Cooperative Robot Teaching

Anonymous Author(s)
Affiliation
Address
email

Abstract: Knowledge and skills can transfer from human teachers to human students. However, such direct transfer is often not scalable for *physical* tasks, as they require one-to-one interaction and human teachers are not available in sufficient numbers. Machine learning enables robots to become experts and play the role of teachers to help in this situation. In this work, we formalize *cooperative robot teaching* as a Markov game, consisting of four key elements: the target task, the student learning model, the teacher model, and the interactive teaching-learning process. Under a moderate assumption, the game reduces to a partially observable Markov decision process (POMDP), with an efficient approximate solution. We illustrate our approach on two tasks, one in a simulated video game and one with a real robot.

Keywords: Robot Teaching, Human Robot Interaction

1 Introduction

Learning complex skills in a multi-player setting is more challenging than its counterpart in a single-agent game: the learner can no longer learn to cooperate by simply observing the demonstrations; it has to act through the motions and practice by interacting with a specialized teacher to master the skills. Learning *collaborative physical skills* forms a major part of this teaching scenario. Referring to table co-reorientation shown in Figure 1 as an example, to cooperatively reorient the table with partners of different intentions and adaptiveness, human normally learns through interacting and practicing the skills with representative teachers, rather than exhaustively collecting demonstrations to cover all possible behaviors of the partners. Due to the reliance on a human teacher, scaling up this teaching process is heavily constrained by the limited human teachers [1]. We aim to fundamentally remove this constraint by finding an alternative source for teachers. Fortunately, with the recent advancement in the machine learning community, robots now can not only master various tasks [2, 3, 4], but are also capable of collaborating with humans and adapting to humans' behaviors efficiently [5, 6, 7]. More importantly, robots are innately scalable. To this end, we propose a conceptual robot teaching framework that aims to teach humans cooperative skills through human-robot interaction (HRI).

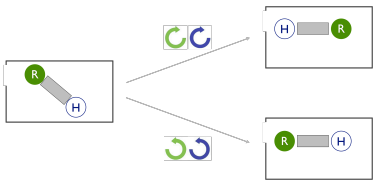


Figure 1. The table co-reorientation task. Blue circle H and green circle R represent the human and the robot, respectively. Both agents are working together to reorientate the table.

The challenge of teaching a skill in a cooperative setting lies in three aspects: representing the skill, identifying a proper mode of interaction, and generating an efficient curriculum. Previous efforts [8, 9, 10, 11] represent the skills by demonstrations. They focus on how to select the demonstrations or training samples optimally. However, given the interactive nature of the task, it is unrealistic to cover all possible demonstrations for the student to learn from. Hence, we need a more compact representation of the skills. In addition, existing works on human-robot interaction, such as shared autonomy [12, 13], are assistive and collaborative in nature, where the robot adapts to humans' actions and assists humans. Surprisingly, we find out that this assistive paradigm in human-robot

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

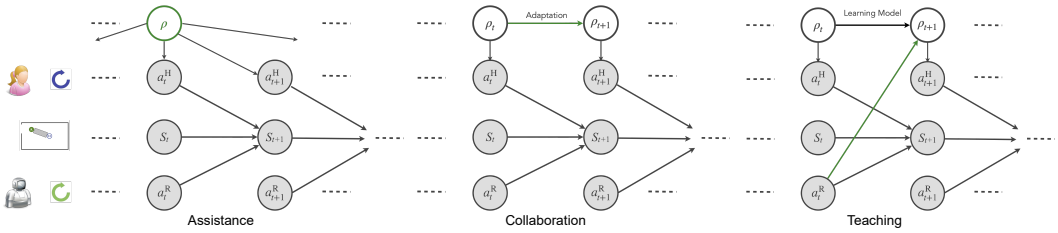


Figure 2. Simplified graphical model for Assistance, Collaboration, and Teaching in the table co-reorientation task. We use ρ to represent human's latent state, s to represnet the task state, a^H and a^R to represent human's action and robot's action respectively. While human's latent state can also be affected by robot action in collaboration, we use the arrow in teaching to emphasize robot's intention to teach human.

interaction could be overly protective and thus hinder humans from acquiring new skills. Therefore, robot teaching requires a new mode of interaction that can 1) hint the human towards the task completion, and 2) motivate the student to acquire new skills. On top of that, customizing the curriculum for individuals is also challenging: it usually demands an experienced teacher to infer students' proficiencies and learning behaviors, so it can design a curriculum tailored to individuals.

Given the aforementioned challenges, we draw insights from student-centered learning [14] and human-robot cross-training [15] to tackle the teaching problem. Rather than representing the skill by its optimal behaviors across all scenarios, we choose to decompose the skill into a fixed set of linear independent sub-skills, whose proficiencies are easy to evaluate. With such a compact and decomposable skill representation, we can naturally derive a partially assistive robot teaching curriculum, whereas each sub-skill is learned one at a time, with the robot assisting the rest of the unlearned sub-skills during the teaching. Moreover, to teach each sub-skill, our teacher induces active learning behavior from the learner so it optimizes the sub-skills by itself [16], removing the burden to specify the optimal actions for the learner to imitate. Our partial assistive mode of interaction no longer suffers from the "lazy student" issue. More importantly, this partially assistive teaching paradigm eases both the overall learning and the proficiency evaluation: we decompose the learning of any complex skill into a set of independent sub-tasks, so they can be tackled one by one.

To this end, we conclude the main contribution of this work as presenting a conceptual framework, Cooperative Robot Teaching, which offers a formal model to describe and analyze robot teaching in a cooperative task. Key elements in the framework are identified, which makes it possible for one to simplify and solve the problem by applying existing algorithms or devising new algorithms. In this work, as a first attempt, we demonstrate an instantiation of solutions to it. We conducted two human-subject experiments to show the effectiveness of Cooperative Robot Teaching.

2 Related Work

Assistance in HRI. One major aspect of HRI is how the robot could assist humans with a hidden human objective [12, 13]. As shown in Figure 2, if the human chooses to rotate the table in a counter-clockwise way and insists on his preference, the objective of the robot is to infer the human's intention and learns to assist the human. In its simplest form, the action selection and human intention inference are separated [17, 18, 19]. A decision-theoretic framework, assistant POMDP, is developed to capture the general notion of assistance in HRI [20]. The robot integrates the reward learning and control modules to perform sophisticated reasoning over human feedback [21, 22]. However, both these two approaches neglect human learning/adaptation and may hinder humans from improving their skills. On the contrary, our work focuses on how to generate behaviors that facilitate human learning during interaction.

Collaboration in HRI. Another important aspect of HRI is to model interactions as the collaboration between the human and the robot [23], for which the human and the robot share the same objective. Consider the table-reorientation task shown in Figure 1, both human and robot's objective is to reorient the table quickly. However, the joint optimal policy, e.g. rotating the table counter-clockwise, is unknown to both agents in the first place. Their interaction is mutually adaptative [24, 25, 26]. Particularly, as pointed out in [7], if one side is only aware of partial information about the task, the

optimal policy pair naturally induces the behavior of active teaching, active learning, and efficient communication between the robot and human. In this work, we focus on the following setting: given that the robot teacher knows the optimal policy, how to design an interactive teaching strategy that can induce active learning from the student.

Teaching Algorithm for Algorithms. Teaching for algorithms aims to facilitate the learning of the algorithm by choosing or generating training samples. Various teaching techniques including curriculum learning [27] and machine teaching [28, 29, 30, 31, 32] have been effectively applied to supervised learning and semi-supervised learning problems. Similar ideas are further extended to train reinforcement learning agents to learn complex skills, e.g., generate training environment for reinforcement learning [33, 34, 35], choose various demonstrations [36] or learn to decompose the skill [37, 38]. Teaching in cooperative multi-agent RL allows agents to simultaneously become teachers and students for each other [39, 40, 41]. However, such approaches generally require relatively more data for training and to some extent the controlled learning behavior of the learner. Transfer of these approaches to human learning is promising but difficult.

Teaching Algorithm for Human. Despite the aforementioned practical challenges, some algorithms have been successfully deployed for human learning. Attempts on teaching the crowd on classification or concepts prove to be successful [11, 42, 43, 44]. While humans can learn concepts from visual or verbal examples, complex skills like motor control skills can hardly be mastered through these signals. Here, we seek to automate the teaching process for humans to cooperate in a physical task, e.g., table co-reorientation.

3 Cooperative Robot Teaching

We formalize Cooperative Robot Teaching by identifying four key elements: (1) the target task, (2) the student learning model, (3) the teacher model, and (4) the interactive teaching-learning process.

The target task. In this paper, we focus on teaching in *a duo cooperative task* and we call it the *Target Task*, which is the original cooperative game both agents aim to solve.

Definition 1 ***(The Target Task)****. The target task is a duo player cooperative Markov game between a teacher, T, and a student, S, that can be described by a tuple, $\mathcal{M} = \langle \mathcal{S}, \{A^T, A^S\}, \mathcal{T}(\cdot|\cdot,\cdot,\cdot), R(\cdot,\cdot,\cdot), \gamma \rangle$ with the following definitions:*

- *$\mathcal{S}$ a set of target task states: $s \in \mathcal{S}$.*
- *A^T a set of actions for the teacher, T: $a^T \in A^T$.*
- *A^S a set of actions for the student, S: $a^S \in A^S$.*
- *$\mathcal{T}(\cdot|\cdot,\cdot,\cdot)$ a conditional distribution on the next target task state, given the previous state and the actions of both agents: $\mathcal{T}(s'|s, a^T, a^S)$.*
- *$R(\cdot,\cdot,\cdot)$ a target task reward function that maps target task states and players' actions to real numbers. $R : \mathcal{S} \times A^T \times A^S \to \mathbb{R}$.*
- *γ the discount factor in the target task.*

At each step, T and S both observe the current task state s_t, then select their actions $a_t^T \sim \pi^T$ and $a_t^S \sim \pi_t^S$ respectively, and receive a joint reward $r_t = R(s_t, a_t^T, a_t^S)$. The next state is updated as $s_{t+1} \sim P_{\mathcal{T}}(s_{t+1}|s_t, a_t^T, a_t^S)$. Next, the student updates its policy π_{t+1}^S by observing the teacher's action a_t^T and the reward r_t, and the process repeats.

Given the definition of the target task, we first answer how to represent the knowledge/skills. In this work, we choose to represent a *skill* by the optimal policy π^* to the target task. The optimal policy maximizes the expected cumulative reward when the student is cooperating with a given partner with policy π, and is defined as $\pi^* = \arg\max_{\pi^S \in \Pi} \mathbb{E}_{\substack{a_t^T \sim \pi, \\ a_t^S \sim \pi^S}} [\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t^T, a_t^S)]$. For example, in the table co-reorientation task, the student needs to learn to deal with either stubborn or adaptive partners. We recognize that there are other ways to represent knowledge/skills, such as a set of demonstrations and the ground-truth reward function. However, such representations are indirectly linked with the skill's performance; therefore, evaluating its proficiency is more obscured. We choose the optimal policy as the representation since it can be directly optimized over and evaluated.

The student learning model. The student policy is non-stationary since it will improve along with teaching. Therefore, we use a tuple of initial policy and updating strategy (learning strategy),

$\langle \pi_0^S, \mathcal{U} \rangle$, to represent the student behavior.

The teacher model. We define the teacher as a knowledgeable agent (expert) who knows the student's optimal policy π^* for a target task. The teacher aims to acquire a teaching policy π^T that can teach π^* to the student effectively. The teacher, the student, and the target task share the same action space. Taking the table co-reorientation task as an example, the teacher may take actions to assist the student or intentionally expose the student to unseen scenarios. In short, the teacher can be described by a tuple of the student's optimal policy and the corresponding teaching policy, $\langle \pi^*, \pi^T \rangle$.

The interactive teaching-learning process. In the target task, T knows the student's optimal policies π^* while the student does not. The task of T is to act in the most informative way so that S learns π^* fastest. To embed the objective of teaching and distinguish it from the *Target Task*, we define it as the *Teaching Task* in the following way:

Definition 2 ***(The Teaching Task)****. The teaching task is a POMDP $\mathcal{M}'$ for the teacher given the target task $\mathcal{M} = \langle \mathcal{S}, \{A^T, A^S\}, \mathcal{T}(\cdot|\cdot,\cdot,\cdot), R(\cdot,\cdot,\cdot), \gamma \rangle$ and the student $\langle \pi_0^S, \mathcal{U} \rangle$. It can be described by a tuple $\mathcal{M}' = \langle \hat{\mathcal{S}}, \hat{A}, \hat{\mathcal{T}}, \mathcal{O}, Z, \hat{R}, \hat{\gamma} \rangle$ with the following definitions:*

- $\hat{\mathcal{S}}$ *a set of teaching task state: for* $\hat{s} \in \hat{\mathcal{S}}$, $\hat{s} = \{s, \pi^S\}$.
- $\hat{A}$ *a set of actions:* $\hat{A} = A^T$.
- $\hat{\mathcal{T}}(\cdot|\cdot,\cdot)$ *a conditional distribution on the next teaching task state:* $\hat{\mathcal{T}}(\hat{s}'|\hat{s}, a^T)$.
- $\mathcal{O}$ *a set of observations: for* $o \in \mathcal{O}$, $o = \{a^S, s, r\}$.
- Z *the observation probability function:* $Z = \{\pi^S, \mathcal{T}(\cdot|\cdot,\cdot,\cdot), R(\cdot,\cdot,\cdot)\}$.
- $\hat{R}(\cdot,\cdot,\cdot)$ *the teaching task reward function that measures the effectiveness of teaching.* $\hat{R} : \hat{\mathcal{S}} \times \hat{A} \times \hat{\mathcal{S}} \to \mathbb{R}$.
- $\hat{\gamma}$ *the discount factor in the teaching task.*

The objective of the *Teaching Task* is to derive a teaching policy enabling student to learn π^* for the *Target Task* fastest. More specifically, the teacher can influence the student through interactive actions a^T and the joint reward r. Given the student initial policy π_0^S and the update function $\mathcal{U}$, the goal of the *Teaching Task* is to find a teaching policy π^T that allows $\pi_0^S \to \pi^*$ as fast as possible.

Next, we introduce our choice of the teaching policy π^T, the update function $\mathcal{U}$, and the reward function $\hat{R}$. To devise a student-aware teaching strategy, apart from the current state s_t and the target policy π^*, our π^T also takes the previous student action a_{t-1}^S and reward r_{t-1} as input, i.e., $a_t^T \sim \pi^T(a_t^T|s_t, a_{t-1}^S, r_{t-1}, \pi^*)$. S updates π^S with any arbitrary iterative functions conditioned on the history of interactive actions and the joint rewards in the following form: $\pi_{t+1}^S = \mathcal{U}(\pi_t^S, [(s_0, r_0, a_0^T, a_0^S), ..., (s_t, r_t, a_t^T, a_t^S)])$. Moreover, to incentivize the teacher to speed up the teaching process, we introduce a step-wise teaching cost to the teacher $c_t = \mathcal{C}(s_t, a_t^T)$ to penalize unnecessary teaching actions. To this end, we define the reward function as

$$\hat{R}(\hat{s}, a_t^T, \hat{s}'; \mathcal{D}, \mathcal{C}, \pi^*, \omega) = \mathcal{D}(\pi_t^S, \pi^*) - \mathcal{D}(\pi_{t+1}^S, \pi^*) - \omega \mathcal{C}(s_t, a_t^T),\ \hat{s} = \{s_t, \pi_t^S\},\ \hat{s}' = \{s_{t+1}, \pi_{t+1}^S\}, \quad (1)$$

where ω is the weighted factor to trade-off the teaching cost and teaching efficiency, and $\mathcal{D}(\cdot,\cdot)$ can be an arbitrary distance measure between two policies, e.g., initial state value in the target task. The solution to the POMDP $\mathcal{M}'$ is a teaching policy π^T that maximizes the expected sum of rewards $\mathbb{E}_{a_t^T \sim \pi^T}[\sum_{t=0}^{\infty} \hat{\gamma}^t \hat{R}(\hat{s}, a_t^T, \hat{s}')]$.

4 Our method

Our solution to the teaching task can be summarized into 3 steps: (1) representing the teacher's actions by decomposition into sub-skills, (2) representing hidden states, transition, and reward with Item Response Theory (IRT) [45] and Knowledge Tracing (KT) [46], and (3) learning the model through interactions and generating training sequences. Our solution is summarized in Algorithm 1.

4.1 Representing Actions by Decomposition into Sub-skills

Sub-skills decomposition is well-studied for single-agent tasks [47, 48, 49]. However, extending the same idea to a multi-agent setting is still challenging since task completion relies on the interaction among multiple parties. We observe that in a multi-agent game, the task naturally comprises several roles, of which each agent takes a subset. The well-established leader-follower model is a particular choice of role-based skill decomposition [50, 51, 52, 53]. Therefore in our work, we propose to

decompose skills based on role allocation. We divide the skill into K locally-independent teachable sub-skills according to the student's potential roles in the task. In the table co-reorientation example, the role of the student can be the proactive agent who leads the rotation or the passive agent who always follows. The teacher's action space $\hat{A} = \{k : k \in \mathbb{Z},\ 0 \leq k < K\}$ consists of teaching each sub-skill. Such a decomposition of skills naturally leads to a partially assistive mode of interaction.

4.2 Representing Hidden States

Item Response Theory (IRT) models the student responses to a certain item [45]. Given the limited interactions, we adopted the simplest form, the one-parameter model (1RL), to model human skills. In the 1RL model, each sub-skill $k \in \hat{A}$ is assigned a parameter β^k representing the difficulty, and a parameter α^k called the *proficiency* representing a student's skill state. The probability that a student has mastered sub-skill k is given by $f(k) := \sigma(\alpha^k - \beta^k)$, where σ is the sigmoid function. Hence, instead of representing the state with student's policy π^S, we use $(\alpha^k, \beta^k)^K$ to represent the hidden state. That is, for $\hat{s} \in \hat{\mathcal{S}}, \hat{s} = \{s, (\alpha^k, \beta^k)^K\}$. In this work, we assume a bounded positive reward function. The ratio between the target task rewards achieved by the student's current and optimal policies when the teacher teaches sub-skill k is used to infer $f(k)$:

Algorithm 1 Approximated Solution to the Teaching Task

Require: Maximum Interactions L, Predefined Interactions N

for $k \in \hat{A}$ **do**:
  Randomly initialize λ^k and α_t^k, β^k, and $C_k = \{\}$
  for $i = 1, 2, ..., N$ **do**:
    C_k.add($f_i(k)$)
  end for
end for
for $i = 1, 2, ..., L$ **do**:
  for $k \in \hat{A}$ **do**:
    Learn λ^k and α_t^k, β^k from C_k
  end for
  $k \leftarrow$ Action selection from λ^k and α_t^k, β^k
  $f_i(k) \leftarrow$ Performance measure from interactions
  C_k.add($f_i(k)$)
end for

$$f(k) := \frac{1}{1 + e^{(\beta^k - \alpha^k)}} = \frac{R(s, a^T = k, a^S)}{R(s, a^T = k, a^*)}, \tag{2}$$

where a^* refers to the action generated by the optimal policy that the teacher aims to teach. For each student and each $k \in \hat{A}$, we assume that α^k changes over time while β^k is a constant.

4.3 Representing the Transition

Knowledge Tracing (KT) is a technique used to model a learner's acquisition of certain knowledge [46]. Following the previous work on online estimation of student proficiency [54, 55], we model the student's proficiencies over time on the sub-skill k as a Wiener process

$$P(\alpha^k_{t+\Delta t} | \alpha^k_t) = \exp\left(-\frac{(\alpha^k_{t+\Delta t} - \alpha^k_t)^2}{2\lambda^k \Delta t}\right), \quad k \in \hat{A}, \tag{3}$$

where Δt refers to the step interval and λ^k is a parameter controlling the "smoothness" with which student's proficiency varies over time. For each student and for each $k \in \hat{A}$, we assume λ^k to be a constant.

4.4 Representing the Reward

The distance between the student's policy and the optimal policies can be represented by $f(k)$. We represent the distance as the average of one minus master probabilities of each sub-skills $\mathcal{D}(\pi^S, \pi^*) = \frac{\sum_{k=0}^{K} 1 - f(k)}{K}$. There are other ways to specify the goal according to the decomposition of the skill, e.g. weakest or multiply [56]. We choose the sum due to our local-independence assumption on sub-skills. In this work, we assume the cost is uniform, thus, given a finite horizon of interactions, maximizing the reward function defined in Equation (1) is equivalent to $\hat{R}(\hat{s}, a_t^T, \hat{s}') = \frac{\sum_{k=0}^{K} f_{t+1}(k) - f_t(k)}{K}$.

4.5 Model Learning and Decision Making

We use the student's performance during the interactions to estimate both λ^k and α_t^k, β^k. Let $f_{1:t}(k)$ denote sequences of student's performance measure against the expert. We have the

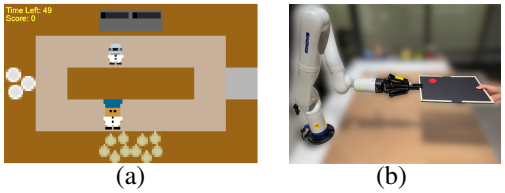
posterior $P(\lambda^k, \alpha_t^k, \beta^k | f_{1:t}(k)) \propto P(f_{1:t}(k) | \lambda^k, \alpha_t^k, \beta^k) P(\lambda^k, \alpha_t^k, \beta^k)$. The conditional probability of the observation and current proficiency can be obtained by integrating out all the previous proficiencies. The likelihood can be approximated through $P(f_{1:t}(k) | \lambda^k, \alpha_t^k, \beta^k) \approx \prod_{t'=1}^{t} \int P(f_{t'}(k) | \lambda^k, \alpha_{t'}^k, \beta^k) P(\alpha_{t'}^k | \alpha_t^k) \mathrm{d}\alpha_{t'}^k$. An approximation of the log posterior over the student's current proficiency given previous responses can be derived to learn the parameters λ^k and α_t^k, β^k. In this work, we employ maximum a posteriori estimation (MAP) to learn these parameters. Given the estimation of current state using the past history, we use one-step look-ahead to reduce the impact of the inaccuracy in the transition function. At timestep t, the teacher's action is given as $a_{t+1}^T = \arg\max_{k \in \hat{A}} \int P(\alpha_{t+1}^k | \alpha_t^k) f_{t+1}(k) \, \mathrm{d}\alpha_{t+1}^k - f_t(k)$. In practice, the student is asked to perform on each sub-skill for N interactions to initialize the parameters.

4.6 Training on Sub-skills

Our overall strategy for training students on each sub-skill is to diversify scenarios the student would encounter during training. Training students on sub-skills naturally leads to a partially assistive partner on unlearned sub-skills, which allows the student to explore the sub-skill freely. We adopt an intuitive assumption: *an agent learns cooperation better with a diverse group of partners*. Such a teaching strategy is effective when dealing with synthetic students [57, 58]. The student could learn from a diverse set of partially assistive partners or learn to cope with them by acquiring new skills.

5 Experiments

We carried out two human-subject experiments to demonstrate how Cooperative Robot Teaching works, one in simulation (Overcooked-AI [59]) and the other with a real robot (Duo Ball Maze). Experiment setups are shown in Figure 3. We investigated the teaching performances of three types of teachers: the **fully-assistive** teacher who performs optimally concerning the student's initial capability, the **student-aware** teacher who behaves according to our teaching strategy, and the **random** teacher who chooses the sub-skill (in Duo Ball Maze) or executes task actions randomly (in Overcooked-AI).



(a) (b)

Figure 3. Experiment setups. (a) Overcooked-AI layout: human participants control the "chef" and the robot controls the "robot". (b) The real robot setup of Duo Maze Ball game.

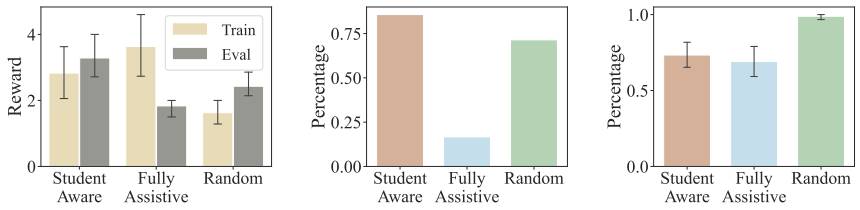
5.1 Setups

Overcooked-AI. Overcooked-AI is a benchmark environment for fully cooperative human-AI task performance and has become a well-established domain for studying coordination [60, 61, 62, 63]. The goal of the game is to cook and deliver as much soup as possible in a limited time. We decompose the policy into two sub-skills: *putting ingredients in the pot* and *delivering the soup*. To put ingredients in the pot, there exists one *efficient strategy* to pass the ingredient through the middle table. We recruited N=20 (8 females and 12 males) participants and randomly assigned them into groups of three, each with a different teaching strategy. Each participant was trained for 5 games and then evaluated for 1 game.

Duo Ball Maze. The Duo Ball Maze game requires coordination from both the robot and the human. Each party will hold one side of the maze board and tilt it to move the ball out from one of the two exits. We define two sub-skills *leading the rotation* and *following the rotation*. We recruited N=19 (6 females and 13 males) participants to carry out human-subject experiments. Data from one male participant was discarded due to a hardware issue during the experiment. The participants were first evaluated in the two sub-skills, then trained for 20 interactions, and finally evaluated in the two sub-skills again. Details can be found in the supplementary materials.

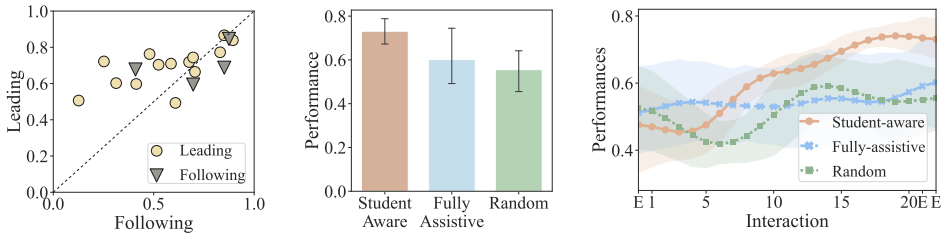
5.2 Results

A fully-assistive teacher impedes human's acquisition of skills. In the Overcooked-AI experiment shown in Figure 4(a), we observe that the students trained with a fully-assistive teacher perform worse than the students with a random teacher: it seems that a student becomes "lazy" and free rides



(a) (b) (c)

Figure 4. Results of the Overcooked-AI experiment. (a) Rewards achieved together by the human-robot pairs during training and evaluation. The error bars correspond to the 95% confidence intervals (95%CI). (b) Percentage of students who found the efficient strategy. None of the students are aware of this strategy at the beginning of the training. (c) Percentage of reward achieved by the human participants during training.



(a) (b) (c)

Figure 5. Results of the Duo Ball Maze experiment. (a) Evaluation performances of the two sub-skills of all participants. The marker styles correspond to the sub-skill preferences of the participants. (b) Evaluation performances. The error bars correspond to the 95%CIs. (c) Performances with respect to training progress. The shaded areas correspond to the 95%CIs. "E" in the horizontal axis means evaluation round.

the teacher when the teacher unilaterally adapts to the student and performs optimally. We further investigate the learning pattern of the "lazy student" problem and find out that *this "laziness" does not lie in the student's reluctance to take actions, but rather in the lack of motivation to explore and improve*. In Figure 4(c), we show the percentage of reward achieved by the student in Overcooked-AI during training. Compared with the student-aware counterpart, the percentage of reward achieved by humans is similar. However, only 17% of the participants of the group find out the efficient strategy (Figure 4(b)), which is crucial to achieve high scores when cooperating with sub-optimal partners. We observe a similar trend in the Duo Ball Maze game shown in Figure 5(c). The performance of students in the fully-assistive group shows marginal improvement with low fluctuations.

Humans may not learn effectively from a random teacher. A random teacher is incapable of teaching a truly cooperative task. Our Overcooked-AI and Duo Ball Maze differ in their requirements for the degree of cooperation: Overcooked-AI can be done by a single agent (one agent can perform both sub-skills to finish the task, at the expense of yielding a lower score), while the Duo Ball Maze need both agents to perform consistent actions. As a result, we observe that while the student can still learn from a random teacher in Overcooked-AI (see Figure 4(a)-(b)), there is no signal of learning in Duo Ball Maze (see Figure 5(b)-(c)). This suggests that a more dedicated teacher is required to teach a truly cooperative task effectively.

Partially assistive or random partner motivates students to explore new strategies. By leaving some/all work to the student, partially assistive and random teachers both motivate the student to acquire new skills. This is shown in Figure 4(b) that most of the students under these two teachers can find out the efficient strategy in Overcooked-AI. However, their performance and the robustness of the learned strategies differ significantly. Though multiple explanations could account for it, we hypothesize the student under the random teacher learns a single fixed strategy to finish the task alone (Figure 4(c)). Such a strategy that completes the task alone cannot utilize the possibly helpful inputs from the partner, therefore resulting in a poorer performance score. On the other hand, the student-aware teacher exposes the student to various scenarios by generating diverse actions on each sub-skill. For example, if the teacher acts as the leader, the student learns to follow; if the teacher does not take the initiative to act, the student explores the leading role, which is directly reflected in the performance improvements shown in Figure 5(c). In addition, students are granted more opportunities to practice and learn a sub-skill dedicatedly. In summary, the role-based partially assistive teacher enjoys the best of both random and fully-assistive teachers. As shown in Figure 4(a), the

student-aware teacher outperformed the fully-assistive and the random teachers in the Overcooked-AI experiment in terms of the evaluation reward (with p-values 0.002 and 0.06). In the Duo Ball Maze game shown in Figure 5(b), the student-aware teacher also outperformed both fully-assistive and random teachers (with p-values 0.133 and 0.017).

An individualized curriculum should be designed for the student. The preferences and biases of each student can differ significantly and they may play a vital role in the teaching task. In the post-experiment survey of Duo Ball Maze, we asked the participants "which mode of the robot is easier to cooperate with?". Out of the 18 participants, 4 participants preferred to follow the robot and 14 participants preferred to lead the robot. Moreover, as we evaluated the student performance with partners of different sub-skills, we found that the student performances were consistent with their declared preferences (Figure 5(a)). That is to say, the student may have a bias over which strategy to acquire, and tailoring the teaching curriculum to focus on that specific strategy is efficient and more intuitive to the student. Indeed, our teaching curriculum first infers the preferred strategy of the student from their proficiency level, then we allocate more training effort to the sub-skills they show strong improvement on. For example, as demonstrated in Figure 6(a), after the first 6 trials that estimated the student's proficiency for each sub-skill, the teacher found out this student improved more as the leader, therefore, the teacher allocated 12 trials to perfect the *leading* sub-skills and only 2 trials for *following*. Moreover, one participant in the random teacher group responded "the robot leading mode is too difficult and I gave up". This demonstrates the importance of an individualized curriculum: though there are multiple equally optimal strategies, the individual may have strong preferences, and teaching a non-preferable strategy will discourage the student from learning anything at all. We view this as a strong call for an individualized curriculum for cooperative robot teaching given human's variance in physical skills. We refer the readers to the Appendix for the complete data of all participants.

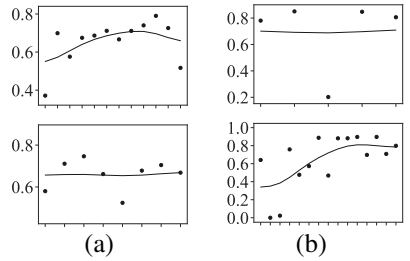


Figure 6. Sub-skill performances with respect to training progress of two example participants trained by the student-aware teacher. The top and bottom figures correspond to leading and following sub-skills respectively. (a) Participant 4. The student improved more when trained in the leading sub-skill. (b) Participant 6. The student improved more when trained in the following sub-skill.

6 Limitation

Decomposition into sub-skills. For many tasks, it is not easy to identify distinct roles to fulfill the local-independence criteria of sub-skills. We manually decompose the skill into a few sub-skills according to the role of the student. Often, such a decomposition may not be possible or requires careful design. **Curriculum design.** In this work, we only design the curriculum over different sub-skills. However, during our experiment, we observe that humans show various responses to the same sub-skill of different difficulties. As a result, a finer-grained curriculum on the sub-skill training shall be found to further facilitate human learning. **Robot teaching in other tasks.** We restrict the teaching to a cooperative task only in this work. Our formulation and approach cannot be applied to the single-agent game or competitive game directly.

7 Conclusion

In this work, we propose a conceptual framework, Cooperative Robot Teaching, that enables robots to teach humans in cooperative tasks. We show that, by abstracting a teaching task over the original duo cooperative task, the robot can learn to act as a specialized teacher to humans. To be more specific, we model the teaching task as a POMDP with hidden student policy and propose a partially assistive teaching curriculum to support human learning. We believe that robot teaching fills in the gap of the bilateral knowledge transfer in HRI: unlike other HRI tasks where the humans instruct the robots how to behave, now the role is reversed and robots try to instill the knowledge back into humans. Despite the great challenges that lie ahead, we believe that robot teaching has great potential and is a necessary step forward to bring robots closer to our daily life.

References

[1] E. García and E. Weiss. The teacher shortage is real, large and growing, and worse than we thought. the first report in" the perfect storm in the teacher labor market" series. *Economic Policy Institute*, 2019.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.

[3] A. Coates, P. Abbeel, and A. Y. Ng. Apprenticeship learning for helicopter control. *Commun. ACM*, 52(7):97–105, jul 2009.

[4] F. Suárez-Ruiz, X. Zhou, and Q.-C. Pham. Can robots assemble an ikea chair? *Science Robotics*, 3(17), 2018.

[5] S. Nikolaidis and J. Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 33–40, 2013.

[6] M. Chen, H. Soh, D. Hsu, S. Nikolaidis, and S. Srinivasa. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction*, 9(2):1–23, 2020.

[7] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

[9] T. Schodde, K. Bergmann, and S. Kopp. Adaptive robot language tutoring based on bayesian knowledge tracing and predictive decision-making. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 128–136, New York, NY, USA, 2017. Association for Computing Machinery.

[10] R. van den Berghe, J. Verhagen, O. Oudgenoeg-Paz, S. van der Ven, and P. Leseman. Social robots for language learning: A review. *Review of Educational Research*, 89(2):259–295, 2019.

[11] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching by pomdp planning. In *Artificial Intelligence in Education*, pages 280–287, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[12] A. D. Dragan and S. S. Srinivasa. A policy-blending formalism for shared control. In *International Journal of Robotics Research*, volume 32, pages 790–805, 2013.

[13] S. Reddy, A. D. Dragan, and S. Levine. Shared autonomy via deep reinforcement learning. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018.

[14] L. Jones. *The Student-centered Classroom*. Cambridge University Press, 2007.

[15] S. Nikolaidis and J. Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. *ACM/IEEE International Conference on Human-Robot Interaction*, pages 33–40, 2013.

[16] B. Yang, G. Habibi, P. Lancaster, B. Boots, and J. Smith. Motivating physical activity via competitive human-robot interaction. In *5th Annual Conference on Robot Learning*, 2021.

[17] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871, 2018.

[18] H. J. Jeon, S. Milli, and A. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

[19] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] A. Fern, S. Natarajan, K. Judah, and P. Tadepalli. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research*, 50:71–104, 2014.

[21] R. Shah, P. Freire, N. Alex, R. Freedman, D. Krasheninnikov, L. Chan, M. D. Dennis, P. Abbeel, A. Dragan, and S. Russell. Benefits of assistance over reward learning, 2021.

[22] O. Macindoe, L. Pack Kaelbling, and T. Lozano-Pérez. Pomcop: Belief space planning for sidekicks in cooperative games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 8(1):38–43, Jun. 2021.

[23] B. J. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[24] S. Nikolaidis, A. Kuznetsov, D. Hsu, and S. Srinivasa. Formalizing human-robot mutual adaptation via a bounded memory based model. In *Proceedings of 11th ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*, pages 75 – 82, March 2016.

[25] S. Nikolaidis, D. Hsu, and S. Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *International Journal of Robotics Research*, 36(5-7):618–634, 2017.

[26] S. Nikolaidis, Y. X. Zhu, D. Hsu, and S. Srinivasa. Human-robot mutual adaptation in shared autonomy. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 294–302, New York, NY, USA, 2017. Association for Computing Machinery.

[27] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[28] X. Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Mar. 2015.

[29] J. Liu, X. Zhu, and H. Ohannessian. The teaching dimension of linear learners. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 117–126, New York, New York, USA, 20–22 Jun 2016. PMLR.

[30] S. Mei and X. Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2871–2877. AAAI Press, 2015.

[31] F. Khan, B. Mutlu, and J. Zhu. How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[32] W. Liu, B. Dai, A. Humayun, C. Tay, C. Yu, L. B. Smith, J. M. Rehg, and L. Song. Iterative machine teaching. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2149–2158. PMLR, 06–11 Aug 2017.

[33] I. Gur, N. Jaques, Y. Miao, J. Choi, M. Tiwari, H. Lee, and A. Faust. Environment Generation for Zero-Shot Compositional Reinforcement Learning. (NeurIPS), 2022.

[34] R. Portelas, C. Colas, K. Hofmann, and P.-Y. Oudeyer. Teacher algorithms for curriculum learning of Deep RL in continuously parameterized environments. (CoRL), 2019.

[35] M. Fontaine*, Y.-C. Hsu*, Y. Zhang*, B. Tjanaka, and S. Nikolaidis. On the Importance of Environments in Human-Robot Coordination. *Robotics: Science and Systems*, 2021.

[36] D. S. Brown and S. Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, (2014):7749–7758, 2019.

[37] B. P. Gerkey and M. J. Matarić. A formal analysis and taxonomy of task allocation in multi-robot systems. *The International Journal of Robotics Research*, 23(9):939–954, 2004.

[38] J. P. González-Brenes and J. Mostow. What and when do students learn? fully data-driven joint estimation of cognitive and student models. In *EDM*, 2013.

[39] S. Omidshafiei, D. K. Kim, M. Liu, G. Tesauro, M. Riemer, C. Amato, M. Campbell, and J. P. How. Learning to teach in cooperative multiagent reinforcement learning. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 6128–6136, 2019.

[40] D. K. Kim, M. Liu, S. Omidshafiei, S. Lopez-Cot, M. Riemer, G. Habibi, G. Tesauro, S. Mourad, M. Campbell, and J. P. How. Learning hierarchical teaching policies for cooperative agents. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2020-May(Aamas):620–628, 2020.

[41] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. A. Ortega, D. J. Strouse, J. Z. Leibo, and N. de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:5372–5381, 2019.

[42] A. Singla, I. Bogunovic, G. Bartók, A. Karbasi, and A. Krause. Near-optimally teaching the crowd to classify. *31st International Conference on Machine Learning, ICML 2014*, 2:1355–1378, 2014.

[43] S. Zilles, S. Lange, R. Holte, and M. Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12:349–384, 2011.

[44] T. Doliwa, H. U. Simon, and S. Zilles. Recursive teaching dimension, learning complexity, and maximum classes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6331 LNAI:209–223, 2010.

[45] R. Hambleton and H. Swaminathan. *Item Response Theory: Principles and Applications*. Evaluation in education and human services. Springer Netherlands, 2013.

[46] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modelling the acquisition of procedural knowledge. volume 4, pages 253–278, 1995.

[47] A. Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in computational intelligence. Springer, Berlin, 2012.

[48] K. Shiarlis, M. Wulfmeier, S. Salter, S. Whiteson, and I. Posner. TACO: Learning task decomposition via temporal alignment for control. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4654–4663. PMLR, 10–15 Jul 2018.

[49] T. Kipf, Y. Li, H. Dai, V. Zambaldi, A. Sanchez-Gonzalez, E. Grefenstette, P. Kohli, and P. Battaglia. Compile: Compositional imitation learning and execution. In *International Conference on Machine Learning (ICML)*, 2019.

[50] P. Evrard and A. Kheddar. Homotopy switching model for dyad haptic interaction in physical collaborative tasks. In *World Haptics 2009 - Third Joint EuroHaptics conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pages 45–50, 2009.

[51] A. Kheddar. Human-robot haptic joint actions is an equal control-sharing approach possible? In *2011 4th International Conference on Human System Interactions, HSI 2011*, pages 268–273, 2011.

[52] N. Jarrassé, T. Charalambous, and E. Burdet. A framework to describe, analyze and generate interactive motor behaviors. *PLOS ONE*, 7(11):1–13, 11 2012.

[53] A. Mörtl, M. Lawitzky, A. Kucukyilmaz, M. Sezgin, C. Basdogan, and S. Hirche. The role of roles: Physical cooperation between humans and robots. *The International Journal of Robotics Research*, 31(13):1656–1674, 2012.

[54] C. Ekanadham and Y. Karklin. T-skirt: Online estimation of student proficiency in an adaptive learning system. *Machine Learning for Education Workshop at ICML*, 2017.

[55] K. H. Wilson, Y. Karklin, B. Han, and C. Ekanadham. Back to the basics: Bayesian extensions of irt outperform neural networks for proficiency estimation. In *EDM*, 2016.

[56] J. Gonzalez-Brenes, Y. Huang, and P. Brusilovsky. General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In *The 7th International Conference on Educational Data Mining*, pages 84 – 91, 2014.

[57] A. Lupu, B. Cui, H. Hu, and J. Foerster. Trajectory diversity for zero-shot coordination. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7204–7213. PMLR, 18–24 Jul 2021.

[58] R. Zhao, J. Song, H. Haifeng, Y. Gao, Y. Wu, Z. Sun, and Y. Wei. Maximum Entropy Population Based Training for Zero-Shot Human-AI Coordination. (NeurIPS):1–18, 2021.

[59] M. Carroll, R. Shah, M. K. Ho, T. L. Griffiths, S. A. Seshia, P. Abbeel, and A. D. Dragan. On the utility of learning about humans for human-ai coordination. In *NeurIPS*, 2019.

[60] P. Knott, M. Carroll, S. Devlin, K. Ciosek, K. Hofmann, A. D. Dragan, and R. Shah. Evaluating the robustness of collaborative agents. In *AAMAS*, pages 1560–1562, 2021.

[61] R. Charakorn, P. Manoonpong, and N. Dilokthanakul. Investigating partner diversification methods in cooperative multi-agent deep reinforcement learning. In *ICONIP*, 2020.

[62] P. Nalepka, J. Gregory-Dunsmore, J. Simpson, G. Patil, and M. Richardson. Interaction flexibility in artificial agents teaming with humans. In *CogSci 2021: program for the 43rd Annual Meeting of the Cognitive Science Society*, pages 112–118. Cognitive Science Society, 2021. Annual Meeting of the Cognitive Science Society (43rd : 2021), CogSci 2021 ; Conference date: 26-07-2021 Through 29-07-2021.

[63] B. Sarkar, A. Talati, A. Shih, and S. Dorsa. Pantheonrl: A marl library for dynamic training interactions. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (Demo Track)*, 2022.