Using Machine Teaching to Investigate Human Assumptions when Teaching Reinforcement Learners

Yun-Shiuan Chuang (yunshiuan.chuang@wisc.edu) [1]
Department of Psychology, University of Wisconsin-Madison, 1202 W. Johnson Street
Madison, WI 53706 USA

Xuezhou Zhang (xzhang784@wisc.edu) [1]
Department of Computer Science, University of Wisconsin-Madison, 1210 W. Dayton Street
Madison, WI 53706 USA

Yuzhe Ma (ma234@wisc.edu)
Department of Computer Science, University of Wisconsin-Madison, 1210 W. Dayton Street
Madison, WI 53706 USA

Mark K. Ho (mho@princeton.edu)
Department of Computer Science and Psychology, Princeton University, 35 Olden Street
Princeton, NJ 08540 USA

Joseph L. Austerweil (austerweil@wisc.edu) [2]
Department of Psychology, University of Wisconsin-Madison, 1202 W. Johnson Street
Madison, WI 53706 USA

Xiaojin Zhu (jerryzhu@cs.wisc.edu) [2]
Department of Computer Science, University of Wisconsin-Madison, 1210 W. Dayton Street
Madison, WI 53706 USA

Abstract

Successful teaching requires an assumption of how the learner learns - how the learner uses experiences from the world to update their internal states. We investigate what expectations people have about a learner with a behavioral experiment: Human teachers were asked to teach a sequential decision-making task to an artificial dog in an online manner using rewards and punishments. The artificial dogs were implemented with either an Action Signaling agent or a Q-learner with different discount factors. Our findings are threefold: First, we used machine teaching to prove that the optimal teaching complexity across all the learners is the same, and thus the differences in human performance was solely due to the discrepancy between human teacher's theory of mind and the actual student model. Second, we found that Q-learners with small discount factors were easier to teach than action signaling agents, challenging the established conclusion from prior work. Third, we showed that the efficiency of teaching was monotonically increasing as the discount factors decreased, suggesting that humans' theory of mind bias towards myopic learners.

Keywords: theory of mind; machine teaching; reinforcement learning

Introduction

People regularly teach other agents (e.g., children, pets, machines) in their environment using evaluative feedback (rewards and punishments). For example, Andrew is teaching his six-year-old daughter Jane to forage for wild berries. To do so, he first brings her to a bush with edible berries. What does he do next? If his goal is purely for her to eat some wild berries, then he has achieved his goal. However, to teach her to forage robustly, he must provide her rewards to incentivize her to leave that bush and seek new ones. How do people teach agents using rewards and punishments, and how does their teaching depend on their knowledge of the internal dynamics of how the learner updates their beliefs?

Although not always presented from this perspective, foraging is an example of the exploration-exploitation problem within reinforcement learning: How should an agent balance exploiting rewards based on their current knowledge while still exploring for berries? This problem has been extensively studies across humans, animals, and idealized agents (Cohen, McClure, & Yu, 2007; Gopnik, 2020; Reid et al., 2016; Stephens, Brown, & Ydenberg, 2007). Many natural agents stop exploiting rewards before the resource is exhausted so that they can explore other states. To do so, there must be some mechanism that provides an implicit or explicit punishment to the agent. Although many natural agents forage efficiently in their environmental niche, the mechanisms used to tune their behavior often are not robust to environments outside of their niche. Although foraging itself (and learning while foraging) has been extensively studied within reinforcement learning and other mathematical frameworks, to the best of our knowledge, teaching others to forage is an open question. In this paper, we explore this question for teaching agents a full policy in a task that requires the teacher to first

[1] Equal Contribution

[2] Co-Senior Author

incentivize sub-optimal actions so that the learner learns what they should do in states they otherwise would not encounter (because they start close to their ultimate goal).

Teaching the correct actions to take in a domain while exploring the domain is a social task involving the interaction of the teacher, a learner, and the environment. Researchers have examined how people teach others, formalized this process, and created automated methods to teach. One unifying computational framework across these areas is the Bayesian pedagogy framework, where the learner and teacher are Bayesian agents that assume both know the teacher is providing information to help the learner (Shafto, Goodman, & Griffiths, 2014). However, this framework is mechanism-agnostic and assumes humans are doing ideal Bayesian updates, which can be a problematic assumption. Instead, we take the perspective of machine teaching, where we examine human teaching from the perspective of optimized teaching for a *particular learning mechanism*.

In this paper, we focus on Q-learning, a family of model-free reinforcement learning algorithms that learns the optimal policy as the agent interacts with a Markov Decision Process (MDP; Sutton & Barto, 2018). We designed a study where a human teacher intervenes in the agent-MDP loop by selecting the reward signal at each time step. They were incentivized to teach the optimal policy as fast as possible. We tested human performance at teaching multiple Q-learning agents (students) with different parameters. If human teachers could teach some agents better than others, it would provide support that those agents' parameters are closer to human teachers' assumptions about how students learn.

Prior Work

From children to adults, when asked to teach, people provide different information than if they are simply asked to convey some information to another learner. In pedagogical situations, the MDP framework captures human performance fairly well. For example, when asked to show how to do a task, people will take actions that are strictly unnecessary for completing the task but convey information to a learner. However, if they are asked to do a task, they only do the necessary actions for completing the task (Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016).

Recent work in cognitive science and human-machine interaction has explored human teaching strategies and to what extent they are optimal. For example, work in Bayesian pedagogy has shown that when teaching a range of simple concepts by example, people can teach others near optimally (Shafto et al., 2014). Similar research on linguistic pragmatics demonstrates that people's language use reflects an intention to be optimally informative (Grice, 1975; Goodman & Frank, 2016). Moreover, these findings on human teaching have been shown to generalize to more complex settings, such as sequential decision making (Ho et al., 2016). However, others have found that in more complex settings, human teaching is misaligned with an MDP framework (Ho, Cushman, Littman, & Austerweil, 2019), which motivates research into learning algorithms tailored to human teaching strategies (Knox & Stone, 2009; MacGlashan et al., 2017).

Although prior work provides a useful perspective on how people ought to teach others who know they are being taught, it does so for an idealized Bayesian agent assumed to be maximizing environmentally provided rewards. Albeit a useful assumption for examining human teaching, people do not teach or learn as an idealized Bayesian agent. In the 1990s, researchers learned how difficult it can be to train a reinforcement learner to complete simple tasks that contained necessary conditions that need to be completed by the learner before the last steps of their task, which are closer in the state space to their current location (e.g., Ng, Harada, & Russell, 1999). Based on this intuition, recent work demonstrated that people fail to teach simple model-free and model-based reinforcement learners how to get from a start state to an end state while staying on a trail in a 3×3 Grid World (Ho et al., 2019). The learners often pick up on "positive net reward cycles", which enable them to get an arbitrarily large reward while not completing the task.

Computational Teaching: Since computational teaching was first proposed by Shinohara and Miyano (1991), optimal teaching has been studied for various learning mechanisms and settings (for a recent review, see Zhu, Singla, Zilles, & Rafferty, 2018). Of particular interest to us are works on teaching online learners such as Online Gradient Descent (OGD) (Liu et al., 2017; Lessard, Zhang, & Zhu, 2018), active learners (Hanneke, 2007; Peltola, Çelikok, Daee, & Kaski, 2019), and other sequential learners (Hunziker et al., 2019; Mansouri, Chen, Vartanian, Zhu, & Singla, 2019). For OGD, an optimal control formulation is required as the learning mechanism updates sequentially. Recent work studied teaching of reward functions using demonstrations (i.e., Inverse Reinforcement Learning; IRL) (Tschiatschek, Ghosh, Haug, Devidze, & Singla, 2019; Kamalaruban, Devidze, Cevher, & Singla, 2019). Finally, computational teaching for reinforcement learning has been studied recently (Zhang, Ma, Singla, & Zhu, 2020; Rakhsha, Radanovic, Devidze, Zhu, & Singla, 2020), where optimal teaching is solved for teachers using rewards and/or state transitions. The present work instead focuses on using computational teaching theory to understand how humans teach.

Computational Analyses of Sequential Decision Making

We study the interaction between three entities: the RL agent (dog learner), the teacher, and the underlying environment. In this work, we assume that the environment is a **reward-less** Markov Decision Process (MDP) parametrized by $M = (\mathcal{S}, \mathcal{A}, P, \mu_0)$ where S is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function where $\Delta_{\mathcal{S}}$ denotes the probability simplex over $\mathcal{S}$, and $\mu_0 \in \Delta_{\mathcal{S}}$ is the initial state distribution. The reward is provided instead by the teacher.

Learners

In our human experiment, we focused on two sets of learners (jointly denoted by $\mathcal{L}$), that have been widely studied as potential models of theory of mind for human learners: (1) standard Q-learning and (2) Action Signaling (AS). The two learners mainly differ in their internal knowledge representation and how they perform learning updates.

A Q-learning agent stores an estimate of the *Q table*, $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which approximates the future cumulative rewards that the agent can receive after performing an action $a \in \mathcal{A}$ in a state $s \in \mathcal{S}$ (Watkins, 1989). The learning update rule is defined by two parameters: the learning rate α and the discount factor γ. α determines how aggressive the learner updates its current belief given the new experience, and γ indicates how much the learner prefers future rewards compared to immediate rewards. Given an experience for a time step t, $e_t = (s_t, a_t, s_{t+1}, r_t)$, Q-learning updates the (s_t, a_t) entry of its Q table as

$$Q_{t+1}(s_t,a_t) = (1-\alpha)Q_t(s_t,a_t) + \alpha\left(r_t + \gamma \max_{a'} Q_t(s_{t+1},a')\right)$$

An Action Signaling (AS) agent stores a multinomial probability distribution over actions for each state, representing its current belief distribution of the optimal action. We included this type of agent because the action-signaling model was proposed to describe people's assumptions about a learning agent when they teach the agent (Ho et al., 2019). In this paper, we represent the multinomial distribution also with a table, $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, where $\sum_{a\in\mathcal{A}} Q(s,a) = 1$. The learning update of an AS agent is defined by a single parameter, the learning rate κ, which serves an analogous as α in Q-learning. Given a new piece of experience $e_t = (s_t, a_t, s_{t+1}, r_t)$, an AS agent updates the belief distribution over the current state s_t. If $r_t > 0$, the probability w.r.t. a_t will increase, whereas if $r_t < 0$, the probability w.r.t. a_t will decrease. In contrast to Q-learning, an AS agent does not make use of the next state s_{t+1}, and does not aim at optimizing the long term reward. Specifically,

$$Q_{t+1}(s_t,a) = \frac{Q_t(s_t,a)e^{1_{[a=a_t]}\kappa r_t}}{\sum_{b\in A} Q_t(s_t,b)e^{1_{[b=a_t]}\kappa r_t}}, \forall a \in A. \qquad (1)$$

We further assume that all learners behave according to the **ε-greedy policy** w.r.t. the current Q table, i.e.

$$a_t = \pi_t^{\varepsilon}(s_t) = \begin{cases} \arg\max_a Q_t(s_t,a), & \text{w.p. } 1-\varepsilon \\ \text{uniform from } A, & \text{w.p. } \varepsilon. \end{cases} \qquad (2)$$

The teacher's goal is to drive the learner to learn a target policy $\pi^\dagger$, which specifies the action $a^\dagger \in A$ the teacher wishes the learner to take at any state $s^\dagger \in S$. This example goal can be expressed as a target set of Q tables that satisfy $\mathcal{Q}^\dagger := \{Q : Q(s^\dagger,a^\dagger) > Q(s^\dagger,a), \forall a \neq a^\dagger\}$. The teaching succeeds if the learner's Q_t falls into $\mathcal{Q}\dagger$ at some time step t, in which case the teaching process terminates. The teachers were asked to achieve the teaching goal as quickly as possible.

The computational difficulty of teaching all learners in $\mathcal{L}$ is the same.

In this subsection, we show that all learners in $\mathcal{L}$ have the exact same teaching complexity, despite taking different forms and having different parameters. To demonstrate that, let us first show that the optimal teaching problem forms a higher-level teaching MDP $\mathcal{N} = (\Xi,\Delta,\rho,\tau)$: (1). The teacher observes the **teacher state** $\xi_t \in \Xi$, which jointly characterizes the environment and the learner at time t: $\xi_t := (s_t, a_t, s'_t, Q_t)$.; (2). The teacher's action space consists of all possible rewards $r_t \in \Delta := \mathbb{R}$; (3). The teacher receives a constant cost of $\rho_t = 1$ for every time step before the teaching goal is accomplished; (4). The teaching state transition probability is specified by $\tau(\xi_{t+1} \mid \xi_t, r_t)$. The resultant teacher state $\xi_{t+1} = (s_{t+1}, a_{t+1}, s'_{t+1}, Q_{t+1})$ is generated as follows: s_{t+1} is copied from s'_t in ξ_t; $a_{t+1} \sim \pi^{\varepsilon}_{t+1}(s_{t+1})$; $s'_{t+1} \sim P(\cdot|s_{t+1}, a_{t+1})$; Q_{t+1} is the learner's updated Q table.

The optimal teaching policy is one where the teacher minimizes its cumulative teaching action cost $\sum_{t=0}^{T} \rho_t$, s.t. $Q_T \in \mathcal{Q}$. Due to the randomness of the MDP as well as the learner's behavior policy, this quantity is a random variable, and thus we instead minimize its expected value. Formally, the teacher seeks a time-invariant **teaching policy** ϕ^*: $\phi^* = \arg\min_{\phi:\Xi\mapsto\Delta} \mathbb{E}_{\mathcal{N}}\left[\sum_{t=0}^{T} \rho_t \text{ , .s.t } Q_T \in \mathcal{Q}^\dagger\right]$. The shortest expected time step to achieve the teaching goal is the **teaching dimension**, i.e.

$$TD(M,L,Q_0,\pi^\dagger) = \min_{\phi:\Xi\mapsto\Delta} \mathbb{E}_{\mathcal{N}}\left[\sum_{t=0}^{T} \rho_t \text{ , s.t. } Q_T \in \mathcal{Q}^\dagger\right] \qquad (3)$$

Using this formalism, we show that all learners in $\mathcal{L}$ have the same teaching dimension for any MDP environment. Specifically,

Theorem 1. *For any two teaching instances defined by $(M,L,Q_0,\pi^\dagger)$ and $(M,L',Q'_0,\pi^\dagger)$, where $L,L' \in \mathcal{L}$, if Q_0 and Q'_0 satisfies that $Q_0(s,a) \geq Q_0(s,a')$ if and only if $Q'_0(s,a) \geq Q'_0(s,a')$ for all $s \in \mathcal{S}$, $a,a' \in \mathcal{A}$, then $TD(M,L,Q_0,\pi^\dagger) = TD(M,L',Q'_0,\pi^\dagger)$.*

Theorem 1 states that an optimal teacher can teach any learner in the family $\mathcal{L}$ equally fast, given that the learner starts with the same initial policy induced by Q_0 (see the Appendices for the proof). This finding suggests that if a teacher tries teaching each kind of dog in $\mathcal{L}$ and practice well enough, he/she can teach each dog with the same efficiency. Notably, the optimal teacher does not need to know the values of the student's internal parameters (e.g., the learning rate α, κ and the discount factor γ) or the student's learning formula (i.e., the update rule) (see the next section for how the optimal teaching policy was derived).

The Comparison of the Machine Teacher to Human Teachers

Unlike the theoretical optimal teacher described in the previous section, our human teachers do not know what learning algorithm or parameters the students use. This allows us to probe what people assume about students learning using evaluative feedback. We expect that when a student algorithm is closer to the human preconception of a student, the human teacher will be better at teaching that student.

How does human teaching compare to machine teaching in a task where a person needs to teach a reinforcement learner a full policy when the learner starts one action away from its environmental goal? Figure 1 presents an idealized exploration-exploitation scenario, where the dog's initial state is one tile left of its goal (the door of its home). The teaching goal is to teach the dog to go home from every tile. Whenever the dog reaches its home, it restarts learning on the initial state. To teach the dog to get home from every tile, the teacher must incentivize the dog to explore, which results in the dog moving away from its ultimate goal: It must punish moving to the goal and/or reward going away from the goal. Once the dog has gotten to the leftmost tile, the teacher can begin to "undo" their prior teaching and teach the dog to go right.

Calculating the optimal teaching policy. To derive the optimal teaching policy, we used the machine teaching method described in the Appendices. The learner starts indifferently with $Q_0(s,a) = 0$ for every feasible state-action pair (s,a). As we have shown in Theorem 1, the optimal teaching dimension is the same for all learners in $\mathcal{L}$. For concreteness, we assume that the learner is parameterized as follows: $\alpha = 0.9$, $\gamma = 0.9$, and $\varepsilon = 0.1$. Given this configuration, the optimal teaching policy is precisely the one discussed above (get the dog to move all the way left with minimal reward/punishment and then teach it to move right at every state). The number of steps to teach is a random variable whose outcomes depend on what actions are selected when the dog is indifferent and when suboptimal actions are taken. Thus, we approximated the expected number of steps via Monte Carlo (simulating teaching dogs using the optimal teaching policy and recording the number of steps it took for the full target policy to be taught). **On average, the optimal teaching policy takes 11 steps to teach the dog the full target policy, with** 100% **success rate.** Recall that by Theorem 1, it is 11 regardless of the learner type or the value of α and γ. Changing ε affects the optimal teaching policy and expected number of steps quantitatively, but the optimal teaching policy follows the same qualitative procedure as before. In light of this, we set $\varepsilon = 0.1$ as a constant throughout the study.

Although our machine teaching results suggest all learners in $\mathcal{L}$ should be equally efficient to train in this task, is this the case for human teachers? Are there some learners that are easier for people to train? Previously, Ho et al. (2019) found that traditional parametrization of Q-learners was extremely challenging for people to train even on simple tasks and that the action-signaling model was much easier for people to train. However, they only examined one learner configuration. In this study, we investigate the role of learner parameters for a different simple teaching task. In particular, we show that the discount factor greatly impacts human teaching success, and a Q-learner with a small discount factor outperforms an action-signaling learner (opposite of their previous work).

Remark (why not change α and κ): Note that in our experiment, we only varied the discount factor γ of the Q-learners, but fixed the learning rate α (for Q-learning) and κ (for AS). This is because we can show formally that the learning rate parameter does not affect teaching performance. In particular,

Proposition 2. *Let L_{AS}, L'_{AS} be two AS-learners with learning rate κ, κ' respectively and assume that $Q_0 = 0$. Then, given any sequence of experiences (s_t, a_t, r_t, s'_t), L_{AS} and L'_{AS} will learn the exact same policy, i.e. $\arg\max_a Q_t(s,a) = \arg\max_a Q'_t(s,a)$ for all s.*

The similar behavior holds for Q-learning. As a result, the learning rate does not provide meaningful variations to the experiments and was therefore fixed to $\alpha = 0.9$ for Q-learners and $\kappa = 1$ for AS-learners.

Human Experiment

Participants. We recruited 330 participants through Amazon MTurk. The number of participants was chosen *a priori* based on the authors' intuition from similar previously conducted studies. We excluded 15 participants (5 for not finishing the experiment, 6 due to experiment error, and 4 who selected "do nothing" for more than 90% of the steps). We analyzed the 915 training sessions from the remaining 305 participants.

Interface/Stimuli. The dog training task took place in a 4 × 1 MDP with an absorbing state on the right. Figure 1 shows the visual interface that the participants interacted with. Four states were represented by four tiles that the dog could walk on, and the absorbing state was represented by a door. At each step, the dog could only go right or left one step. If the dog went left at the leftmost tile, the dog would stay at the same tile. If the dog went right at the rightmost tile, the dog would enter its home, and learning restarted with the dog placed back at the rightmost tile. A dog training session ended either if the dog learned the full target policy (successful training), or the dog had already taken 40 steps (unsuccessful training). Once the training of one dog ended, participants continued to train a new dog (a dog in a different color). The internal states were displayed to participants as two rows provided by a "brain scanner", which corresponded to the dog's current Q table Q_t and the current policy for non-greedy actions (i.e., $\arg\max_a Q_t(s_t,a)$, $\forall a \in A$). At the beginning of each training session, the dog was placed at the rightmost tile with an initial Q table Q_0 where the Q value of each state-action pair was zero for Q-learners and 0.5 for the AS-learner. Participants responded via a continuous slider (feedback of -1 to 1), or could select a button to "do nothing" (feedback of zero).

Procedure. Participants taught three dogs to go home from any tile. They were given an extensive quiz about the instructions and could not continue until every question was answered correctly. There was one between-subjects condition: *learner type*. Each participant was assigned to either one of the four traditional Q-learners ($\gamma \in \{0.0, 0.1, 0.45, 0.9\}$) or

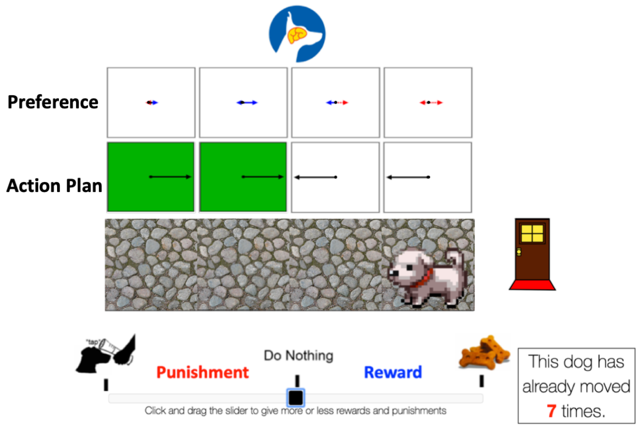
Preference
Action Plan
Punishment
Do Nothing
Reward
Click and drag the slider to give more or less rewards and punishments
This dog has already moved 7 times.

Figure 1: The dog training task took place in an environment comprised of four tiles and a door on the right. The dog's current Q table (labeled "Preference") and policy (labeled "Action Plan") were shown above the garden. In the Q table, four cells corresponded to the Q values of each of the four tiles. In each cell, arrows pointing to the left/right encoded the Q value of moving to the left/right at that tile. Blue solid arrows encoded positive Q values, whereas red dotted arrows encoded negative Q values. The policy was derived from the current Q table, where the arrow pointed in the direction with maximal Q value for the given state. When the dog's policy at a tile dictated going towards the door, the background of the cell turned green, indicating the policy at that state matched the target policy. After the dog took its action, the participant was asked to provide feedback to the dog. The participant could use the slider to select the feedback value or clicked the "do nothing" button to give zero feedback. After the participant chose the feedback, the dog's Q table and target would be updated, and the dog performed its next action. The number of steps was displayed in the bottom-right corner.

the AS-learner.

The learning dynamics were shown to the participant in the form that the dog's internal state was displayed to the participant while the participant was sliding to decide the feedback value. The reason that we showed the internal states and the learning dynamics was to ensure the human teachers could access the as much information as the optimal teacher did. On each trial, the dog would move from one state to another. It followed an ε-greedy policy ($\varepsilon = 0.1$) with respect to the current Q-table. When a random action was selected for the dog, a squirrel would appear in the direction of that action (participants were told that when a squirrel appeared, the dog would move towards it no matter what its internal states were). After the dog moved, the participant would respond by dragging the feedback slider or hitting the "do nothing" button. After finishing training three dogs, they took a short survey to ensure that they treated the slider symmetrically and to gathered standard demographic data.

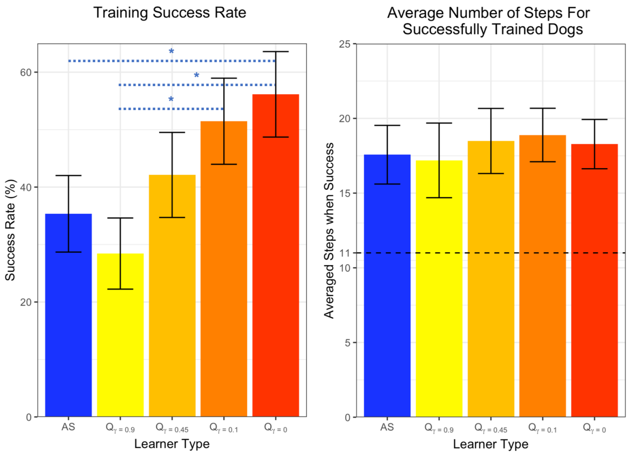

Figure 2: Participant success rate (right panel) and teaching length at training the full policy(left panel) for different learner types. The blue dotted horizontal lines above the bars (in the left panel) indicate significant differences between the bars, tested by Tukey post-hoc pairwise comparisons (see the main text for details). The dotted horizontal line (in the right panel) denotes the teaching length based on the optimal teaching policy. Error bars denote 95% confidence intervals. AS: Action Signaling; $Q_{\gamma=x}$: Q-learner with $\gamma = x$.

Results and Discussion

Figure 2 shows the success rate of participants at training the full policy and the average number of steps taken when a dog is successfully trained. To analyze the success rate, we fitted a mixed-effects logistic regression model where the binary dependent variable was whether the training session was successful, with learner type being the fixed effect, and participant being the random effect. The learner type significantly influenced the training success rate, $\chi^2(4) = 24.35, p < .001$. Follow-up pairwise comparisons were performed with Tukey adjustment. Critically, the training success rate was significantly larger for $Q_{\gamma=0}$ than the success rate for AS, $z = 3.25, p = .010$. The training success rate for $Q_{\gamma=0}$ was also significantly larger than the training success rate for $Q_{\gamma=0.9}$, $z = 4.28, p < .001$, so did $Q_{\gamma=0.1}$ compared to $Q_{\gamma=0.9}$, $z = 3.61, p < .005$. All other pairwise comparisons were not significantly different.

On the other hand, we also fitted a mixed-effects linear regression model to analyze the number of steps taken when successful (the right panel of Figure 2), with learner type being the fixed effect, and participant being the random effect. The learner type had no significant effect on the number of steps when successful, $F(4,171) = 0.43, p = .786$.

Are humans good teachers? How well did people teach the dogs? Clearly, they were suboptimal – the success rate ranged from 28.43% to 56.14% (AS: 35.35%, $Q_{\gamma=0.9}$: 28.43%, $Q_{\gamma=0.45}$: 42.11%, $Q_{\gamma=0.1}$: 51.46%, $Q_{\gamma=0}$: 56.14%), and when successful it took 17.19 to 18.89 steps (AS: 17.57, $Q_{\gamma=0.9}$:

17.19, $Q_{\gamma=0.45}$: 18.49, $Q_{\gamma=0.1}$: 18.89, $Q_{\gamma=0}$: 18.28). In contrast, the optimal teaching policy had a success rate of 100% and took an average of 11 steps regardless of the learner type. In sum, we observed that human teachers were suboptimal: they did not always succeed in teaching the policy to an agent; and when they did, they required more steps compared to the 11 steps of the optimal teaching algorithm.

Which student aligns best with humans' theory of mind? Our experimental results suggest that the human teachers expect students to be myopic (i.e. $\gamma = 0$, which is its smallest possible value). One interpretation is that the teacher are better at teaching students that are *predictable* (without any complications arising from discounted future rewards when $\gamma > 0$). Critically, our results were inconsistent with the results of Ho et al. (2019), showing that teaching a Q-learning agent with a small discount factor was easier than teaching an action-signaling agent. One limitation of our results is that the task, the MDP, and the population were different between the present study and Ho et al. (2019)'s study, and future work is required to clarify the generalizability of our findings.

Why is $Q_{\gamma=0}$ easier to teach than an AS-learner? At the first glance, the AS-learner is similar to a Q-learner with $\gamma = 0$ in that they both don't take future expected award into account. The AS-learner, however, has a critical difference from $Q_{\gamma=0}$. Note that according to the AS's update rule (Equation 1), when $Q(s,a)$ approaches 1 for any (s,a), the effect of the feedback r on the Q value of other actions a' at the same state s diminishes. In other words, once the AS-learner has developed a strong belief for a specific action a at a state s, it becomes harder for the learner to change its belief if the teacher decides to teach a new action. This indeed poses a challenge when the teacher wants to "undo" what the learner has learned before, e.g., once the dog reaches the leftmost tile, or when a wrong policy was learned. In contrast, the $Q_{\gamma=0}$ learner does not suffer from this complication.

Conclusions

In this paper, we investigated how people taught learners to solve the exploration-exploitation tradeoff, which is a common problem in everyday life. We did so by first formulating it as a machine teaching problem. Our first contribution was a theoretical result showing that, from the point of view of the computational difficulty, both the Q-learner and the AS-learner considered were equally hard to teach regardless of their internal parameters (e.g., learning rate,). We then ran a behavioral experiment of an idealized scenario to see how well people trained artificial agents to solve the exploration-exploitation tradeoff when the teaching goal was an entire policy. We found that people were suboptimal and that teaching was the easiest with a Q-learner that had a small discount factor γ. Future work is required to clarify the generalizability of our findings.

Acknowledgments

This work was funded by the VCRGE at University of Wisconsin - Madison with funding from the WARF.

References

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942.

Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*(1803), 20190502.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics (vol. 3* (pp. 41–58). Academic Press.

Hanneke, S. (2007). Teaching dimension and the complexity of active learning. In *International conference on computational learning theory* (pp. 66–81).

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teaching with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520–549.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In *Advances in neural information processing systems* (pp. 3027–3035).

Hunziker, A., Chen, Y., Mac Aodha, O., Rodriguez, M. G., Krause, A., Perona, P., ... Singla, A. (2019). Teaching multiple concepts to a forgetful learner. In *Advances in neural information processing systems.*

Kamalaruban, P., Devidze, R., Cevher, V., & Singla, A. (2019). Interactive teaching algorithms for inverse reinforcement learning. In *Ijcai* (pp. 2692–2700).

Knox, W. B., & Stone, P. (2009). Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on knowledge capture* (pp. 9–16).

Lessard, L., Zhang, X., & Zhu, X. (2018). An optimal control approach to sequential machine teaching. *arXiv preprint arXiv:1810.06175.*

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., ... Song, L. (2017). Iterative machine teaching. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 2149–2158).

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wuang, G., Roberts, D. L., ... Littman, M. L. (2017). Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th international conference on machine learning (vol. 70)* (pp. 2285–2294).

Mansouri, F., Chen, Y., Vartanian, A., Zhu, J., & Singla, A. (2019). Preference-based batch and sequential teaching:

Towards a unified view of models. In *Advances in neural information processing systems* (pp. 9195–9205).

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml* (Vol. 99, pp. 278–287).

Peltola, T., Çelikok, M. M., Daee, P., & Kaski, S. (2019). Machine teaching of active sequential learners. In *Advances in neural information processing systems* (pp. 11202–11213).

Rakhsha, A., Radanovic, G., Devidze, R., Zhu, X., & Singla, A. (2020). Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. *arXiv preprint arXiv:2003.12909.*

Reid, C. R., MacDonald, H., Mann, R. P., Marshall, J. A. R., Latty, T., & Gamier, S. (2016). Decision-making without a brain: how an amoeboid organism solves the two-armed bandit. *Journal of the Royal Society: Interface*, *13*, 20160030.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shinohara, A., & Miyano, S. (1991). Teachability in computational learning. *New Generation Computing*, *8*(4), 337–347.

Stephens, D. W., Brown, J. S., & Ydenberg, R. C. (2007). *Foraging: Behavior and ecology*. Chicago, USA: University of Chicago Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tschiatschek, S., Ghosh, A., Haug, L., Devidze, R., & Singla, A. (2019). Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *Advances in neural information processing systems.*

Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Zhang, X., Ma, Y., Singla, A., & Zhu, X. (2020). Adaptive reward-poisoning attacks against reinforcement learning. *arXiv preprint arXiv:2003.12613.*

Zhu, X., Singla, A., Zilles, S., & Rafferty, A. N. (2018). An overview of machine teaching. *CoRR*, *abs/1801.05927*.

Appendices

Proof of Theorem 1. First, we denote $Q_t \equiv Q'_t$ if $Q_t(s,a) \geq Q_t(s,a')$ if and only if $Q'_t(s,a) \geq Q'_t(s,a')$ for all $s \in \mathcal{S}$, $a, a' \in \mathcal{A}$. The proof is based on a key technical lemma:

Lemma 3. *For any teaching policy ϕ for teaching instance $(M, L, Q_0, \pi^\dagger)$, there exists a matching teaching policy ϕ' for teaching instance $(M, L', Q'_0, \pi^\dagger)$, such that for any time t, if $Q_t \equiv Q'_t$ and $\xi_t = \xi'_t$, then $Q_{t+1} \equiv Q'_{t+1}$.*

Lemma 3 implies that with the same random seed, for $L, L' \in \mathcal{L}$, the relationship $Q_{t+1} \equiv Q'_{t+1}$ will remain invariant through teaching when ϕ is used on $(M, L, Q_0, \pi^\dagger)$ and ϕ' used on $(M, L', Q'_0, \pi^\dagger)$. Thus, $Q_t \in Q^\dagger$ if and only if $Q'_t \in Q^\dagger$. Therefore, if there exists an optimal ϕ that achieves the teaching dimension for $(M, L, Q_0, \pi^\dagger)$, then the matching ϕ' also achieves the same teaching dimension for $(M, L', Q'_0, \pi^\dagger)$. Thus, the TD for both teaching instances must match. What remains is to prove Lemma 3. ■

Proof of Lemma 3. Given a particular Q_t and an experience tuple $\xi_t = (s_t, a_t, s'_t)$, the Q-learning will only modify the value of $Q_t(s_t, a_t)$ based on the teacher provided reward $r_t = \phi(\xi_t)$. Define the rank of (s_t, a_t) in Q_t as $rank_{Q_t}(s_t, a_t) = |\{a | Q_t(s_t, a) > Q_t(s_t, a_t)\}|$. On the other hand, while the AS update rule changes the $Q_t(s_t, a)$ for all actions, the rank among all the other actions $a \neq a_t$ remains unchanged because their values are only renormalized with a shared denominator. Therefore, under either the Q-learning update rule

$$Q_{t+1}(s_t, a_t) = (1-\alpha)Q_t(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q_t(s_{t+1}, a')),$$

or the AS-learner update rule

$$Q_{t+1}(s_t, a) = \frac{Q_t(s_t, a) e^{\mathbb{1}_{[a=a_t]}\kappa r_t}}{\sum_{b \in A} Q_t(s_t, b) e^{\mathbb{1}_{[b=a_t]}\kappa r_t}}, \ \forall a \in A.$$

it is obvious that there exist r_t such that the rank of (s_t, a_t) in Q_{t+1} can be any of $0, 1, ..., A-1$. Assume now that the rank of (s_t, a_t) in Q_{t+1} becomes k after updating with $r_t = \phi(\xi_t)$. Define ϕ' such that $r'_t = \phi'(\xi'_t)$ will also update the rank of (s_t, a_t) in Q'_{t+1} to be k. Then, since $Q_t \equiv Q'_t$ and for both Q_{t+1} and Q'_{t+1} the rank of (s_t, a_t) becomes k, we have $Q_{t+1} \equiv Q'_{t+1}$. This concludes the proof. ■

Deriving the Optimal Teaching Policy

To derive the optimal teaching policy, we used Twin Delayed DDPG (TD3) (Fujimoto, van Hoof, & Meger, 2018), a state-of-the-art Deep Reinforcement Learning (DRL) algorithm that solved continuous control problems. The hyperparameters for TD3 were described in Table 1. The hyperparameters were selected via grid search on the Dog MDP. Each experiment was run for 5000 episodes, where each episode is 200 iterations long. The learned policy was evaluated for every 50 episodes, and the policy with the best evaluation performance was used for the computation of the teaching dimension. In the computation of the teaching dimension, we ran the best-found TD3 policy for 1000 episodes, and took the average number of steps to teach the target policy. This gave 11.0.

Parameters	Values	Description
exploration noise	0.5	Std of Gaussian exploration noise.
batch size	100	Batch size for both actor and critic
discount factor	0.99	discount factor for the attacker problem.
policy noise	0.2	Noise added to target policy during critic update.
noise clip	$[-0.5, 0.5]$	Range to clip target policy noise.
action L2 weight	50	Weight for L2 regularization added to the actor network loss function.
buffer size	10^7	Replay buffer size, larger than total number of iterations.
optimizer	Adam	Use the Adam optimizer.
learning rate critic	10^{-3}	Learning rate for the critic network.
learning rate actor	5^{-4}	Learning rate for the actor network.
τ	0.002	Target network update rate.
policy frequency	2	Frequency of delayed policy update.

Table 1: Hyperparameters for TD3.