considered a total of six ordering preferences, each of which was a feature in the preference weight vector $\mathbf{w}$. Second, the human can have *workspace preferences* over the number of ingredients that can be on the shared workspace at any given time, including one ingredient maximum, two ingredients maximum, or up to four ingredients. We describe these workspace preferences via two features in $\mathbf{w}$. Thus, the preference weight vector has a total of eight dimensions.

We set the components of the shared reward (eq. (1)) as follows. The goal reward is most positive when a topping that should be on the pizza is moved from the storage to the workstation or added to the pizza. The preference reward is defined as in eq. (3). The appendix includes a more detailed description of the reward, state, action, and preference space of the pizza-making task.

5.2 Simulating the Human in the Collaboration

Following common practice for evaluations in the preference learning literature [17, 26], we model human behavior in our simulations with a Boltzmann rationality model. We assume that the human tends to take rational actions per eq. (6). When a new time-step of the interaction occurs in simulation, the simulated human always chooses to give explicit feedback – later in Section 6, we demonstrate our preference learning approach with real human feedback, which can vary in frequency over time [9].

Prior work in interactive learning shows that binary human feedback tends to be "noise-reducing" in comparison to other types of explicit feedback [55]; thus, we simulated explicit feedback as rational feedback for studying RQ1 and RQ2. The simulated human decides which binary feedback signal to provide based on how well the robot performs relative to the best possible reward. To compute the best possible reward, we leverage the fact that the simulated human knows the reward of the robot R_R, as in eq. (4), because they know the goal reward and the true preference weights $\mathbf{w}^*$. Then, when the robot takes action a_R at a given time-step with a state $\mathbf{s}$, the simulated human compares the actual reward induced by the robot's action, $R_R(\mathbf{s}, a_R)$, with the highest possible reward $\max_{\hat{a}_R} R_R(\mathbf{s}, \hat{a}_R)$ the robot could receive at state $\mathbf{s}$, over all possible actions $\hat{a}_R$. If $R_R(\mathbf{s}, a_R) == \max_{\hat{a}_R} R_R(\mathbf{s}, \hat{a}_R)$, then the human gives positive binary feedback, indicating acceptable robot behavior. Otherwise, the simulated human gives negative feedback.

5.3 Evaluation Setup

We evaluate learning via two metrics:

L_2 error: As is common in preference learning, we evaluated learning in terms of the L_2 error with respect to the ground truth preference weights $\mathbf{w}^*$ as the interaction progresses. At each timestep, we compute $\|\mathbf{w}^* - \sum_i \mathbf{w}_i/M\|$, with $\mathbf{w}_i$ samples from the belief $\mathbf{W}$.

Conflict percentage: At each time step, we use the current, average estimate of the weights $\tilde{\mathbf{w}} = \sum_i \mathbf{w}_i/M$, with $\mathbf{w}_i \in \mathbf{W}$, to define a greedy policy for the robot, where it selects actions that maximize the reward according to $R_R(\mathbf{s}, a_R)$ parameterized by $\tilde{\mathbf{w}}$. Then, we simulate a full interaction with a target pizza where the robot takes actions according to the greedy policy, and calculate how many of the actions taken by the robot do not match the true optimal robot action, based on $\mathbf{w}^*$. This metric allows us to quantify the practical effect of learning the preferences, because a wrong estimate for the true weight $\mathbf{w}^*$ may still result in similar behavior to $\mathbf{w}^*$, or perhaps small deviations from $\mathbf{w}^*$ (not captured by the L_2 error) could change the robot's behavior in an undesired way.

5.4 Results

5.4.1 (RQ1) Type of Feedback. Figure 2 shows the results for 100 simulated human-robot interactions, each of which consisted of 10 pizzas and where the simulated human had a specific ground truth preference that was randomly sampled. We evaluated learning using different values of β, representing different rationality constants for the actions the simulated human took (β_H in eq. (6)) and for the robot's assumptions when reasoning about explicit and implicit human feedback ($\hat{\beta}_{exp}$ and $\hat{\beta}_{imp}$ in eq. (8)). Overall, the figure shows that PIE tended to result in better or similar performance than only using explicit feedback or only using implicit feedback.

We conducted a statistical analysis of the final L_2 error in Figure 2, after the 10 pizzas – due to limited space, we omit this analysis for the conflict percentage metric. Specifically, we used a linear mixed model analysis, estimated with REstricted Maximum Likelihood (REML), to evaluate the L_2 error. The model considered Interaction ID (100 levels) as random effect, Feedback Modality (*Explicit*, *Implicit*, or *Combined* with (PIE)) and Rationality ($\beta \in \{0.1, 1, 10\}$) as main effects, and the interaction effect of the latter two variables.

Feedback Modality had a significant effect on the L_2 error ($p < 0.0001$). As we hypothesized, a Tukey HSD post-hoc test indicated that learning preferences with *Combined* feedback (using PIE) led to significantly lower error than using a single feedback modality only. At the end of learning, the L_2 error for *Combined* feedback was $M = 0.15$ ($SE = 0.007$). The error for *Explicit* feedback was $M = 0.25$ ($SE = 0.01$), and for *Implicit* feedback was $M = 0.24$ ($SE = 0.007$).

The analysis also indicated a significant effect of the Rationality parameter (β) on the L_2 error ($p < 0.0001$). A Tukey HSD post-hoc test indicated that $\beta = 0.1$ (with $M = 0.33$, $SE = 0.008$) led to significantly higher L_2 error than $\beta = 1$ ($M = 0.15$, $SE = 0.006$) and $\beta = 10$ ($M = 0.16$, $SE = 0.007$). As discussed later for RQ2, this finding can be due to an important mismatch between how the robot modeled the rationality of the human's explicit feedback when $\beta = 0.1$ (which implied $\hat{\beta}_{exp} = 0.1$ for RQ1) and the perfectly-rational approach used by the simulated human to provide explicit feedback (as explained in Sec. 5.2).

Finally, we also found the Feedback Modality × Rationality interaction to have a significant effect on the L_2 error at the end of the interactions ($p < 0.0001$). Significant pairwise differences from a Tukey HSD post-hoc test are shown in Fig. 3. Notably, using $\beta = 0.1$ and *Explicit* feedback only led to the highest error ($M = 0.44$; $SE = 0.01$) of all combinations of Modality and Rationality. Meanwhile, the *Combined* feedback led to significantly smaller error with $\beta = 1$ ($M = 0.08$; $SE = 0.004$) and $\beta = 10$ ($M = 0.09$; $SE = 0.006$) compared to all other combinations.

5.4.2 (RQ2) Assumptions for $\hat{\beta}_{imp}$ and $\hat{\beta}_{exp}$. Figure 4 shows the L_2 error for PIE over 100 interactions, where each interaction consisted of 10 pizzas. Visually, it is clear that $\hat{\beta}_{imp}$ affects performance. The dashed lines show preference learning performance when there is a mismatch between how rational the human is at taking actions (β_H) and how the robot modeled this rationality ($\hat{\beta}_{imp}$); while the solid lines indicate performance when the robot's assumption was correct and $\hat{\beta}_{imp} = \beta_H$. In most cases, the dashed line leads to

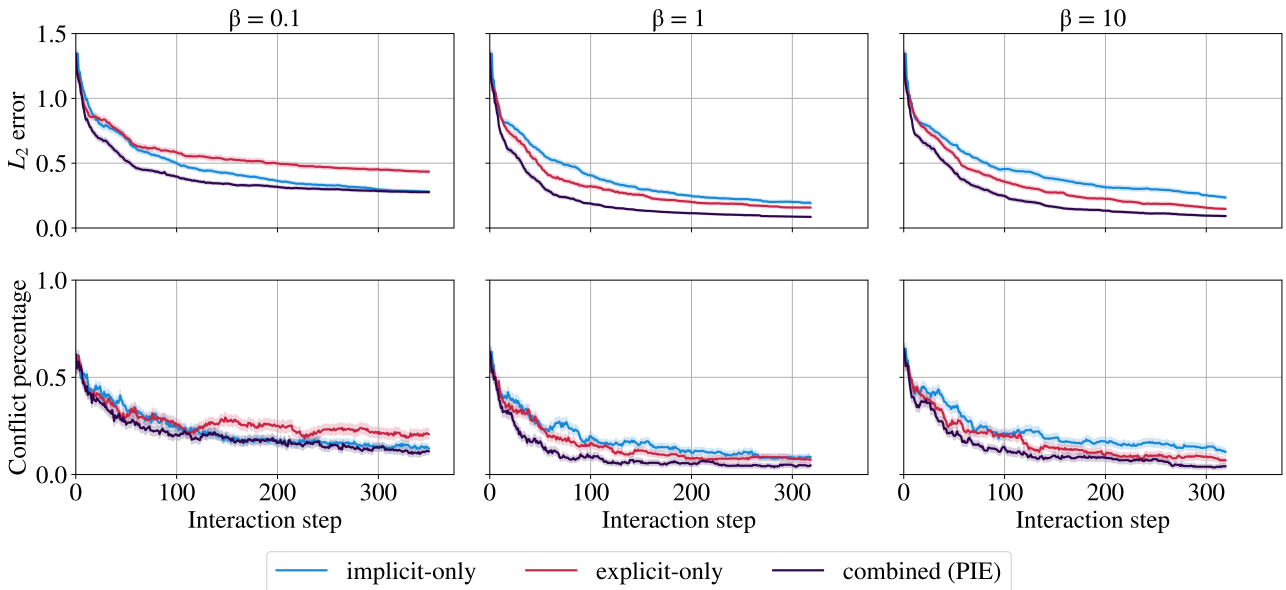

Figure 2: Results for RQ1: L_2 error with respect to the ground truth preference weights (top) and conflict percentage (bottom) as the interaction progresses. Lines represent average L_2 error and conflict percentage, and shading represents std. error across 100 interactions. Results are shown for implicit feedback only (blue), explicit feedback only (red), and both implicit and explicit feedback (dark purple). Columns show results considering different β values, where $\beta_H = \hat{\beta}_{imp} = \hat{\beta}_{exp} = \beta$ for $\beta \in \{0.1, 1, 10\}$.

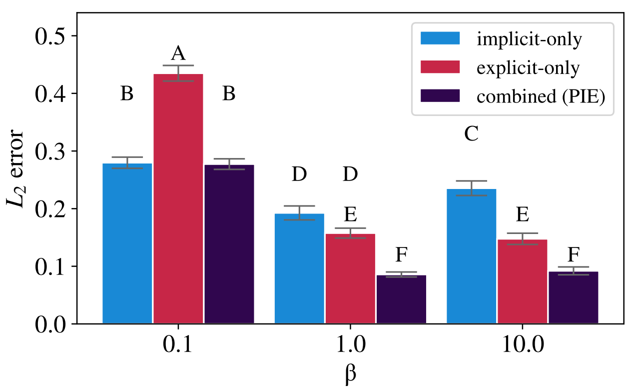

Figure 3: Results for RQ1: L_2 error at the end of the interactions based on Feedback Modality and Rationality ($\beta = \beta_H = \hat{\beta}_{imp} = \hat{\beta}_{exp}$) . Error bars are std. error. Bars labeled with different letters (A-F) have significantly different error based on a Tukey HSD post-hoc test. See the text for more details.

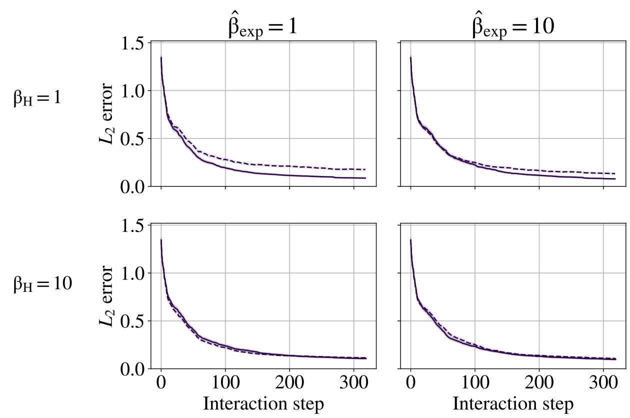

Figure 4: Results for RQ2: Average L_2 error as the robot learns with PIE. The two lines in the plots indicate whether the robot's assumption about the rationality of the human's action ($\hat{\beta}_{imp}$) match the oracle's behavior (β_H): the solid lines correspond to $\hat{\beta}_{imp} = \beta_H$; and the dashed lines correspond to an erroneous assumption $\hat{\beta}_{imp} \neq \beta_H$. In the top row, when $\hat{\beta}_{imp} \neq \beta_H$, $\hat{\beta}_{imp} = 10$. In the bottom row, when $\hat{\beta}_{imp} \neq \beta_H$, $\hat{\beta}_{imp} = 1$. Shaded areas are the std. error in 100 interactions.

higher error, especially when $\beta_H = 1$. The results in Fig. 4 seemed less susceptible to the choice of $\hat{\beta}_{exp} \in \{1, 10\}$.

We conducted a linear mixed model analysis on the L_2 error at the end of the interactions, considering Interaction ID as random effect. The main effects were $\hat{\beta}_{imp}$ Alignment (which had a value of 1 when $\hat{\beta}_{imp} = \beta_H$, and 0 otherwise), and $\hat{\beta}_{exp}$ Alignment (which had a value of 1 when $\hat{\beta}_{exp} = 10$ and was 0 otherwise, because 10 better approximated the purely-rational feedback from the simulated human than $\hat{\beta}_{exp} = 1$). The analysis also considered the interaction effect between $\hat{\beta}_{imp}$ Alignment and $\hat{\beta}_{exp}$ Alignment.

The analysis indicated that $\hat{\beta}_{imp}$ Alignment had a significant effect on the L_2 error at the end of the interactions ($p < 0.0001$). As expected, a post-hoc t-test showed that $\hat{\beta}_{imp} = \beta_H$ led to significantly lower error than the misaligned $\hat{\beta}_{imp}$. Similarly, $\hat{\beta}_{exp}$ Alignment had a significant effect on the L_2 error ($p < 0.0001$). The post-hoc test indicated that $\hat{\beta}_{exp} = 10$ (more aligned) led to lower error than $\hat{\beta}_{exp} = 1$ (less aligned), although the difference was small ($M = 0.10$; $SE = 0.003$ vs. $M = 0.12$; $SE = 0.003$).

Finally, the interaction effect between $\hat{\beta}_{imp}$ Alignment and $\hat{\beta}_{exp}$ Alignment was significant ($p = 0.036$). A Tukey HSD post-hoc test indicated that the error was significantly higher when $\hat{\beta}_{imp}$ and $\hat{\beta}_{exp}$ were both misaligned. Also, when $\hat{\beta}_{imp}$ was misaligned but $\hat{\beta}_{exp}$ was not, the error was significantly higher than when $\hat{\beta}_{imp}$ was aligned (whether $\hat{\beta}_{exp} = 1$ or $\hat{\beta}_{exp} = 10$).

Overall, these results reinforce findings for β with RQ1, and are consistent with prior work showing that having incorrect assumptions about rationality harms performance [10].

6 Real-World Evaluation

Having validated our approach in simulation, we conducted a real-world evaluation with 21 people. Each person collaborated on the pizza-making task with a robot, as illustrated in Fig. 1, while the robot tried to estimate their preferences for the collaboration. Through the real-world demonstration, we investigated:

(***RQ3***) *How well can the robot learn preferences over collaborations in real-world human-robot interactions?* The main challenge in this setup is learning from realistic human feedback, which may be noisy and sparse in more complicated ways than modeled in Section 5.

6.1 Experimental Protocol

We conducted the real-world evaluation with approval from our local Institutional Review Board. An experimental session typically lasted 45 min. Each participant was compensated US$15 for taking part in the real-world evaluation. People participated in the same pizza-making task described in Section 5.1.

Experimental Setup. As shown in Fig. 1, the participants interacted with a robot system made of two robots: a Franka Emika Panda arm, and a table-top robot called Shutter [50]. The Panda executed pick and place actions planned within the MoveIt Task Constructor framework [19]. During interactions, Shutter engaged with participants through its gaze and speech. Following Candon et al. [9], Shutter occasionally reminded participants to give explicit feedback. The reminders were framed as helping the robot to improve as a teammate (e.g., *"Remember that you can give me feedback so we can collaborate better in the future"*), and were only issued before the robot picked an object. High-level, multimodal robot behavior was controlled with behavior trees[14].

The table served as the collaborative workspace with defined storage and hand-off areas for the pizza ingredients. Each pizza ingredient was stored in a clear plastic container that could be grasped by the Panda hand. The table included two illuminated buttons that participants could press to give explicit feedback.

Procedure. The participants consented to participate in the interaction and to be audio- and video-recorded. The experimenter explained the goal of the interaction, introduced the robot, and started a tutorial. During the tutorial, the robot explained the person's role and the robot's role, the workstation, and how the person could provide explicit feedback. The participant then constructed a simple, practice pizza with a basic preference to see how preferences influenced the pizza-making interaction. The experimenter finally explained the set of preferences to chose from, had the participant select a preference, and went through hypothetical scenarios to ensure the participant understood their preferences. Each participant then worked with the robot to construct three different pizzas.

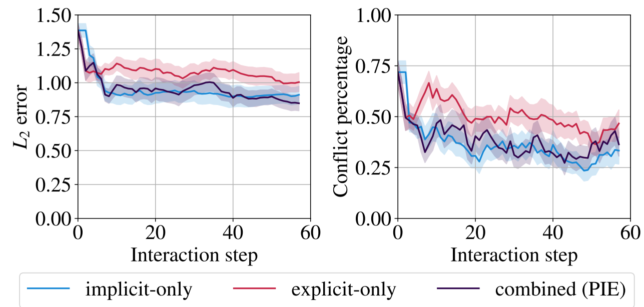

Figure 5: L_2 error with respect to the ground truth preference weights (left) and conflict percentage (right) as the interaction progresses. Lines represent average L_2 error and conflict percentage, and shading represents the standard error across the 21 participants in the real-world evaluation. Results are shown for implicit feedback only (blue), explicit feedback only (red), and our combined PIE approach (dark purple).

Participants. We recruited 21 participants via flyers, online postings, and word of mouth. They were required to be at least 18 years of age, be fluent in English, and have normal or corrected-to-normal hearing and vision. Sixteen of the evaluation participants (76%) were undergraduate or graduate students.

Evaluation. We conducted offline preference learning on the data collected from the 21 participants. We evaluated learning via L_2 error and conflict percentage, as in Sec. 5.3. For each participant, we first fit $\hat{\beta}_{imp}$ and $\hat{\beta}_{exp}$ using the data from the practice pizza, as an individual calibration step. We then used the fitted values for the three pizzas in the participant's interaction.

6.2 Results

Figure 5 shows the results of running 3 pizzas each with 21 different participants. We analyzed the L_2 error at the end of the interactions with the 21 participants with a linear mixed model that considered Participant ID as random effect, and Feedback Modality (*Explicit*, *Implicit*, or *Combined* with PIE) as main effect. We found a trend for Feedback Modality having an effect on the L_2 error ($p = 0.07$). At the end of the interactions, the average L_2 error was $M = 0.911$ ($SE = 0.062$) for *Implicit* feedback, $M = 1.005$ ($SE = 0.070$) for *Explicit* feedback, and $M = 0.849$ ($SE = 0.055$) for *Combined* feedback with PIE. The conflict percentage results were similar, with $M = 0.333$ ($SE = 0.067$) for *Implicit* feedback, $M = 0.466$ ($SE = 0.067$)

for *Explicit* feedback, and $M = 0.364$ ($SE = 0.058$) for *Combined* feedback.

7 Discussion

Our PIE approach outperforms single-modality baselines, achieving lower L_2 error and fewer conflicts in simulation. Even small reductions in conflict—such as 2–3 versus 5 incorrect actions in a 50-step pizza task—can meaningfully improve both task performance and human perception of the robot. Real-world results show a similar trend: leveraging multiple, naturally occurring feedback signals enhances a robot's ability to infer human preferences, highlighting the value of richer feedback in human-robot collaboration. Our results also highlight the importance of incorporating accurate assumptions when reasoning about feedback. Future work could explore jointly learning preference weights and individualized betas, and addressing modality-specific effects on gradient updates.

We opted for a myopic objective so that we could reason about human feedback in relation to high-level actions with dense rewards. This limits applicability in sparse-reward settings, where reasoning over longer horizons would be beneficial. This direction would require reasoning about the gradients of value functions, which we leave to future work. Our presented work is also limited to only two modalities of feedback: binary evaluative feedback via button presses and human task actions. Future efforts could integrate other types of feedback signals (e.g., corrections or language [26]).

By learning from naturally occurring, multimodal feedback, PIE moves toward more seamless, adaptive human-robot interaction, reducing the teaching burden on humans and fostering intuitive, productive collaboration.

Acknowledgments

This work was partly funded by Tata Sons Private Limited, Tata Consultancy Services Limited, Titan, and the National Science Foundation (Grant No. IIS-2143109). LQ was funded by YES Scholars Summer Fellowship and AQ was funded by Yale College First-Year Summer Research Fellowship in the Sciences & Engineering.

References

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.
[2] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.
[3] Inbal Avraham and Reuth Mirsky. 2025. Shared Control with Black Box Agents using Oracle Queries. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*. IEEE, 1–8.
[4] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113, 3 (2009), 329–349.
[5] Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. 2013. Teamwork with limited knowledge of teammates. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 27. 102–108.
[6] Andrew G Barto. 2021. Reinforcement learning: An introduction. by richard's sutton. *SIAM Rev* 6, 2 (2021), 423.
[7] Jake Brawer, Debasmita Ghose, Kate Candon, Meiying Qin, Alessandro Roncone, Marynel Vázquez, and Brian Scassellati. 2023. Interactive Policy Shaping for Human-Robot Collaboration with Transparent Matrix Overlays. In *Proceedings of HRI*.
[8] Andrew G Brooks and Ronald C Arkin. 2007. Behavioral overlays for non-verbal communication expression on a humanoid robot. *Autonomous robots* 22 (2007), 55–74.
[9] Kate Candon, Helen Zhou, Sarah Gillet, and Marynel Vázquez. 2023. Verbally Soliciting Human Feedback in Continuous Human-Robot Collaboration: Effects of the Framing and Timing of Reminders. In *Proceedings of HRI*.
[10] Lawrence Chan, Andrew Critch, and Anca Dragan. 2021. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956* (2021).
[11] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 307–315.
[12] Susan EF Chipman. 2016. *The Oxford handbook of cognitive science*. Oxford University Press.
[13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).
[14] Michele Colledanchise and Petter Ögren. 2018. Behavior Trees in Robotics and AI.
[15] Yuchen Cui, Qiping Zhang, Alessandro Allievi, Peter Stone, Scott Niekum, and W. Bradley Knox. 2020. The EMPATHIC Framework for Task Learning from Implicit Human Feedback. In *CoRL*.
[16] David Feil-Seifer and Maja J. Matarić. 2011. Socially Assistive Robotics. *IEEE Robotics & Automation Magazine* 18, 1 (2011), 24–31. doi:10.1109/MRA.2010.940150
[17] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. 2022. INQUIRE: INteractive querying for user-aware informative REasoning. In *6th Annual Conference on Robot Learning*.
[18] Cedric Goubard and Yiannis Demiris. 2023. Cooking up trust: Eye gaze and posture for trust-aware action selection in human-robot collaboration. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–5.
[19] Michael Görner, Robert Haschke, Helge Ritter, and Jianwei Zhang. 2019. MoveIt! Task Constructor for Task-Level Motion Planning. In *2019 International Conference on Robotics and Automation (ICRA)*. 190–196.
[20] Bradley Hayes and Brian Scassellati. 2016. Autonomously constructing hierarchical task networks for planning and human-robot collaboration. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5469–5476.
[21] Hui-Ru Ho, Edward M Hubbard, and Bilge Mutlu. 2024. " It's Not a Replacement:" Enabling Parent-Robot Collaboration to Support In-Home Learning Experiences of Young Children. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
[22] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–35.
[23] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems* 31 (2018).
[24] Anil Kumar Inkulu, MVA Raju Bahubalendruni, Ashok Dara, and SankaranarayanaSamy K. 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 (2022), 226–239.
[25] I. Iturrate, L. Montesano, and J. Minguez. 2010. Robot reinforcement learning using EEG-based reward signals. In *2010 IEEE International Conference on Robotics and Automation*. 4822–4829. doi:10.1109/ROBOT.2010.5509734
[26] Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems* 33 (2020), 4415–4426.
[27] W. Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: the TAMER framework. In *Proceedings of the Fifth International Conference on Knowledge Capture* (Redondo Beach, California, USA) *(K-CAP '09)*. Association for Computing Machinery, New York, NY, USA, 9–16. doi:10.1145/1597735.1597738
[28] Kheng Lee Koay, Dag Sverre Syrdal, Michael L Walters, and Kerstin Dautenhahn. 2009. Five weeks in the robot house–exploratory human-robot interaction trials in a domestic setting. In *2009 second international conferences on advances in computer-human interactions*. IEEE, 219–226.
[29] Przemyslaw A Lasota and Julie A Shah. 2015. Analyzing the effects of human-aware motion planning on close-proximity human–robot collaboration. *Human factors* 57, 1 (2015), 21–33.
[30] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. 2021. B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026* (2021).
[31] Jia-Min Li, Ke-Xi Liu, Ji-Fei Xie, and Hao-Yu Wu. 2024. How Does Human-Robot Collaboration Affect Hotel Employees' Proactive Behavior? *International Journal of Human–Computer Interaction* (2024), 1–15.
[32] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. 2022. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research* 41, 1 (2022), 20–44.
[33] Dylan P Losey and Marcia K O'Malley. 2018. Including uncertainty when learning from human corrections. In *Conference on Robot Learning*. PMLR, 123–132.

[34] Daniel Marta, Simon Holk, Christian Pek, Jana Tumova, and Iolanda Leite. 2023. Aligning human preferences with baseline objectives in reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA), MAY 29-JUN 02, 2023, London, ENGLAND*. Institute of Electrical and Electronics Engineers (IEEE).

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

[36] Debasmita Mukherjee, Kashish Gupta, Li Hsin Chang, and Homayoun Najjaran. 2022. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing* 73 (2022), 102231.

[37] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 61–68.

[38] Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. 2022. Learning multimodal rewards from rankings. In *Conference on robot learning*. PMLR, 342–352.

[39] Austin Narcomey, Nathan Tsoi, Ruta Desai, and Marynel Vázquez. 2024. Learning human preferences over robot behavior as soft planning constraints. *arXiv preprint arXiv:2403.19795* (2024).

[40] Carl Oechsner, Sven Mayer, and Andreas Butz. 2022. Challenges and Opportunities of Cooperative Robots as Cooking Appliances. *AutomationXP@ CHI* (2022).

[41] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. 2019. Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In *Proceedings of the 12th ACM international conference on PErvasive technologies related to assistive environments*. 247–255.

[42] Mariah L. Schrum, Erin Hedlund-Botti, Nina Moorman, and Matthew C. Gombolay. 2022. MIND MELD: Personalized Meta-Learning for Robot-Centric Imitation Learning. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 157–165. doi:10.1109/HRI53351.2022.9889616

[43] Thomas B Sheridan. 2016. Human–robot interaction: status and challenges. *Human factors* 58, 4 (2016), 525–532.

[44] Herbert A Simon. 1955. A behavioral model of rational choice. *The quarterly journal of economics* (1955), 99–118.

[45] Yue Song, Mengying Zhang, Jiajing Hu, and Xingping Cao. 2022. Dancing with service robots: The impacts of employee-robot collaboration on hotel employees' job crafting. *International Journal of Hospitality Management* 103 (2022), 103220.

[46] Maia Stiber, Russell H Taylor, and Chien-Ming Huang. 2023. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 222–230.

[47] Yuta Sugiura, Daisuke Sakamoto, Anusha Withana, Masahiko Inami, and Takeo Igarashi. 2010. Cooking with robots: designing a household system working in open environments. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2427–2430.

[48] Prasanth Sengadu Suresh, Siddarth Jain, Prashant Doshi, and Diego Romeres. 2024. Open human-robot collaboration using decentralized inverse reinforcement learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7092–7098.

[49] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737.

[50] Sydney Thompson, Austin Narcomey, Alexander Lew, and Marynel Vázquez. 2024. Shutter: A Low-Cost and Flexible Social Robot Platform for In-the-Wild Deployments. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. 94–96.

[51] Kristina Tornbjerg, Anne Marie Kanstrup, Mikael B Skov, and Matthias Rehm. 2021. Investigating human-robot cooperation in a hospital environment: Scrutinising visions and actual realisation of mobile robots in service work. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. 381–391.

[52] Konstantinos Tsiakas, Maher Abujelala, and Fillia Makedon. 2018. Task engagement as personalization feedback for socially-assistive robots and cognitive training. *Technologies* 6, 2 (2018), 49.

[53] Mikel Val-Calvo, José Ramón Álvarez-Sánchez, José Manuel Ferrández-Vicente, and Eduardo Fernández. 2020. Affective robot story-telling human-robot interaction: exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access* 8 (2020), 134051–134066.

[54] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research* 18, 136 (2017), 1–46.

[55] Hang Yu, Reuben M Aronson, Katherine H Allen, and Elaine Schaertl Short. 2023. From "Thumbs Up" to "10 out of 10": Reconsidering Scalar Feedback in Interactive Reinforcement Learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4121–4128.

[56] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of HRI* (Stockholm, Sweden). 10 pages. doi:10.1145/3568162.3576986

[57] Qiping Zhang, Nathan Tsoi, Mofeed Nagib, Booyeon Choi, Jie Tan, Hao-Tien Lewis Chiang, and Marynel Vázquez. 2025. Predicting Human Perceptions of Robot Performance during Navigation Tasks. *ACM Transactions on Human-Robot Interaction* 14, 3 (2025), 1–27.

[58] Michelle D Zhao, Reid Simmons, and Henny Admoni. 2023. Learning Human Contribution Preferences in Collaborative Human-Robot Tasks. In *Conference on Robot Learning*. PMLR, 3597–3618.

[59] Zuyuan Zhu and Huosheng Hu. 2018. Robot learning from demonstration in robotic assembly: A survey. *Robotics* 7, 2 (2018), 17.