Adaptive group machine teaching for human group inverse reinforcement learning

Suresh Kumaar Jayaraman [1], Aaron Steinfeld[1], Henny Admoni[1], and Reid Simmons[1]

***Abstract*— For safe and effective collaboration between a robot and a human group, the challenge arises in teaching a diverse group of individuals about the robot's decision-making process in a potentially time-sensitive and resource-limited environment. We explore the use of robot demonstrations as a means of effectively teaching groups. We extend prior work developed for teaching individuals to teaching groups of human learners. Group teaching introduces challenges in personalization because of differing individual knowledge. We address these challenges by using aggregated team knowledge representations and developing models of team beliefs. We present several strategies for group teaching and finally propose a user study design to evaluate the learning performance and experience under these different strategies. We expect the results from this study will inform context-dependent adaptation of teaching strategies for human-robot teams.**

I. INTRODUCTION

Robots are getting more ubiquitous increasingly transitioning from tools that assist humans to collaborative teammates. For safe and effective human-robot collaboration, humans need to understand robot decision-making, often taught through demonstrations [1] in which the human, frequently modeled as inverse reinforcement learners [2], learns a robot policy from demonstrations of desired behavior. Machine teaching [3] aims to generate informative demonstrations for human learners, but most of this research has focused on single learners. In contrast, human-robot teams involve multiple people, posing additional challenges due to diverse learning abilities as the robot would have to teach its decision-making to this entire group. In this work, we explore how can robots teach a group of people using demonstrations.

Teaching a group as a whole instead of teaching each person individually is preferable, especially in large groups with limited time and resources. Take, for instance, ad hoc emergency response team (see Fig. 1 (a)) tasked with building shelters after an earthquake and aided by a robot that can bring requested items. The robot has limited maneuverability over rubble, limited range, and may prefer to recharge when possible. These capabilities and preferences (i.e. its decision-making) must be taught to the team quickly because of the time-sensitive situation. A challenge in group teaching is accommodating individuals with varied learning abilities (e.g., a mix of amateur volunteers and trained professionals in the ad hoc team) by generating common demonstrations for all. Prior work has shown that it is possible to teach a heterogeneous class using common examples [4], albeit for simple concepts. While groups can also learn from each other through communication and information sharing, here, we focus on learning from common examples only. Although group heterogeneity could also imply variations in prior knowledge, here, we assume similar prior knowledge, focusing solely on differences in learning ability.

Melo and Lopes [5] generated personalized demonstrations for each of the learners, although at a high teaching cost. In an educational setting, a teacher could develop personalization strategies based on various objectives — they could focus on slow learners, or on fast learners, or consider the class as a whole using class average or similar measures and adapt their teaching accordingly. Drawing parallels from the education literature, our key insight is that group teaching can be tailored by considering the team as a whole and can generate demonstrations based on common representations of team knowledge. While it is preferable for every team member to have perfect knowledge about the robot decision-making, not all tasks or situations might demand this. An active teacher that personalizes and adapts to the learner based on human feedback can improve learning [6], [7]. But the challenge in groups is identifying which feedback to use and to whom the personalization should cater.

In this work, we develop team belief models that facilitate group teaching focusing on the team as a whole. We introduce a closed-loop teaching framework that effectively incorporates human feedback to improve group teaching. We propose a user study design to explore how the various teaching strategies affect team learning in a situation where perfect learning is not necessary. This work on adaptive group machine teaching could generate interesting discussions surrounding bi-directional human-robot learning, use of human feedback, use of robot feedback for robot learning from humans, etc., for safe reinforcement learning practices in human-robot teams.

II. BACKGROUND

Markov Decision Process We model the environment as a Markov Decision Process (MDP), given by the tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma, \mathcal{S}_{\prime} \rangle$, representing the state space, action space, 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 [8], $R = {\mathbf{w}^*}^\top \phi(s, a, s')$ is represented as a weighted linear combination of reward features. We define a group of MDPs that share $R, \mathcal{A}$, and γ but differ in T_i, $\mathcal{S}_i$, and S_i^0, as a

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

[1]The authors are with the Robotics Institute at Carnegie Mellon University, USA. email: sureshkj, steinfeld, hadmoni, rsimmons@andrew.cmu.edu.

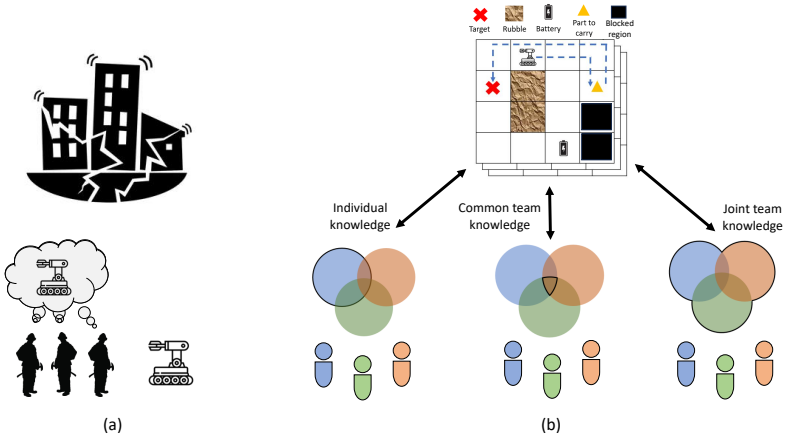

Fig. 1. (a) An ad hoc team of emergency response personnel with potentially diverse individuals is provided with a robot to help with their team tasks. To work well together safely the team has to understand the robot's decision-making, expressed through its reward function. (b) To all team members, the robot shows demonstrations to teach its reward function and gives tests to evaluate their knowledge about the robot's reward function. The demonstrations are generated based on belief about an individual's knowledge, or the team's common knowledge, which is knowledge everyone in the team has, or the team's joint knowledge, which is knowledge that at least one person in the team has. The provided demonstrations are, in turn, used to estimate the updated knowledge they would gain from seeing the demonstrations and the test responses are used to update these estimates.

domain. Sharing the same R, ensures that all demonstrations within the domain support inference over a common $\mathbf{w}^*$. We use the MDP formulation to model an *item delivery* task where a robot is tasked with delivering an item through an environment that has rubble, blocked regions, and a battery recharge station (see Fig.1 (b)).

Machine teaching for policies: We adapt the machine teaching framework for policies [9] to select a set of demonstrations $\mathcal{D}$ of size n that maximizes the similarity ρ between optimal policy π^* and the policy $\hat{\pi}$ recovered using a computational model $\mathcal{M}$ (e.g., IRL) on $\mathcal{D}$, $\arg\max_{\mathcal{D}\subset\Xi} \rho(\hat{\pi}(\mathcal{D},\mathcal{M}),\pi^*) \quad \text{s.t.} \quad |\mathcal{D}| = n$, where Ξ is the set of all demonstrations of π^* in a domain. Once $\mathbf{w}^*$ is approximated through IRL, this approach assumes that the learner can deduce π^* by planning on the underlying MDP. Thus, the objective reduces to selecting demonstrations that are informative in conveying $\mathbf{w}^*$, which can be measured using behavior equivalence classes.

Behavior equivalence class: *The behavioral equivalence class (BEC)* of a policy π is the set of reward functions under which π is optimal. For a reward function that is a weighted linear combination of features, the BEC of a demonstration ξ of π is the intersection of half-spaces [10] formed by the exact IRL equation [11]

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

where $\mu_\pi^{(s,a)} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \mid \pi, s_0 = s, a_0 = a\right]$ is the vector of reward feature counts accrued from taking action a in s, then following π after. Any demonstration can be converted into a set of constraints on $\mathbf{w}$ using (1), with each constraint being a *knowledge component (KC)* [12] that captures a facet of the reward function (e.g., tradeoffs between the underlying reward features). Consider the item delivery domain, which has binary reward features $\phi = [$*traversed rubble*, *battery recharged*, *action taken*$]$. In practice, we require $||\mathbf{w}^*||_2 = 1$ to bypass both 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 span the surface of the $n-1$ sphere (n is number of domain features) due to the L^2 norm constraint on $\mathbf{w}^*$. We instead assume that learners begins with a prior that action weight is negative (e.g. favoring shortest path, see Fig. 3 (a)).

Team modeling: A common way to represent a team characteristic such as knowledge is by aggregating knowledge of individuals. Team characteristics are normally represented as average, median, sum, range, minimum, or maximum values of the characteristic of individuals [13]. More recently, team knowledge is represented using a latent *collective intelligence* parameter that is highly correlated with team process and performance [14]. However, operationalizing such a latent parameter is difficult and we choose to represent team knowledge by aggregating individual knowledge. We focus on two aggregated representations of team knowledge — common knowledge and joint knowledge. We define **common team knowledge** as the knowledge that all team members have. It can be visualized as the intersection of individual knowledge. We define **joint team knowledge** as the knowledge that at least one individual in the team has, visualized as the union of individual knowledge (see Fig. 1 (b) for visual representations of these).

III. Methods

In this section, we discuss a particle filter-based model for human beliefs proposed in [7] that supports iterative Bayesian updates and sampling for counterfactual reasoning. We extend this approach to group teaching problem to model aggregated team beliefs. We use this model in a closed-loop teaching framework that leverages insights from the education literature, akin to [7], to adaptively generate demonstrations based on beliefs of individual and aggregated team knowledge.

A. Particle filter human belief model

We model human belief about the robot's reward weights using a particle filter, where each particle represents a potential belief about the robot's reward function and the particle weights are updated in a Bayesian manner based on constraints. The constraints correspond to expected knowledge gain for demonstrations (that the robot expected the learner to have gained after seeing the demonstration) and actual knowledge gain for tests (that the robot estimates from their test responses). This formulation enables iterative updates on human belief from demonstrations and tests.

The particle filter updates after each demonstration or test. Each demonstration generates multiple constraints by comparing the optimal demonstration against possible counterfactuals (robot behaviors that the human would've expected based on their beliefs). The test response, if incorrect, will generate a constraint by comparing the optimal trajectory with the test response. Each constraint c_i can be converted to a probability distribution $p(x_i|c_i)$ that in turn is used to update the particle weights. We use the custom probability distribution (refer Fig. 2 (a)) proposed in [7], which is a combination of a uniform distribution for the correct half-space of the constraint (indicating that any particle lying in this space is equally valid for the demonstration) and a von Mises-Fisher distribution for the incorrect half-space (indicating that particles farther away from the constraint are exponentially less likely to have generated the demonstration). We refer the reader to [7] for more details on the particle filter model. While we assume that all team members have the same prior knowledge for simplification, our approach can easily incorporate different prior knowledge.

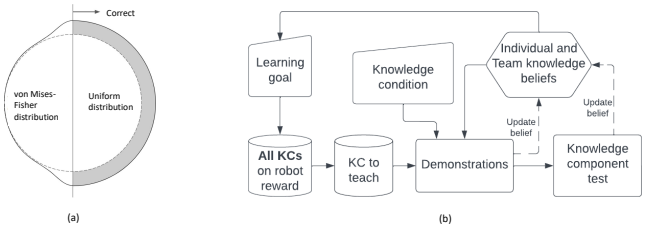

Fig. 2. (a) The custom probability density function (pdf) for updating particle weights based on a constraint generated. Figure reused from [7] with author permission. (b) The proposed closed-loop teaching framework.

B. Modeling team beliefs

We extend the human belief particle filter model to model beliefs on aggregated team knowledge. In this context, the region spanned by the particles represents aggregated team knowledge rather than individual knowledge. The primary distinction in modeling team beliefs is how the particles are updated, specifically in how individual constraints are aggregated and used for updating the particle weights. For the ad hoc team, let us assume that team 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 $\mathcal{C}_3 = \{c_3^1, c_3^2, ...c_3^n\}$. The update probability for each individual is given by, $P_i = \prod_{j=1}^{n} p(x_i^j|c_i^j)$.

We operationalize common team knowledge by considering all individual constraints and representing it as $\mathcal{C}_{ck} = \{c_1^1, c_2^1, c_3^1, c_1^2, c_2^2, c_3^2, ..., c_1^n, c_2^n, c_3^n\}$. We assume individuals to be independent. Consequently, the particle filter representing common team knowledge is updated based on the joint probability of all aggregated constraints across all tests in the set for all the individuals. $P = \prod_{i=1}^{3}\prod_{j=1}^{n} p(x_i^j|c_i^j)$. This aligns with our definition of common knowledge as the knowledge that everyone on the team has. On the other hand, joint team knowledge is operationalized by considering the set of constraints for all individuals for each test separately and is represented as $\mathcal{C}_{jk} = \{\{c_1^1, c_2^1, c_3^1\}, \{c_1^2, c_2^2, c_3^2\}, ..., \{c_1^n, c_2^n, c_3^n\}\}$. Update probabilities are calculated individually for each team member. The particles are then updated based on the maximum probability of any of the individuals, given by, $P = \arg\max_{i\in[1,2,3]} \prod_{j=1}^{n} p(x_i^j|c_i^j)$. This corresponds to our definition of joint knowledge as the knowledge that at least one team member has.

C. Closed-loop teaching

We develop a closed-loop teaching framework (see Fig. 2 (b) to sequentially generate demonstrations and tests to teach and evaluate the team's understanding of the robot policy. Using the scaffolding techniques from [15], we select individual knowledge components (KCs) that incrementally increase in information across an increasing subset of 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 these three. The demonstrations are selected based on the KCs and the knowledge condition (which individual/team knowledge to cater towards). The demonstrations and test responses are used to update the individual and team knowledge belief. Fig. 3 shows a demonstration-test loop for teaching one KC. The sequence of such demonstration-test loops are repeatedly provided by the robot until all knowledge components have been sufficiently learned by all the team members.

IV. Planned User Study

We are currently developing an online user study to explore the impact of various group teaching strategies on team understanding of the robot policy. We are considering four strategies as the *between subjects* study condition to generate demonstrations that are based on — (i) the individual with the lowest knowledge, (ii) the individual with the highest knowledge, (iii) team common knowledge, and (iv) team joint knowledge, and a baseline condition of teaching each person individually. We recognize that the baseline approach might yield the highest knowledge gain, but it would need more interactions (demonstrations and tests). The study will teach the robot policy in the *item delivery* domain for which we consider perfect knowledge is not necessary. The differences in demonstrations among these strategies arise from distinct belief models sampled from the knowledge space spanned by the particles of corresponding knowledge.

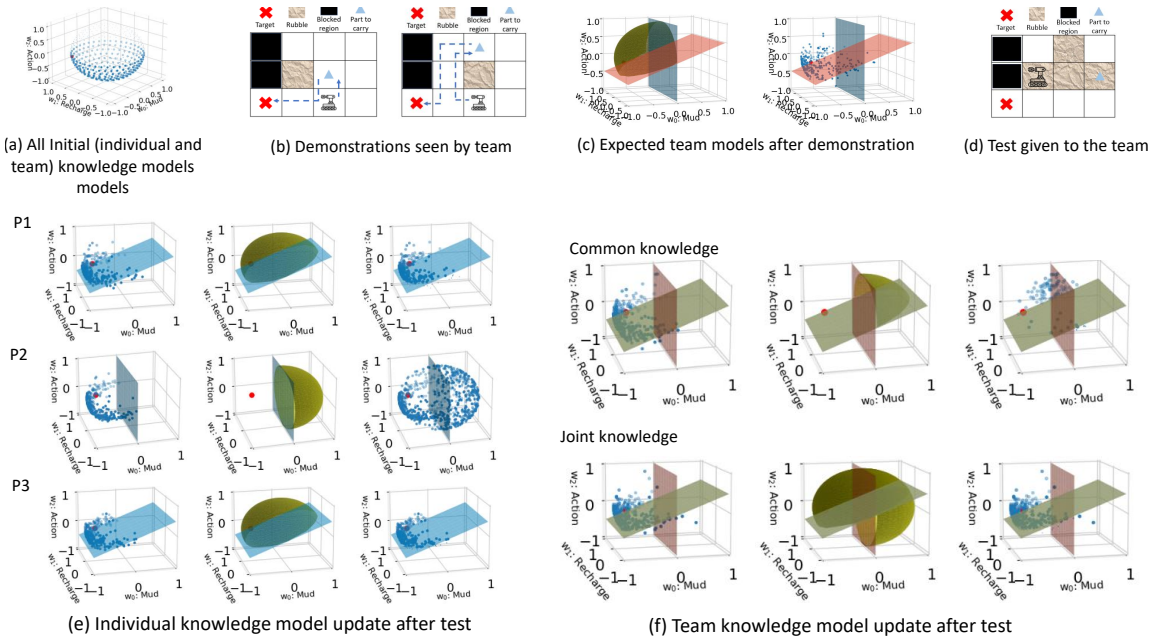

Fig. 3. A snippet of a sample teaching sequence. The individual and team knowledge starts with the same prior (a) since everyone is assumed to have the same prior knowledge. The team sees a set of demonstrations (b) and is tested if they have understood the knowledge component in the demonstrations (d). (c) The expected knowledge is the same for all individuals and the team models since everyone started with the same prior and is expected to have learned the knowledge component completely. However, in this case, person P2 got the test response incorrectly, indicating they did not learn the knowledge component correctly. The updated knowledge for each person is shown in (e) and the aggregated team knowledge is shown in (f). In both these figures, the left plot shows the current particle weights and distribution, the center plot shows the correct half-space from the corresponding constraints, and the right plot shows the updated particle distribution. The updated knowledge is used to the generate next set of demonstrations.

The associated distinct counterfactuals of the sampled belief models are used for generating informative demonstrations.

A sample demonstration and test sequence is shown in Fig. 3 for the item-delivery task domain. Initially, with individuals assumed to have the same prior knowledge, the first set of demonstrations and expected knowledge are similar across teaching strategies. The first set of tests also tends to be similar. Individual and team knowledge remain alike until someone answers a test question incorrectly. Suppose team members P1 and P3 answer correctly, while P2 answers incorrectly. In this scenario (as shown in Fig. 3 (e) and (f)), the knowledge for P1 and P3 converges toward the true reward, while P2's response skews the team's common knowledge away from it. Conversely, the joint team knowledge representation has particles close to the true reward since atleast one person answered test correctly although the particles are more spread out than P1 and P2 because of the less precise information from the joint constraints. Following the study condition, the next set of human belief models is sampled from the corresponding knowledge space. These demonstrations can vary significantly based on the study condition for the same knowledge component. This process continues until the team has learned the robot policy sufficiently.

To measure knowledge learned, we utilize the commonly used *Jaccard Index* [16]. We define knowledge learned, K_i for member i as the ratio of the intersection of the true reward region and the region spanned by the constraints/particles representing their knowledge over the union of the same. A perfect knowledge would have a value of 1 or 100%. This knowledge metric also establishes the learning goal for the user study, for example, achieving 80% of possible knowledge. Additionally, we aim to collect subjective feedback regarding workload, perception of understanding, comfort level, etc.

Our hypothesis is that group teaching based on aggregated team models will likely reach the learning goal with fewer demonstrations and receive more positive feedback compared to group teaching based on individual models, and especially when compared to the baseline individual teaching approach. We also anticipate that teaching individuals will yield the highest knowledge gain, aligning with current findings.

V. Conclusion

We explore group machine teaching in this paper, as the ability to safely collaborate and utilize robots in human-robot teams depend on the team's understanding of the robot decision-making. Leveraging the insight that team knowledge can be viewed as an aggregate representation of individual knowledge, we extend the particle filter-based human belief modeling and counterfactual reasoning-based demonstration generation developed in [7] to group teaching scenarios. We introduce methods to aggregate individual knowledge and represent team knowledge in the context of our teaching framework. Furthermore, we present a closed-loop teaching framework to effectively incorporate human feedback for generating tailored demonstrations. Finally, we propose a user study to assess the performance of various teaching strategies, based on individual and team knowledge, in learning the robot policy.

REFERENCES

[1] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Autonomous Robots*, vol. 43, pp. 309–326, 2019.

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

[3] 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.

[4] 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.

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

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

[7] 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.

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

[9] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Exploring computational user models for agent policy summarization," in *International Joint Conference on Artificial Intelligence*, 2019.

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

[11] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *International Conf. on Machine Learning*, 2000.

[12] 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*, vol. 36, no. 5, pp. 757–798, 2012.

[13] 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.

[14] 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, p. e2005737118, 2021.

[15] 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.

[16] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.