2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3503-7770-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IROS58592.2024.10801370

Understanding Robot Minds: Leveraging Machine Teaching for Transparent Human-Robot Collaboration Across Diverse Groups

Suresh Kumaar Jayaraman
Robotics Institute
Carnegie Mellon University
Pittsburgh, USA

Reid Simmons
Robotics Institute
Carnegie Mellon University
Pittsburgh, USA

Aaron Steinfeld
Robotics Institute
Carnegie Mellon University
Pittsburgh, USA

Henny Admoni
Robotics Institute
Carnegie Mellon University
Pittsburgh, USA

***Abstract*—In this work, we aim to improve transparency and efficacy in human-robot collaboration by developing machine teaching algorithms suitable for groups with varied learning capabilities. While previous approaches focused on tailored approaches for teaching individuals, our method teaches teams with various compositions of diverse learners using team belief representations. We investigate various group teaching strategies, such as focusing on individual beliefs or the group's collective beliefs, and assess their impact on learning robot policies for different team compositions. Our findings reveal that team belief strategies produce less variation in learning duration and better accommodate diverse teams compared to individual belief strategies, suggesting their suitability in mixed proficiency settings with limited resources. In contrast, individual belief strategies provide a more uniform knowledge level, particularly effective for homogeneously inexperienced groups. Our study indicates that the effectiveness of the teaching strategy is significantly influenced by team composition and learner proficiency, highlighting the importance of real-time assessment of learner proficiency and adapting teaching approaches based on learner proficiency for optimal teaching outcomes.**

***Index Terms*—explainable decision-making, human-robot teams, group machine teaching, adaptive explainability, team modeling**

I. Introduction

Robots are increasingly becoming an integral part of people's lives, evolving from human assistants to collaborators. For safe and effective human-robot collaboration, it is crucial that humans understand how robots make decisions. Explainable decision-making aims to clarify the underlying robot decision-making process to human collaborators.

Our approach focuses on explaining the robot policy to the human learner using the framework of machine teaching [1]. In prior work, the human is frequently modeled as an inverse reinforcement learner [2] who learns a robot policy from robot behavioral demonstrations [3]. The objective of the machine teaching problem is to identify the minimum set of examples/demonstrations of robot behavior that will aid the human to learn the robot policy. Huang et al. [3] used informative demonstrations with approximate Bayesian IRL, which requires a fixed set of candidate reward functions. On the other hand, Cakmak and Lopes [4] focused on reducing uncertainty in IRL learners by selecting demonstrations that maximally reduce uncertainty over reward parameters. More recently, interactive methods for policy explanation [5] have been used in which humans request specific demonstrations, showing that a hybrid strategy of AI-selected and human-selected demonstrations yields the best results.

Explanations are more useful when personalized to the individual [6]–[8]. While many machine teaching methods discussed earlier focus on individuals, real-world scenarios

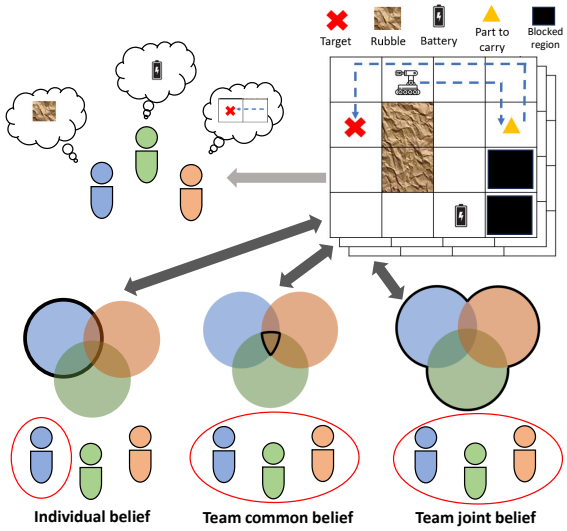

Fig. 1. The figure illustrates the complexity of group machine teaching, highlighting the disparity within diverse individuals in interpreting and understanding the various concepts of the robot's decision-making from common examples. Personalizing examples to a group is challenging due to varied individual beliefs and learning abilities. Our approach utilizes estimations of individual and collective team beliefs to tailor demonstrations for effective communication of the robot's policy to the entire group.

This work was supported by the Office of Naval Research award N00014-181-2503.

often involve the robot working with groups of people. In such cases, teaching the entire group simultaneously is preferable, as most real world situations have limited time and resources. Consider, for example, an ad hoc emergency response team tasked with building shelters after an earthquake. They are given a robot that can bring the requested items to the team members. The robot has limited maneuverability over rubble, a limited range, and may prefer to recharge when possible. These capabilities and preferences inform the robot's behavior and thus must be taught to the team quickly because of the time-sensitive situation. However, a challenge in group teaching is accommodating individuals with varied learning abilities through a common set of demonstrations. This work investigates adaptive approaches for robots to effectively communicate their decision making through demonstrations to such diverse groups of human learners.

Prior work has shown that it is possible to teach a heterogeneous group of learners using common examples [9], albeit for simple concepts. While groups can also learn from each other through communication and information sharing, we focus only on learning from common examples. Also, group heterogeneity could imply differences in prior knowledge, but we assume similar prior knowledge, focusing solely on varying learning abilities. There are several ways to teach a group, for example, in a classroom, a teacher could personalize the lessons to the class based on various strategies — focus on naive learners who need more support, or on proficient learners, or aggregate the class as a whole — and adapt their teaching accordingly. Yeo et al. [10] explored categorizing learners based on learning rates and provided personalized teaching to each category. Melo and Lopes [11], on the other hand, generated ultra-personalized demonstrations for each individual learner, but at a high teaching cost. An active teacher who personalizes and adapts to the learner can improve learning [12], [13]. But a challenge in groups is to identify to whom the personalization should be directed.

Drawing inspiration from the pedagogical literature on teaching [14], our key insight is that *machine teaching can be tailored to a group of learners by considering the group as a whole and generating demonstrations based on the aggregation of the group's understanding.* In this work, we develop team belief models that facilitate group teaching focusing on the entire team. We utilize a closed-loop teaching framework that effectively incorporates feedback from the robot teacher to assist human learning. We adapt a human belief model to generate simulated human learners with varying learning abilities. We conducted a simulation study to explore how different group teaching strategies affect the group's learning and how team composition of learners with varying learning abilities, naive and proficient, moderate group learning. Our findings suggest that teaching methods designed for individual beliefs weren't much affected by how knowledgeable team members were. However, these methods did affect how long the team members took to learn the robot policy, depending on team composition. On the other hand, teaching methods that focused on team beliefs helped increase knowledge, especially in groups with more proficient learners.

II. Background

Machine teaching for policies: We represent the environment as a Markov Decision Process (MDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma, \mathcal{S}_\iota \rangle$, representing the state and action spaces, transition function, reward function, discount factor, and initial state distribution, respectively. An optimal trajectory ξ^* is a sequence of (s_i, a, s_i') tuples obtained by following the robot's optimal policy π^*. Similar to prior work [15], the reward function R is expressed as a weighted linear combination of reward features $R = {\mathbf{w}^*}^\top \phi(s, a, s')$. We define a domain as a set of MDPs that share $R, \mathcal{A}$, and γ but differ in T_i, $\mathcal{S}_i$, and S_i^0. The consistency of R across the domain allows for demonstrations that facilitate inference over a common $\mathbf{w}^*$. We apply this MDP framework to model an *item delivery* task, where a robot is required to deliver an item within an environment featuring rubble, blockages, and a battery recharge station (see Fig.1).

We adapt the machine teaching framework for policies [16] to select a subset of demonstrations $\mathcal{D}$ of size n that maximizes the similarity ρ between the optimal policy π^* and the policy $\hat{\pi}$ recovered using a computational model $\mathcal{M}$ (e.g., IRL) on $\mathcal{D}$. This is formulated as $\arg\max_{\mathcal{D} \subset \Xi} \rho(\hat{\pi}(\mathcal{D}, \mathcal{M}), \pi^*) \quad \text{s.t.} \quad |\mathcal{D}| = n$, where Ξ is the set of all demonstrations of π^* in a domain. Once $\mathbf{w}^*$ is approximated through IRL, it is assumed that the learner can infer π^* by planning on the underlying MDP. Therefore, the objective becomes selecting demonstrations that are most informative in conveying $\mathbf{w}^*$, which can be evaluated using behavior equivalence classes.

Behavior equivalence class: *The behavior equivalence class (BEC)* of a policy π is the set of reward functions under which π is optimal. When the reward function is a weighted linear combination of features, the BEC of a demonstration ξ of π is determined by the intersection of half-spaces [17] formed by the exact IRL equation:

$$\text{BEC}(\xi|\pi) := \mathbf{w}^\top \left(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)} \right) \geq 0, \forall (s,a) \in \xi, b \in \mathcal{A}. \tag{1}$$

where $\mu_\pi^{(s,a)} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi\left(s_t\right) \mid \pi, s_0 = s, a_0 = a\right]$ represents the vector of reward feature counts accrued from taking action a in s and subsequently following policy π. Any demonstration can be transformed into a set of constraints on $\mathbf{w}$ using (1), with each constraint representing a knowledge component (KC) [18] that encapsulates an aspect of the reward function, such as trade-offs between the underlying reward features. For instance, consider the item delivery domain with binary reward features ϕ = [*traversed rubble*, *battery recharged*, *action taken*]. In practice, we enforce $||\mathbf{w}^*||_2 = 1$ to avoid issues related to the scale invariance of IRL and the degenerate all-zero reward function. If no prior knowledge is assumed, the potential belief space on reward weights would uniformly cover the surface of the $n-1$ sphere (where n is the number of domain features) due to the L^2 norm constraint on $\mathbf{w}^*$. However, we assume that learners begin with a prior belief

that the action weight is negative (e.g., favoring the shortest path, see Fig. 2).

Team modeling: A common way to represent a team characteristic such as team knowledge is by aggregating the knowledge of individuals. Team characteristics are normally represented as average, median, sum, range, minimum, or maximum values of the characteristic of individuals [19]. More recently, team knowledge is represented using a latent *collective intelligence* parameter that is highly correlated with team process and performance [20]. However, operationalizing such a latent parameter is challenging and we thus represent team belief through observable behaviors by aggregating individual beliefs. We focus on two aggregated representations of team belief — common belief and joint belief. We define **common team belief** as the belief that all team members have. It can be visualized as the intersection of individual beliefs. We define **joint team belief** as the knowledge that at least one individual in the team has, visualized as the union of individual beliefs (see Fig. 3 (b) for visual representations of these).

III. Methods

In this section, we discuss an approach using particle filters (PF) for modeling individual and team beliefs about the robot's decision-making, i.e. its reward. We use these different beliefs to select corresponding demonstrations that are shown to the entire team. Lee et al. [13] originally proposed a PF-based approach to model individual human belief that supports iterative Bayesian updates and sampling for generating informative and tailored demonstrations using counterfactual reasoning. We extend this approach to group teaching to model aggregated team beliefs in addition to individual beliefs. We use this model in a closed-loop teaching framework leveraging insights from the education literature and adaptively generating demonstrations based on individual and aggregated team beliefs. In addition, after seeing demonstrations, we provide tests, which collect responses on expected optimal robot trajectory in an unseen enrivonment to evaluate their understanding.

A. Particle filter model of human learner belief

Humans generally perform approximate inference from demonstrations [3] and thus we model the human learner's belief about the robot's reward weights using a particle filter, where each particle represents a potential belief about the reward weights [13] and the particle weight represents the belief probability. The particle filter follows a Bayesian update process that uses constraints of the corresponding demonstrations and tests. This formulation enables iterative updates on learner belief from demonstrations and tests.

From demonstrations, constraints on reward weights can be obtained using Eq. 1, by comparing the optimal demonstration with possible counterfactuals. Similarly, the correct test response can be compared with the incorrect learner response to get these constraints using Eq. 1. Each constraint c_i generated from the demonstrations and test responses is a half-space constraint, meaning, one side is consistent with the demo or test response and the other is not. Each constraint c_i can be converted to a probability distribution $p(x_i|c_i)$ that is used to update the particle weights w_i of particles x_i.

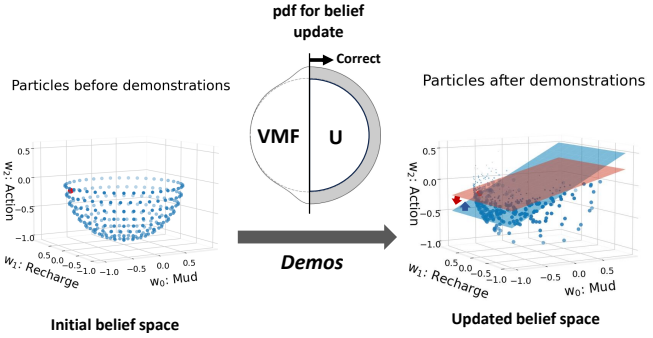

Fig. 2. Update process of a learner's belief space represented by a particle filter, with the red particle indicating the true reward weights. The left figure shows the initial belief space, which is a hemisphere since action costs are always negative. A cross-section of the custom probability density function (pdf) used for the particle filter belief update is shown. The updated belief distribution is shown on the right with the constraint planes corresponding to the demonstrations seen. Particles consistent with the demonstrated behavior, lying on the side of the half-space constraint planes indicated by the arrows, receive higher weights via a uniform distribution (U), while those on the inconsistent side are weighted less, decreasing exponentially with distance from the constraint, via a von-Mises Fisher (VMF) distribution.

We use the custom probability distribution (refer Fig. 2) proposed in [13], designed such that any particle on the consistent side of the constraint (yellow region in Fig. 2) is equally valid and could have generated the demonstration (represented by the uniform distribution) and the particles on the inconsistent side are exponentially less likely to have generated the demonstration (represented by the von Mises-Fisher distribution). Fig. 2 shows how the initial learner belief for the delivery robot's reward gets updated using this custom pdf after seeing the demonstrations. The belief updates after demonstrations and feedback are similar to the *predict* stage and the update based on the learner's test response is similar to the *correct* stage of Bayesian estimation. The robot teacher uses the learner belief space to sample possible counterfactuals for generating learner-centric demonstrations. To handle potential sample degeneracy in particle filtering, we add a Gaussian noise η when updating the particles [21]. We also add a small Gaussian noise ν while updating the particles after test to account for the teacher's estimation noise. For more details on the particle filter model refer [13].

B. Team belief modeling

A demonstration's ability to reveal the underlying reward function via IRL is highly dependent on the counterfactuals considered. Thus the learner's belief space from which the counterfactuals are sampled critically influences the demonstrations generated. For groups, we model team belief as aggregations of individual member beliefs [19], [22]. The main difference in modeling team beliefs is how the particles are updated, specifically how individual constraints are aggregated and used for updating the particle weights.

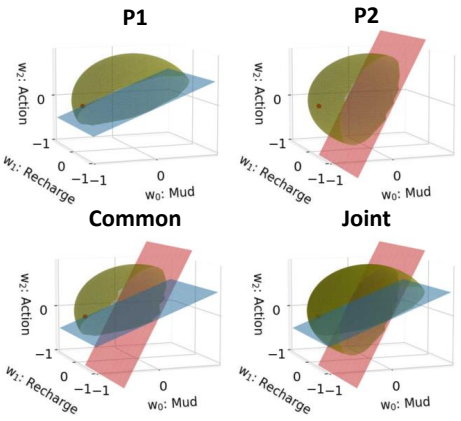

Fig. 3. This figure illustrates an example set of test responses for a team with two individuals, P1 and P2. The test responses are transformed to constraints. The yellow partial spheres show the regions that are consistent with their test response, i.e. agree with the constraint. When their responses are different, the constraints space of their common belief of their tests is their intersection of individual beliefs and that of their joint belief is the union of individual beliefs as depicted. These constraints spaces are used to update the weights of the PF distributions.

We envision a group teaching scenario akin to classroom teaching and assume that each team member will have the same kind of interactions, i.e. see the same demonstrations and are provided the same tests. Thus the constraints from the demonstrations will be similar for all individuals. However, let us assume that the team with m members had different responses to a set of n tests, and their constraints are denoted as $\mathcal{C}_1 = \{c_1^1, c_1^2, ...c_1^n\}$, $\mathcal{C}_2 = \{c_2^1, c_2^2, ..., c_2^n\}$, and $\mathcal{C}_m = \{c_m^1, c_m^2, ...c_m^n\}$. The update probability for each member is given by, $P_i = \prod_{j=1}^{n} p(x_i^j | c_i^j)$.

We model common team belief by considering the constraints of all members and representing it as $\mathcal{C}_{ck} = \{c_1^1, c_2^1, ..., c_m^1, c_1^2, c_2^2, ..., c_m^2, ..., c_1^n, c_2^n, ..., c_m^n\}$. We assume individuals to be independent. Consequently, the particle filter representing common team belief is updated based on the probability of all aggregated constraints across all tests in the set for all the individuals, given by, $P = \prod_{i=1}^{m} \prod_{j=1}^{n} p(x_i^j | c_i^j)$. This aligns with our definition of common belief as the belief that everyone on the team has.

Joint belief, on the other hand, is modeled by considering the set of constraints for all individuals for each test separately and is represented as a set of disjointed subsets, $\mathcal{C}_{jk} = \{\{c_1^1, c_2^1, ..., c_m^1\}, \{c_1^2, c_2^2, ..., c_m^2\}, ..., \{c_1^n, c_2^n, ..., c_m^n\}\}$, where each subset represents the constraints of the individuals. Update probabilities are calculated individually for each team member and the particles are updated based on the maximum probability of any of the individuals, given by, $P = \arg\max_{i \in [1,2,...,m]} \prod_{j=1}^{n} p(x_i^j | c_i^j)$. This corresponds to our definition of joint belief as the belief that at least one team member has.

Feedback is an effective learning mechanism [23]. Fig. 4 shows the effects of the demonstrations, tests, and feedback that the team has during one teaching session. A teaching session aims to teach a specific KC, for example, the trade-off between mud cost and step cost for the delivery robot. Each teaching session consists of a set of demonstrations, a set of tests, and a set of feedback (corrective or confirmatory feedback based on whether the response was *incorrect* or *correct*). Confirmatory feedback reinforces the learner's knowledge while corrective feedback informs them that their learning is incorrect and also provides the correct response.

Fig. 4 shows that for individuals *P1* and *P3* who responded correctly, the belief distribution becomes more concentrated within the constraints area. This is further strengthened by the confirmatory feedback they receive. On the other hand, individual *P2* responded incorrectly, indicating that they may not have learned the KC. Thus, their belief distribution moves away from the constraint region of the KC. However, receiving corrective feedback brings the distribution closer to the constraint region. The common team belief moves slightly further away due to the incorrect response of *P2* but is still mostly close to the correct reward and the constraint region.

C. Teaching curriculum development

We employ the methods discussed in [16] to generate demonstrations and tests for teaching the robot's policy. The approach primarily considers likely learner counterfactuals by estimating the learner's belief about robot policy (i.e. reward weights) and sampling n possible counterfactuals from this belief space. For every possible robot demonstration in a domain, and for each reward weight, we simulate what the "human" counterfactual to each demonstration would be if the human had this reward weight in mind and generate the corresponding constraints using Eq. 1 and consolidate all these constraints. We select the demonstration from this set of constraints that maximizes knowledge gain before and after seeing the demonstration. We use the consolidated set of constraints to identify test environments that examine the concept taught in the demos (refer [16] for additional information).

D. Simulated learner model

Learners have different cognitive capabilities and understand at different levels the same information provided. Furthermore, the teaching process is likely to be an adaptive and varied process, catered to the specific learner. Thus our model of the learner should be able to *encompass a wide variety of learner abilities in different teaching contexts*. We again employ a particle filter-based approach to simulate a learner's learning dynamics and belief updates. Similar to the teacher's model, each particle represents a potential belief about the robot's reward function, but they are the learner's self-belief as opposed to the teacher's estimated learner belief. There are two key differences between the teacher and the learner models — first, the learner model updates after seeing the demonstration and feedback only and not after tests since we expect that learners will get information only from the demonstrations and feedback and not from just knowing if they got test responses correct/incorrect [23] and second, the learner model does not have any estimation Gaussian noise (ν) and only the resampling noise (η).

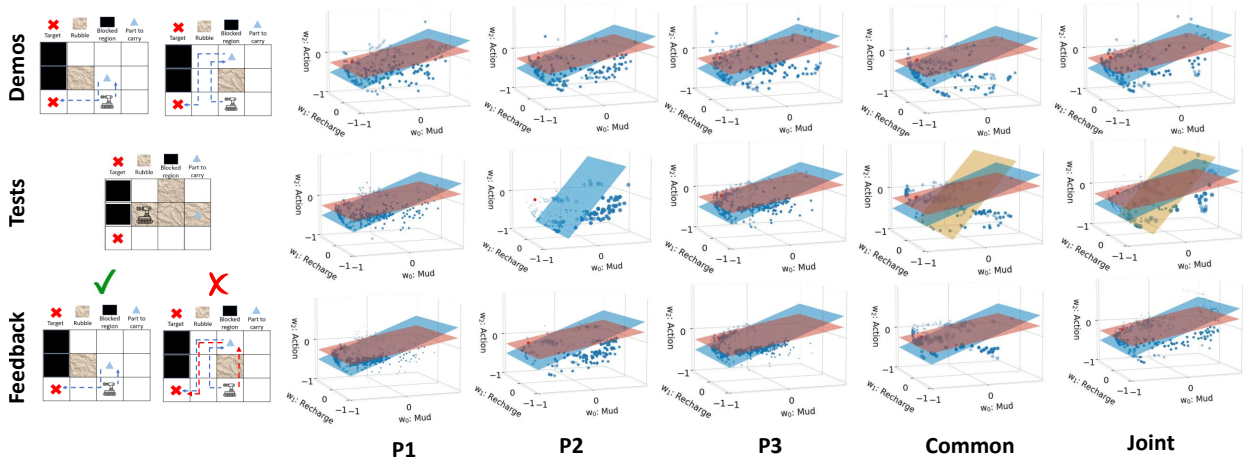

Fig. 4. Interactions and corresponding PF belief updates for a team with three people for the first teaching session. The red particle represents the true reward weight. A teaching session consists of one set of demos related to a KC, followed by a set of tests, and then feedback (corrective or confirmatory). After the demos, all individual and team beliefs are updated based on expected information gain from demos. The distributions are similar since all individuals are expected to learn equally. After the demos, they are provided with test(s) to evaluate their understanding and their responses are used directly for updating individual beliefs and aggregated to update team beliefs. In this case, P1 and P3 got the response correctly and P2 got it incorrect. The difference in the constraint spaces and the updated distributions after tests can be observed for the individual and team beliefs. Confirmatory or corrective feedback is given after the tests and they are expected to learn from either feedback. The distributions are updated to reflect this learning from feedback.

Each individual has a different ability to understand the information conveyed through visual demonstrations. We parametrize this ability as β that can vary for each individual. It is operationalized as the probability mass on the uniform (consistent) side of the custom distribution of the underlying constraint (see Fig. 2). It moderates how the demonstrations (and feedback) modify the belief space used for particle updates, $IG_{pf} = f(\beta, IG_{demo})$. A higher β implies that learners assign more weights to particles on the consistent side of the constraints, indicating certainty over particles that likely resulted in the demonstrations. More details on how various values of β affect the belief space can be found in [24].

We initialize $\beta = \beta_0$, as an individual's ability to learn from demonstrations. People learn a concept better when they receive feedback and are repeatedly exposed to the concept [23]. To incorporate the effects of feedback, we define the feedback dynamics of β as

$$\beta_t = \beta_{t-1} + \Delta\,\beta_{t-1}, \text{ with,} \tag{2a}$$

$$\Delta\,\beta_{t-1} = \begin{cases} \delta\,\beta_c & \text{if test at } t-1 \text{ is } \textit{correct} \\ \delta\,\beta_i & \text{if test at time } t-1 \text{ is } \textit{incorrect} \end{cases} \tag{2b}$$

where β_c is change in β due to confirmatory feedback, when the learner's test response is correct and β_i is due to corrective feedback, when the learner's test response is incorrect. Corrective and confirmatory feedback have different effects on learning [23]. Fazio et al. [23] found that feedback on incorrect responses led to more learning than feedback on correct responses, particularly in more difficult tasks. Thus we define $\beta_i > \beta_c$. β_t resets to β_0, for each new concept as we assume the concepts to be independent. Thus improvement in β due to feedback is contained within the specific concept or KC.

IV. Simulation Study

We ran a simulation study to evaluate the effects of different teaching strategies on group learning. The strategies differed in the belief space they utilized for sampling possible counterfactuals to generate informative demonstrations. We used N=8 counterfactuals in this study. We also examined the effects of the various team compositions of diverse group members on group learning.

A. Metrics

We measure the teaching-learning performance using two measures — (1) the number of teaching sessions (N_t) taken to learn the policy, and (2) the average team knowledge level at the end of learning. For each individual, their knowledge level is defined as the proportion of belief space that lies within the BEC region of the robot's policy at the end of the learning session. It is given by $p_{BEC} = \sum_{j=1}^{m} p_i, \forall\, i \in \epsilon_{BEC}$, where j is the individual, ϵ_{BEC} is the BEC region and i indicates an individual particle and p_i its normalized weight. The team knowledge level, $\overline{p_{BEC}}$ is the average knowledge level of all individuals.

B. Study conditions

The primary condition that we examined was the *teaching strategy* employed to generate the demonstrations and tests. We considered four strategies that samples counterfactuals from four different belief spaces for each interaction period — *individual low*, the belief space is of the individual with the lowest knowledge about the robot's reward, *individual high*, the belief space is of the individual with the highest knowledge about the robot's reward, *common*, the belief space is the common team belief, and *joint*, where the belief space is the joint team belief. The individual, common and joint

belief spaces are visualized in Fig. 3. The bigger the belief space, the more diverse the sampled counterfactuals would be. The individuals with lowest or highest knowledge are identified at the end of each interaction period and their corresponding belief spaces are used for the next interaction period corresponding to the strategy employed. We compared these different strategies with a baseline strategy of separately teaching each individual sequentially.

We also examined the influence of *team composition* on the group's learning. Particularly, we considered two categories of learners, *novice* and *proficient*. Novice learners are considered beginners and have a low ability to learn from demonstrations, given by a low β_0. On the other hand, proficient learners have a higher β_0. For the simulation study, we estimate the distribution of the learner parameters for both the types of learners (see Table I). We considered four team compositions, *all novice*, *majority novice*, *majority proficient*, and *all proficient*.

We expect that group teaching strategies based on group belief, especially the ones based on team belief, target the group as a whole and hence would be able to cater to the entire group resulting in faster group teaching. However, individual teaching strategies, specifically "individual low" is likely to result in higher knowledge as it focuses on the learners who have the least knowledge. So demonstrations catered for them will also improve the knowledge of others. We also expect that teams with more 'Proficient' learners will learn quicker than other teams. We expect that the baseline strategy, which teaches each learner separately would take more interactions, would result in better learning because it personally focuses on each learner. Thus we expect, *H1: Group-belief based strategies to have fewer interactions than individual-belief based strategies. H2: Individual low strategy to have the highest knowledge level apart from baseline because of its personalization. H3: Teams with more high learners to have both learn faster and high knowledge levels.*

C. Study setup

We conducted the study for a team with 3 learners. We ran a 5 $\times$ 4 study for the two conditions of *teaching strategy* and *team composition* including the baseline strategy. For each combination, we collected 15 'simulated' teams' data totaling to 300 teams.

- *Teaching strategy*: baseline, individual low, individual high, common, and joint
- *Team composition*: [N, N, N], [N, N, P], [N, P, P], and [P, P, P], where 'N' denotes Novice and 'P' denotes Proficient learners.

The learning parameters for each type of learner was estimated (see Table I) from human learner data collected from a user study discussed in [13]. We performed a grid search of the parameters by simulating the teaching interactions from the dataset for all possible grid combinations and choosing the runs that have performance error with $p_{BEC} < \epsilon$

The demonstrations are selected based on the KCs and the teaching strategy (which individual or team belief to adapt). The tests are similarly identified to assess how well they have understood the constraints conveyed by the demonstrations. For the conditions based on individual beliefs, the individuals with the lowest and highest knowledge are identified at the end of each teaching session, where a teaching session consists of a set of demonstrations, tests, and feedback related to a specific KC. The teaching moves to the next KC only after all individuals in the team learn the current KC. If the team does not learn the KC, the same KC is taught in the subsequent teaching sessions. The demonstrations and tests for subsequent sessions are calculated in real-time based on the updated belief of the learners. For aggregated group belief, for example, common belief, it is possible that the responses of learners are such that there is no common intersecting constraint region. In such cases, we exclude the conflicting individual(s) constraint spaces and consider the plausible intersecting region formed by most team members.

Learner type	$\beta_0\,(\mu\,,\sigma)$	$\delta\beta_c\,(\sigma)$	$\delta\beta_i\,(\sigma)$
Novice	0.703 (0.034)	0.033	0.056
Proficient	0.809 (0.025)	0.022	0.052

TABLE I
LEARNER PARAMETERS ESTIMATED FROM HUMAN LEARNER DATA.

We developed a closed-loop teaching framework to sequentially generate demonstrations, tests, and feedback to teach and evaluate the team's understanding of the robot policy. Utilzing scaffolding techniques discussed in [16], we select and sequentially introduce knowledge components (KCs). KCs are broadly defined in the education literature as "a concept, principle, fact, or skill inferred from performance on a set of related tasks" [18]. In our case, KCs represent specific characteristics or distinct constraints of the reward features. For example, the KCs could incrementally teach the bounds on the cost of traveling through rubble given the step cost, followed by bounds on the reward for recharging given the step cost, and then trade-offs between rubble and battery.

After the demonstrations, the group members are given test(s) related to the current KC, to evaluate their understanding. Ideally, human learners will be provided a test environment (see Fig. 4) and asked to map the trajectory they believe the robot will take. For our simulation study, we sample a reward weight based on the individual's PF distribution and use that as the response to the test environment. The more proportion of particles that lie within the consistent area of the test environment, the more likely the sampled weight vector gets the test response correct. In case they get the response correct, a confirmatory feedback is provided and if they get the response incorrect, a corrective feedback is provided. The β_t of the individual learners are updated according to Eq. 2b.

V. RESULTS AND DISCUSSION

Manipulation check: We wanted to ascertain the distinctiveness of the demonstrations and information provided by the various demonstration strategies. We used the surface area formed by the constraints space (yellow region in Fig. 3) of the demonstrations at each teaching session for comparison, the lower the area, the higher the information conveyed by

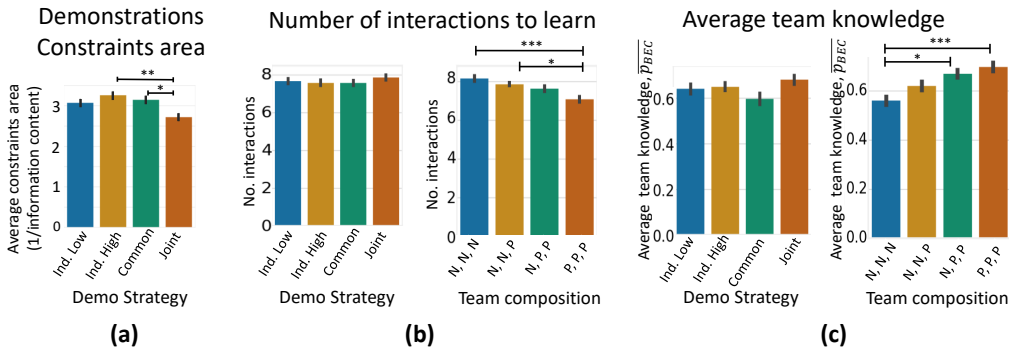

Fig. 5. Experimental results on the effects of demonstration strategy and team composition. (a) All group teaching strategies performed better than the baseline strategy of teaching individuals sequentially in terms of number of teaching sessions. No discernable difference due to strategies for number of teaching sessions. Expected differences in teams, teams with more proficient learners learned quicker, observed. (b) Noticeable differences in average team knowledge observed for strategy but differences are not statistically significant. We also observed that teams with more proficient learners had higher knowledge level. Error bars represent standard error.

the demonstrations. Fig. 5 (a) illustrates that the constraints area for the various demonstration strategies are significantly different ($p < 0.01$). Demonstrations based on joint belief result in the most informative demonstrations, whereas those based on the belief of the individual with highest knowledge result in the least informative demonstrations. This is likely because joint belief is a union of individual beliefs, and has a broader distribution resulting in diverse counterfactuals which in turn generate more informative demonstrations.

Experimental conditions results: For the four group teaching strategies, the average number of teaching sessions to learn the reward weights is $N_g = 7.67(1.68)$. Unsurprisingly, the baseline strategy of teaching each learner separately takes more teaching sessions, almost twice as much, ${N_t}^b = 17.13(2.32)$. However, contrary to our expectations, the baseline condition had a lower knowledge level, $k_b = 0.59(0.12)$ than the group conditions, $k_g = 0.64(0.21)$, though the difference is not statistically significant.

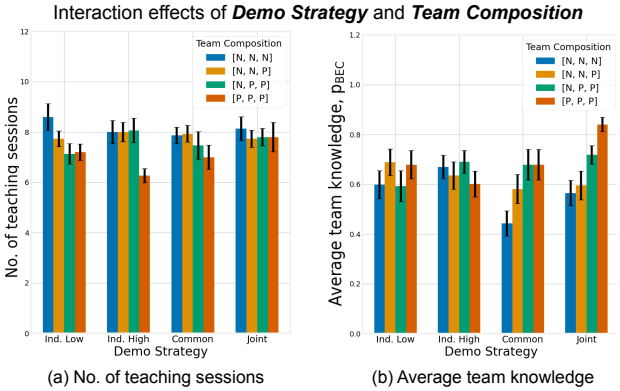

Fig. 6. Interaction effects of demo strategy and team composition on number of teaching sessions and average team knowledge. Error bars indicate standard error.

Since the baseline condition is distinctly different from the other group teaching strategies, we performed a two-way ANOVA only on the four group teaching strategies to examine their effects on team learning (see Fig. 5 (b) and (c)). The demonstration strategy did not significantly influence either the number of teaching sessions ($F = 0.42$, $p = 0.74$) or team knowledge level ($F = 1.73$, $p = 0.16$). Team composition, on the other hand, significantly influences both the number of teaching sessions ($F = 4.67$, $p = 0.00$) and team knowledge level ($F = 4.64$, $p = 0.00$), as expected. This is not surprising, as with more proficient learners, the team is likely to learn faster and better. We also found a significant interaction effect between the demonstration strategy and the team composition for team knowledge level ($F = 2.32.$, $p = 0.02$).

To further understand the significant interaction effect we found for team knowledge level, we perform a pairwise Tukey's Honestly Significant Difference (HSD) post-hoc test [25] for each combination of team composition and demonstration strategy. We found significant differences ($p < 0.05$) between teams with more proficient learners [P,P,P], [N,P,P] and no proficient learners [N,N,N], specifically for group belief strategies. For teams with more proficient learners, they perform better with group strategies, while teams with more naive learners perform similarly across all strategies.

Fig. 6 shows the interaction trends. Although there are no significant interaction effects on the number of teaching sessions, some interesting trends are observed. Individual belief strategies have more variance in the number of teaching sessions taken to learn compared to group belief strategies. However, group belief strategies, particularly the joint belief strategy, is able to accommodate diverse teams and have similar teaching durations. Group belief strategies might therefore be more suitable in situations where the proficiencies of the learners are unclear but the group has to be taught as a whole.

The interaction trends for team knowledge level shows less variance in the team's knowledge level for individual belief and more variation for group belief. This could be because of the targeted nature of the individual belief demonstrations, which adaptively samples counterfactuals for upcoming demonstrations from individual belief spaces based on their current knowledge level. This could result in bringing a

uniformity of knowledge in the individual belief conditions, particularly for the 'individual low' condition. While not significant, we can also see that individual strategies work better than group strategies for teams with more naive learners. The robot can choose the appropriate strategy from the start if proficiency information is available a priori but is more robust if it starts with a strategy and then adaptively changes the strategy based on real-time information about the learners' proficiency. These nuanced results thus highlight the need to observe and estimate learner proficiency in real-time for more effective teaching.

VI. Conclusion

In this study, we aimed to enhance the transparency and efficacy of human-robot collaboration among human groups through explainable robot demonstrations. We developed machine teaching algorithms that cater to teams with diverse learning abilities, employing team belief representations aggregated from individual beliefs, represented through particle filters. Our findings revealed that teaching strategies tailored to group or individual beliefs significantly benefit distinctly different groups characterized by varying levels of learner capabilities. Specifically, we observed that the group belief strategy and joint belief, in particular, was advantageous for groups composed mostly of proficient learners. Individual strategies were better suited for groups with mostly naive learners, though they would take more interactions. We gained deeper insights into the dynamics of group learning, thus laying the foundation for adaptively selecting teaching strategies to facilitate collaborative decision-making in real-time scenarios. However, our study has several limitations. In particular, the simulation did not evaluate the teaching algorithms across multiple domains, nor did it involve actual human learners. Our study had isolated simulated learners and does not consider the nuances of interaction within the group. These limitations highlight the simulation-centric nature of our investigation and suggest the need for empirical validation in real-world settings. Moving forward, we plan to extend this simulation study by conducting a human-subjects user study. This future research will involve actual human learners with diverse learning abilities to assess the efficacy of our teaching strategies. Additionally, we aim to explore the applicability of our methods across various domains and settings to ensure generalizability. By addressing the interaction dynamics within groups, we aspire to refine our teaching algorithms further, ensuring that they are not only effective but also adaptable to the needs of different types of learners and group compositions. By continuing to explore and refine machine teaching approaches, we anticipate contributing to the development of robots that can seamlessly integrate into human teams, enhancing both efficiency and understanding in collaborative tasks.

References

[1] X. Zhu, "Machine teaching: An inverse problem to machine learning and an approach toward optimal education," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[2] J. Jara-Ettinger, "Theory of mind as inverse reinforcement learning," *Current Opinion in Behavioral Sciences*, vol. 29, pp. 105–110, 2019.

[3] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Autonomous Robots*, 2019.

[4] M. Cakmak and M. Lopes, "Algorithmic and human teaching of sequential decision tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012, pp. 1536–1542.

[5] P. Qian and V. Unhelkar, "Evaluating the role of interactivity on improving transparency in autonomous agents," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022, pp. 1083–1091.

[6] J. Schneider and J. Handali, "Personalized explanation in machine learning: A conceptualization," *arXiv preprint arXiv:1901.00770*, 2019.

[7] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[8] A. Silva, P. Tambwekar, M. Schrum, and M. Gombolay, "Towards balancing preference and performance through adaptive personalized explainability," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 658–668.

[9] X. Zhu, J. Liu, and M. Lopes, "No learner left behind: On the complexity of teaching multiple learners simultaneously." in *IJCAI*, 2017, pp. 3588–3594.

[10] T. Yeo, P. Kamalaruban, A. Singla, A. Merchant, T. Asselborn, L. Faucon, P. Dillenbourg, and V. Cevher, "Iterative classroom teaching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5684–5692.

[11] F. S. Melo and M. Lopes, "Teaching multiple inverse reinforcement learners," *Frontiers in Artificial Intelligence*, vol. 4, p. 625183, 2021.

[12] P. Kamalaruban, R. Devidze, V. Cevher, and A. Singla, "Interactive teaching algorithms for inverse reinforcement learning," *arXiv preprint arXiv:1905.11867*, 2019.

[13] M. S. Lee, H. Admoni, and R. Simmons, "Closed-loop reasoning about counterfactuals to improve policy transparency," in *International Conference o Machine Learning (ICML) Workshop on Counterfactuals in Minds and Machines*, 2023.

[14] J. M. Carpenter, "Effective teaching methods for large classes," *Journal of Family & Consumer Sciences Education*, vol. 24, no. 2, 2006.

[15] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *ICML*, 2004.

[16] M. S. Lee, H. Admoni, and R. Simmons, "Reasoning about counterfactuals to improve human inverse reinforcement learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 9140–9147.

[17] D. S. Brown and S. Niekum, "Machine teaching for inverse reinforcement learning: Algorithms and applications," in *AAAI*, 2019.

[18] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning," *Cognitive science*, 2012.

[19] N. J. Cooke, E. Salas, J. A. Cannon-Bowers, and R. J. Stout, "Measuring team knowledge," *Human factors*, vol. 42, no. 1, pp. 151–173, 2000.

[20] C. Riedl, Y. J. Kim, P. Gupta, T. W. Malone, and A. W. Woolley, "Quantifying collective intelligence in human groups," *Proceedings of the National Academy of Sciences*, vol. 118, no. 21, 2021.

[21] T. Li, S. Sun, T. P. Sattar, and J. M. Corchado, "Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches," *Expert Systems with applications*, 2014.

[22] S. K. Jayaraman, A. Steinfeld, H. Admoni, and R. Simmons, "Adaptive group machine teaching for human group inverse reinforcement learning," *Presented at the 3rd RL-CONFORM workshop at the International Conference on Intelligent Robots and Systems*, 2023.

[23] L. K. Fazio, B. J. Huelser, A. Johnson, and E. J. Marsh, "Receiving right/wrong feedback: Consequences for learning," *Memory*, vol. 18, no. 3, pp. 335–350, 2010.

[24] S. K. Jayaraman, A. Steinfeld, R. Simmons, and H. Admoni, "Modeling human learning of demonstration-based explanations for user-centric explainable ai." in *Presented at the Explainability for Human-Robot Collaboration workshop at ACM/IEEE International Conference on Human-Robot Interaction*, 2024.

[25] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (hsd) test," *Encyclopedia of research design*, vol. 3, no. 1, pp. 1–5, 2010.