Machine Teaching of Collaborative Policies for Human Inverse Reinforcement Learning

Nyomi Morris[1], Michelle Zhao[2], Reid Simmons[2], Henny Admoni[2]

***Abstract*— In order for humans and robots to collaborate effectively, robots must be able to communicate their objectives to human partners. Understanding the robot's objective enables human partners to coordinate with the robot on a shared approach to the task. Prior work in machine teaching has examined how robots can communicate reward functions to human learners in a single-agent environment. We consider the problem of teaching a human partner a joint reward function, which captures how both human and robot should contribute to the task. This reward, which is known only to the robot, is joint over human and robot actions, and encompasses constraints over how the human and robot should contribute to a task. By adapting existing machine teaching frameworks for our collaborative domain, we seek to provide a minimal number of demonstrations such that a human can learn the rewards. We test the ability of human partners to learn an optimal collaborative policy based on demonstrations from the robot, and evaluate the effect of learning on team performance in a collaborative task. We additionally compare the effectiveness of using demonstrations at different levels of complexity to explicitly providing numeric rewards values on human learning. Results of our preliminary user study validate demonstrations as a method for teaching humans collaborative policies on both performance and comprehension levels.**

I. Introduction

As AI agents become increasingly involved in the daily lives of humans [1], it is more important than ever for humans and robots to share task responsibilities [2]. Our ability to fluently collaborate with robots is contingent on our ability to understand their objectives and decision-making. People unfamiliar with robots may not know how they should work with them, nor understand how a particular robot is designed to help them. This leads to questions such as, "How do I collaborate with this robot? Which tasks are designated for it versus me?" Consequently, it is important for the robot to communicate a collaborative approach–how human partners can or should collaborate with it. In this work, we examine how robots can communicate a joint reward function and in turn, a joint policy, to human partners through demonstrations.

A new partnership between a human and robot agent requires a period of adjustment for each collaborator to understand the capabilities of the other. We examine a paradigm where the robot partner is an expert on the optimal policy for a collaborative task. When a new human partner is paired with this robot, there must be some way for the robot to impart its expertise and constraints during the initial acclimation period. One way to facilitate this knowledge is through demonstration.

We propose an approach that extends the capabilities of machine teaching for humans in collaborative tasks. By generating a set of demonstrations, human partners will be able to learn an optimal policy. We explore how environment complexity affects the demonstrations selected and shown to the human learner. We designed and propose a user study that compares how demonstrations at different levels of complexity affect human learning and task performance. We additionally compare the efficacy of demonstration and explicit rewards. We conducted a preliminary user study which suggests that easier demonstrations are effective in teaching humans over harder demonstrations and over explicit reward values.

II. Related Work

1) Machine Teaching for Humans: Machine Teaching for Humans is an approach to forming demonstrations featuring an agent so that a human learner can understand a robot's policy [3]. The framework generates a minimal set of demonstrations to teach a robot learner the optimal reward function for a task. It modifies the Set Cover Optimal Machine Teaching (SCOT) algorithm [4], which was originally devised to form demonstrations for robot learners, to be more palatable for human learners.

Lee [3] extends this work by shifting the paradigm to generating demonstrations for human learners. In these demonstrations, the human observes a single agent and infers its policy based on how the rewards influence the agent's behavior in the environment. This work proposes a general machine teaching framework for collaborative tasks where rewards are joint across the human and robot. Therefore, the generated demonstrations feature joint actions that are distinct and independently chosen by each agent based on their own understanding of the environment.

2) Value Alignment: Value alignment is the problem of finding and maintaining a shared objective between a human-robot team [5]. Prior work in value alignment among agents exists [6], but doesn't address the complexity of humans who project additional mental states into their decision-making [5]. The value-alignment problem is therefore bi-directional, requiring each agent to behave as a teacher and learner simultaneously throughout an interaction. Techniques in inverse reinforcement learning have been used to learn agent reward functions. In Bayesian inverse reinforcement learning, the goal is to learn a maximum reward from an

[1]Authors are with Rose Hulman University, Pittsburgh, PA, USA morrisnc@rose-hulman.edu

[2]Authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA {mzhao2, rsimmons, hadmoni}@andrew.cmu.org

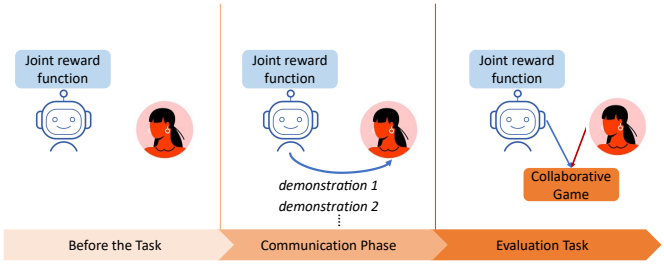

Fig. 1: Teaching humans collaborative policies is broken into 3 stages: (1) The robot becomes an expert on a joint reward function for the task. (2) The robot teaches the human the optimal policy with demonstrations generated from the reward function. (3) The human learner is evaluated based on their performance of the collaborative task.demonstration

expert [7]. This work seeks to make progress on the value alignment problem by producing a method for humans to infer the optimal policy on a collaborative task by inferring rewards through the robot expert's demonstrations.

III. Preliminaries

1) Task Domain: We developed a 3D simulated environment for a pick and place task where agents are tasked together to clean up objects from a table. At every timestep, each agent simultaneously chooses an item (a duck or lego) to pick up and place in the tray. Each agent's strategy is based on picking up items that are most rewarding for the team as a whole. The team's objective is to clean up all the items within the given time steps while also maximizing the overall reward.

2) Strategies: The clean-up task is performed optimally when each agent understands the role of their partner. Agents are free to pick up any item or choose no item. Instances where agents must choose "no item" should allow the other partner with more reward to gain for that item to pick it up. Suboptimal scenarios occur when two agents reach for the same item. This results in a failed pickup that forces the team to fully incur the step cost. This scenario represents disfluency between the agents and the human partner's misunderstanding of the rewards assigned to the object.

IV. Demonstration Generation

Our proposed interaction for teaching human collaborative task policies occurs in three stages. Prior to interaction, the robot expert develops demonstrations through the method described below. Then, the demonstrations are shown to the human, and finally, the robot and human perform the task, where the human applies their knowledge learned. (Figure 1).

A. Joint Machine Teaching for Humans

Markov Game Formulation. We model the environment as a two-player Markov game $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^h, \mathcal{A}^r, \mathcal{T}, R, \gamma, \mathcal{S}_0 \rangle$. $\mathcal{S}$ is the set of states, fully observable to both human and robot. The human and robot each have an action space $\mathcal{A}^h$ and $\mathcal{A}^r$, though in our task we consider these action spaces to be the same. $\mathcal{T} : \mathcal{S} \times \mathcal{A}^h \times \mathcal{A}^r \times S \rightarrow [0, 1]$ is the transition function, which is dependent on the joint action of human and robot, which act simultaneously in the game. γ is the discount factor, and $\mathcal{S}_0$ is the initial state distribution. The reward function $R : \mathcal{S} \times \mathcal{A}^h \times \mathcal{A}^r \rightarrow \mathbb{R}$ is linear. $R = \mathbf{w}^{*T}\phi(s, a^h, a^r, s')$ is represented as a weighted linear combination of state and joint action features. $\mathbf{w}^*$ is known only to the robot, and must be taught to the human through demonstration. We assume the human is aware of the full Markov game aside from the reward function weights $\mathbf{w}^*$. $\pi^* : \mathcal{S} \rightarrow \mathcal{A}^h \times \mathcal{A}^r$ is the optimal joint policy with respect to the reward weights $\mathbf{w}^*$ in $\mathcal{M}$.

For generating demonstrations, we aim to show the minimum set of demonstrations such that $\mathbf{w}^*$ is taught to the human. Each demonstration in the set will show the team performing actions according to an optimal policy in a particular Markov game instance, which we term a *context*. We define a context, c, as a particular instance of a Markov game $\mathcal{M}_c$, with a fixed $R, \mathcal{A}^h, \mathcal{A}^r$, and γ, but a unique $\mathcal{S}, \mathcal{T}$, and $\mathcal{S}_0$. The set of all possible contexts, C, is a group of Markov games instances that share $R, \mathcal{A}^h, \mathcal{A}^r$, and γ, but differ in $\mathcal{S}, \mathcal{T}$, and $\mathcal{S}_0$. In our collaborative decluttering domain, there is one demonstration of π^* in a context c, which is just the full, noiseless rollout of π^*.

Machine Teaching for Policies. Our objective is to select the minimum set of informative demonstrations to teach the joint policy π^*. We use the machine teaching for inverse reinforcement learning framework [4] to determine the best set of initial states from which we roll out the optimal policy which will completely teach the human the reward function R. We aim to select the set of demonstrations $\mathcal{D}$ that minimizes the following optimization problem:

$$\text{argmin}_{\mathcal{D} \in C} |\mathcal{D}| \text{ s.t. } Loss(\mathbf{w}^*, \hat{\mathbf{w}}) \leq \epsilon, \hat{\mathbf{w}} = IRL(\mathcal{D}) \quad (1)$$

Since there is one demonstration of π^* in a context c, demonstration set $\mathcal{D}$ is equivalently as set of contexts $\{c_{(0)}, c_{(1)}, ..., c_{(|\mathcal{D}|)}\}$ selected from C. The policy loss of an estimated weight vector $\hat{\mathbf{w}}$ compared with the true weight vector $\mathbf{w}^*$ is defined:

$$Loss(\mathbf{w}^*, \hat{\mathbf{w}}) = \mathbf{w}^{*T}(\mu_{\pi^*} - \mu_{\hat{\pi}}) \quad (2)$$

where π^* is the optimal policy under $\mathbf{w}^*$ and $\hat{\pi}$ is the optimal policy under $\hat{\mathbf{w}}$. This loss represents the difference in expected return between the robot's optimal policy and expected return from the policy learned by the human observer, when evaluated under the true reward weights $\mathbf{w}^{*T}$. $\mu_\pi^{(s,a)} = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) | \pi, s_0 = s, a_0 = a]$ is the vector of expected features counts that result from taking action a in state s and following π from then on. The Set Cover Optimal Teaching algorithm from [4] gives us the optimal set of demonstrations and contexts which comprise the robot's set of demonstrations which teach $\hat{\mathbf{w}}$.

B. Trading Off Context Complexity vs. Informativeness

As demonstrations are presented to the human, there may be certain kinds of demonstrations that are more effective than others. We observe a Pareto frontier (Figure 2) as we tradeoff between complexity of contexts in the demonstrations and number of demonstrations needed to teach the

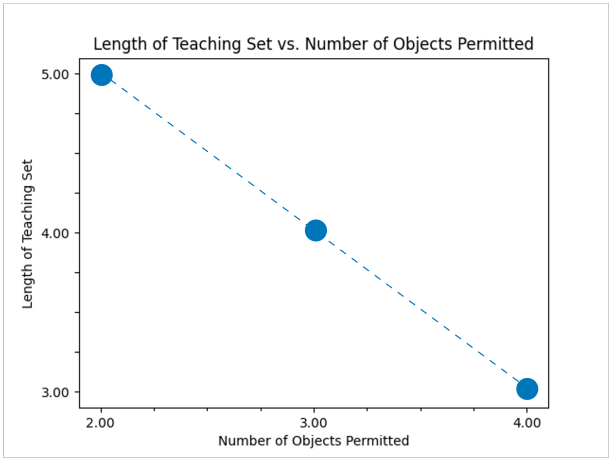

Fig. 2: This Pareto front shows an inverse relationship between number of objects permitted per type and the length of the teaching set. For our purposes, fewer objects per object type result in teaching sets with higher complexity.

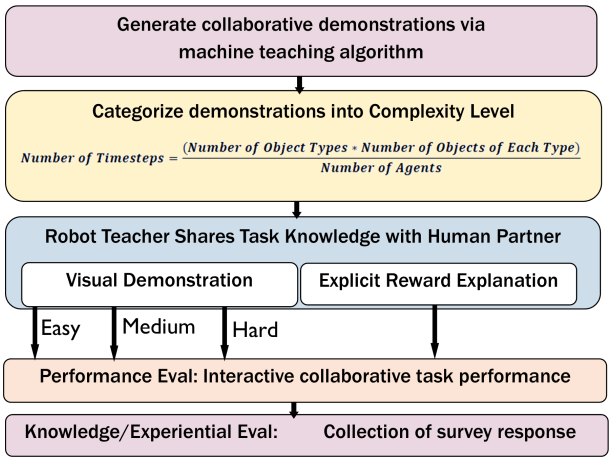

Fig. 3: The study design answers two questions: (1) How does the visual demonstration complexity influence human learning? and (2) How do visual demonstrations compare to explicit reward values in enabling humans to comprehend rewards?

optimal policy. Complexity is defined by the number of objects of each type we allow to be present in the environment. Operationally defined, when we increase complexity, allowing for more objects of each type to be present in the environment, we increase $|C|$ the size of the set of contexts over which we search for the minimum set of demonstrations. For example, when we allow for at most 4 items of each type to be present in a context c (Complexity=4), the space of possible contexts includes all contexts where at most 4 objects of each type are present, up to 3 objects of each type, up to 2 objects of each type, and up to 1 object of each type. This is necessarily more than the Complexity=2 and Complexity=3 cases.

The number of demonstrations is determined by the SCOT algorithm termination condition, where the optimal policy is taught and no additional demonstration offers more information. The Pareto frontier is a set of solutions that represents the best trade-off between all the objective functions. Every point along the Pareto frontier is optimal with respect to being able to teach the optimal policy using SCOT.

Observing that there are levels of complexity in the sets of demonstrations we can potentially show to the human partner, we aim to test whether these levels of complexity affect the human's ability to learn. Our user study thus aims to answer this question of whether the complexity affects the human's ability to learn from joint demonstrations, and whether these joint demonstrations are more effective than explicitly giving joint reward functions.

V. User Study Design

We conducted a piloted study with 12 participants on the interactive table cleanup task to evaluate the effectiveness of teaching humans through demonstration. We also evaluated which complexity level for demonstrations are best for humans to learn from. In each experimental condition, participants were then instructed to maximize the team reward by choosing an optimal action that would jointly collaborate with their robot partner's choice.

Each participant was first assigned some form of explanation provided by the robot partner based on one of four between-subjects conditions. The Pareto optimal between length of the teaching set and the number of objects per object type allowed for demonstrations to be categorized by context complexity. Those complexity levels, easy, medium, and hard, were three of the conditions that the participant could be assigned to learn from. Video demonstration sets presented the entirety of an optimal cleanup task from start to finish featuring joint actions by a robot agent and an optimal human partner. Before performing the task themselves, participants viewed demonstrations to gather implicit information about each agent's rewards.

The fourth experimental condition was in the form of explicit reward explanations where participants were provided with a score sheet of agent-specific object rewards and the step cost for actions which they could reference as they chose actions for the task (See Figure 6). This serves as a baseline against the effectiveness of learning from demonstrations alone.

Additionally, four within-subjects reward configurations were used to generate demonstrations and served as the basis for interactive games. This required participants to understand various combinations of constraints for each game and

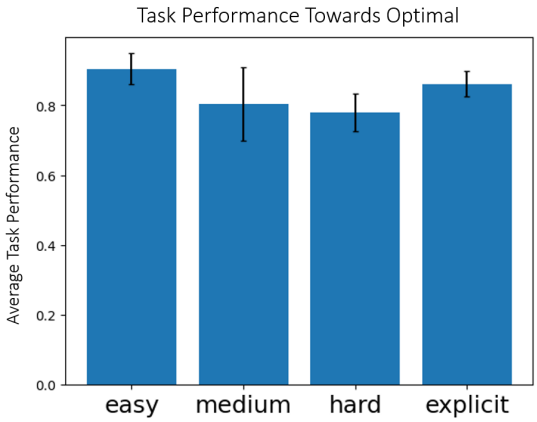

Fig. 4: Easy demonstrations proved the best for providing enough information for human partners to apply directly to the collaborative task. Errors bars represent standard error.

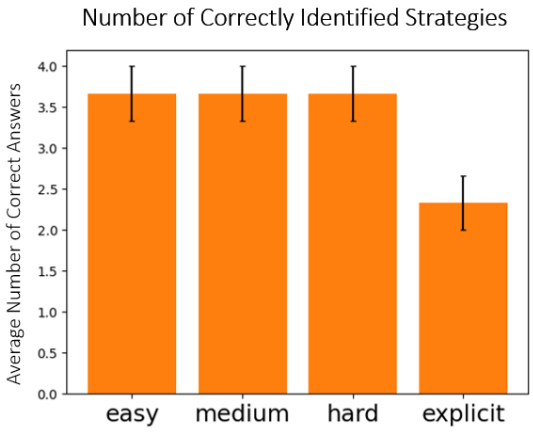

Fig. 5: Participants with explicit reward explanations were the least accurate when selecting the strategy used by the team

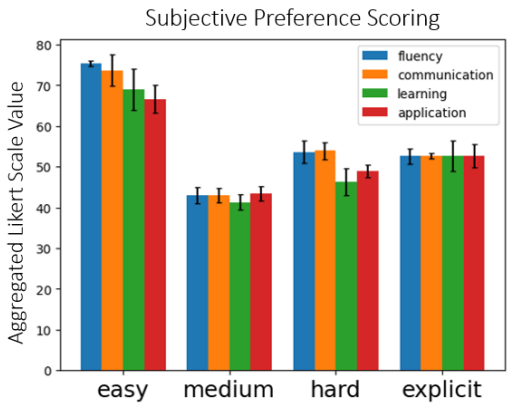

Fig. 6: Easy demonstrations out-perform all other conditions in subjective measures

respond with the appropriate action selection. After viewing the demonstration set for a given reward configuration, participants were then asked to interactively collaborate with the robot partner in a simulated table cleanup task.

A. Hypotheses

Collaborative interactions with demonstration-based robot teachers will guide users to more optimal task performance, supporting the following hypotheses:

H1: Participants will collaborate more optimally after learning from demonstrations than explicit reward explanations.

H2: Participants who see the easy demonstrations will perform the task better than those who see medium and hard demonstrations.

H3: Participants will have higher confidence in robotic agents as teachers when given demonstrations rather than explicit reward explanations.

B. Measures

Several objective and subjective measures were used to support the aforementioned hypotheses. For the subjective measures, we gathered alignment through surveying to draw experiential data about the human-robot collaboration. The set of questions related to collaborative fluency borrow from Hoffman's established set [8] and additionally include new questions for the following metrics (fluency in Fig. 6). We measured task performance by the overall team score achieved by the team after cleaning objects within the limited number of timesteps. We additionally measured participants confidence in their learning (learning), and affinity towards the robot's communication style (communication). We measured whether participants felt they were able to apply their knowledge learned from the demonstrations (application). Users asked to explicitly identify the strategy taught by the robot from a multiple choice list. This gives indication of the conscious comprehension by the user on the constraints for each agent, represented in Fig. 5.

VI. Results

We gathered preliminary results via pilot study from 12 in-person participants. Participant age ranged from 18 to 21 years old (M = 22.67, SD = 2.66). Additionally, about 66% of participants self-reported as male and 33% reported themselves as female. The four experimental conditions (easy, medium, and hard demonstrations and explicit reward explanations) distributed evenly among the participants at random.

To observe generalized learning behavior, all participants additionally interacted with four environment configurations requiring them to infer different interaction strategies. The overall team reward from each game was compared in a ratio to the known optimal score. This is how we define performance. Participants were assigned one of the four experimental conditions and data from the interactive task as well as the survey response were analyzed. An average of the performance across all participants for each experimental condition was taken. For subjective measures, the responses for each query in the subjective category was aggregated and averages were also taken across users.

Data from the pilot study indicates a subjective preference for easy demonstrations along with the most optimal task performance among all the experimental groups. Participants with explicit reward explanations performed the worst on conscious indication of the strategy (See Figure 5) being employed. Easy demonstration sets, then, seem to satisfy the capability of improving conscious awareness of the task constraints and improving task performance.

VII. Discussion

We have presented a method for robots to teach human partners collaborative policies by providing demonstrations of varying complexity levels to evaluate the human's learning. We ran a pilot study featuring 12 participants exploring the effectiveness of policy teaching via demonstration and evaluating task performance in comparison to explicit reward explanations. Initial results indicate humans perform well with demonstrations and primarily favor demonstrations at the easy level over all other explanation groups. Future work will explore more closely how the complexity of the environment influences task performance in demonstration groups versus explicit reward groups.

We are supported by DARPA FP00002636, and NSF IIS-2112633. Thank you to Ravi Pandya for directional guidance.

REFERENCES

[1] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of human-robot interaction*, vol. 3, no. 2, p. 74, 2014.

[2] L. Fiorini, M. De Mul, I. Fabbricotti, R. Limosani, A. Vitanza, G. D'Onofrio, M. Tsui, D. Sancarlo, F. Giuliani, A. Greco *et al.*, "Assistive robots to improve the independent living of older persons: Results from a needs study," *Disability and Rehabilitation: Assistive Technology*, vol. 16, no. 1, pp. 92–102, 2021.

[3] M. S. Lee, H. Admoni, and R. Simmons, "Machine teaching for human inverse reinforcement learning," *Frontiers in Robotics and AI*, vol. 8, p. 693050, 2021.

[4] D. S. Brown and S. Niekum, "Machine teaching for inverse reinforcement learning: Algorithms and applications," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7749–7758.

[5] J. F. Fisac, M. A. Gates, J. B. Hamrick, C. Liu, D. Hadfield-Menell, M. Palaniappan, D. Malik, S. S. Sastry, T. L. Griffiths, and A. D. Dragan, "Pragmatic-pedagogic value alignment," in *Robotics Research: The 18th International Symposium ISRR*. Springer, 2020, pp. 49–57.

[6] T. K. Büning, A.-M. George, and C. Dimitrakakis, "Interactive inverse reinforcement learning for cooperative games," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2393–2413.

[7] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning." in *IJCAI*, vol. 7, 2007, pp. 2586–2591.

[8] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019.