Teaching by Intervention: Working Backwards, Undoing Mistakes, or Correcting Mistakes?

Mark K Ho (mark_ho@brown.edu)
Department of Cognitive, Linguistic and Psychological Sciences, Brown University, 190 Thayer Street
Providence, RI 02912 USA

Michael L. Littman (mlittman@cs.brown.edu)
Department of Computer Science, Brown University, 115 Waterman Street
Providence, RI 02912 USA

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

Abstract

When teaching, people often intentionally intervene on a learner while it is acting. For instance, a dog owner might move the dog so it eats out of the right bowl, or a coach might intervene while a tennis player is practicing to teach a skill. How do people *teach by intervention*? And how do these strategies interact with learning mechanisms? Here, we examine one *global* and two *local* strategies: working backwards from the end-goal of a task (*backwards chaining*), placing a learner in a previous state when an incorrect action was taken (*undoing*), or placing a learner in the state they would be in if they had taken the correct action (*correcting*). Depending on how the learner interprets an intervention, different teaching strategies result in better learning. We also examine how people teach by intervention in an interactive experiment and find a bias for using local strategies like *undoing*.

Keywords: teaching, intervention, reinforcement learning

Introduction

When attempting to teach another agent, people have many tools at their disposal. They may choose to explain (Callanan & Oakes, 1992), give a demonstration (Brugger, Lariviere, Mumme, & Bushnell, 2007; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Király, Csibra, & Gergely, 2013), or offer rewards and punishments for taking certain actions (Knox & Stone, 2015; Ho, Littman, Cushman, & Austerweil, 2015). Another way in which people teach a learner is by *intervening* on the learner or the learner's environment. For example, if a puppy urinates on the carpet when a person is trying to teach the puppy to urinate on a pad, a person might move the puppy to the pad or move the pad to the puppy. When teaching another person a skill like tennis, a teacher might intervene on the trainee mid-movement and either adjust their arm to match the target movement or stop them to start over. The space of possible ways in which a teacher could change a learner's situation for pedagogical purposes is large. This raises several questions: First, what is the effectiveness of different intervention strategies? Second, how could learners interpret interventions and how does the interpretation affect a teaching strategy's efficacy? And, finally, what teaching strategies do people tend to use?

In this work, we examine three teaching by intervention strategies from a reinforcement learning perspective (Sutton & Barto, 1998). The first, *backward chaining*, is motivated by algorithms such as value iteration (Bellman, 1957) that solve multi-stage decision-problems by propagating information about rewards to previous states that lead to those rewards. Intuitively, this is akin to teaching a task by "working backwards", first ensuring that the learner knows how to reach a goal from the penultimate state, and then reach the penultimate state from the antepenultimate state, and so on. We consider this a *global* intervention strategy since it involves changing the learner's state in a manner that reflects the structure of the entire task, rather than a small part of it. The second strategy, *undoing*, is motivated by the intuition that interventions prevent learners from executing an undesirable action by having them restart from the state they performed the undesirable action. The third strategy, *correcting*, intervenes on a learner when she executes an undesirable action (like undoing), but places her in the state she would have gone to if she had taken the desired action. Unlike *backwards chaining*, *undoing* and *correcting* involve *local* changes to an agent's state.

How could a learner interpret an intervention? In a typical reinforcement learning setting, an agent takes an action in a state, and then the environment rewards or punishes her and moves her to a new state (Figure 1). We formalize four ways that an intervention can be interpreted. First, the intervention may simply *reset* the learner in a new location from which the next action will be taken. Second, the next state that the learner is moved to could be interpreted as part of a *transition* in the environment. Third, the intervention could be treated as an *interruption* in a learner's stream of behavior such that the undesirable action just taken never happened. Fourth, the intervention could be treated as a *disruption*, in which the intervention is experienced negatively. Each of these accounts may interact with a teacher's training strategy in different ways, meaning that the best teaching strategy may be dependent on the learner's intervention interpretation.

The outline of the paper is as follows. First, we review the reinforcement learning framework. Second, we formalize four different ways that a reinforcement learning algorithm

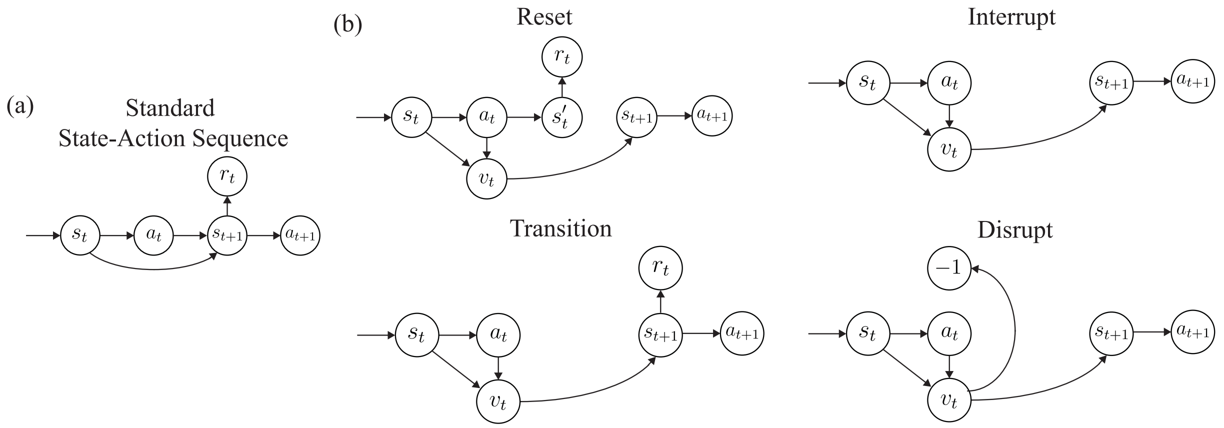


Figure 1: (a) Standard state, action, reward, next state sequence of a Markov Decision Process at a given time step. (b) Four different interpretations of a teacher intervening to place the learner in state v_t in response to a learner's action a_t from state s_t. When interventions are interpreted as *reset*, *transition*, or *disrupt*, r_t is respectively determined by the environmental next state, s'_t, the teacher's next state, s_{t+1}, or the teacher's intervention, v_t. When the the intervention is treated as *interrupt*, no reward experienced and no learning occurs for that time step.

could interpret an intervention and three teaching strategies. Third, we conduct simulations to examine how efficacious different teaching strategies are depending on how a learner interprets their interventions. Fourth, we conduct an experiment to investigate how *people* teach by intervention. We find that undoing, a local intervention strategy, is often effective and that people tend to teach most often by undoing, occasionally correcting, and rarely backward chaining.

Computational Modeling

In this section we present the standard reinforcement learning (RL) formalism, discuss the four intervention interpretations, and define the three teaching strategies.

Reinforcement Learning RL describes how an agent interacts with an environment and learns reward-maximizing behaviors (Sutton & Barto, 1998). Formally, an RL algorithm learns to take actions in a Markov Decision Process (MDP), defined by the tuple $< S, A, T, R, \gamma >$: a set of states in the world S; a set of actions for each state $\mathcal{A}(s)$; a transition function that maps state-action pairs to a probability distribution over next states, $P(s' \mid s, a)$; a reward function that maps states to scalar rewards, $R : S \to \mathbb{R}$; and a discount factor $\gamma \in (0, 1]$.

At each time step t, an RL agent takes an action a_t from a state s_t, which results in moving to next state s_{t+1} and a reward $r_{t+1} = R(s_{t+1})$ (Figure 1). Actions are determined by the agent's policy π that maps states to distributions over actions. For a policy π, the value at each state, $V^{\pi}(s)$, is:

$$V^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right]. \qquad (1)$$

The optimal policy, π^*, is one that maximizes the value function in every state, $V^*(s) = \max_{\pi} V^{\pi}(s), \forall s \in S$. An agent uses state, action, next state, reward tuples to learn an optimal policy.

Q-Learning One algorithm for learning an optimal policy is Q-learning, which is an off-policy temporal difference control algorithm. Under mild assumptions, Q-learning converges to the true action-value function (Watkins & Dayan, 1992). Moreover, humans and animals both engage in the type of error-driven reward learning found in Q-learning, making it a useful model with which to test different human teaching strategies (Niv, 2009). We use one form of this algorithm, one-step Q-learning, which is defined by the following update rule given a tuple (s, a, s', r):

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s,a)]. \qquad (2)$$

where α is the learning rate. We convert the estimated action-value function to a policy using the softmax decision-rule $\pi(a \mid s) = \exp\{Q(s,a)/\lambda_Q\} / \sum_{a'} \exp\{Q(s,a')/\lambda_Q\}$, where λ_Q is a temperature parameter controlling the probability that an agent takes the action estimated to yield the largest reward depending on the relative rewards she could get by taking other actions.

Teaching by Intervention

Interpreting Interventions The standard RL formulation does not define how interventions should be interpreted. Thus, we posit four different possible interpretations here, depicted in Figure 1. The four interpretations are motivated by formalizing the following two intuitions in different ways. First, a teacher could be treated as a part of the environment such that her intervention directly changes the next state of the learner (possibly stopping the feedback she would have received had she gone to the next state had the intervention not happened). Second, a teacher is distinct from the standard MDP environment, and intervenes as a direct response to a learner having taken an action and ended up in a next state.

Formally, at a time step t, the learner in state s_t takes an action a_t and ends up in new state s'_t. If the teacher does not intervene, $s_{t+1} = s'_t$. Otherwise, a teacher intervenes to place the learner in state $v_t \in S$. For all intervention types, $s_{t+1} = v_t$. However, if the teacher's intervention is interpreted as a *reset*, then the learner performs a Q-learning update using the tuple $(s_t, a_t, s'_t, R(s'_t))$, meaning that she still receives the reward she would have gotten had she reached s'_t as the next state. If it is interpreted as a *transition*, then the learner updates with $(s_t, a_t, v_t, R(v_t))$, meaning that she gets the reward had she taken the action that would move her from s_t to v_t. If it is an *interruption*, then the learner does not update the state-action value function the state-action pair that was intervened on, and she takes her next action in $s_{t+1} = v_t$. If it is interpreted as a *disruption*, then the learner updates with $(s_t, a_t, v_t, -1)$.

Teaching Strategies We discuss three teaching strategies: *backward chaining*, *undoing*, and *correcting*. A teacher using *backward chaining* has an n-length trajectory $J =< (s_1, a_1), ..., (s_n, a_n) >$ that she uses to teach the learner. We denote the states in the trajectory as $S_J = \{s_i : i = 1, 2, 3..., n\}$. The teacher also has a utility function over different interventions, where initially $U_0(s_i) = i$ for $i = 1, 2, 3, ..., n$ and $U_0(s) = -\infty$ for $s \in S \setminus S_J$. On each time step, the teacher's utility function is updated as:

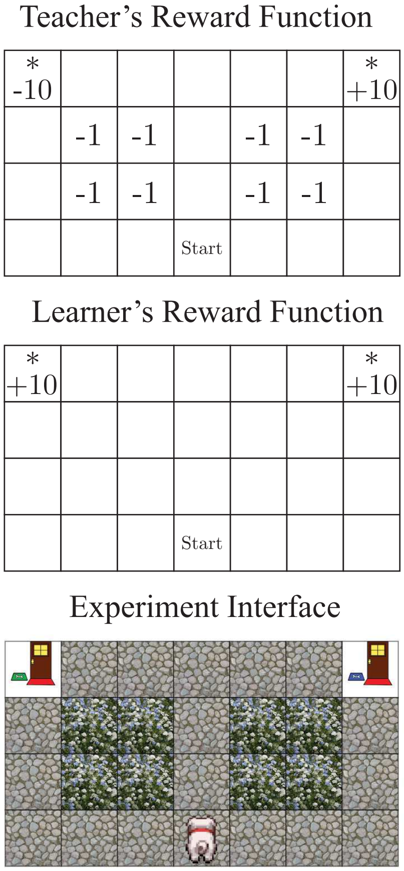
$$U_{t+1}(s_t) = \begin{cases} U_t(s_t) - 1 \text{ if } (s_t, a_t) \in J \\ U_t(s_t) \text{ otherwise.} \end{cases} \tag{3}$$

Teachers only intervene when the agent performs an action inconsistent with the trajectory (i.e. $(s_t, a_t) \notin J$) and place the agent in a next state according to a softmax decision rule over their utilities: $P(v) \propto \exp\{U_t(v)/\lambda\}$, where λ is a temperature parameter. The *backward chaining* teacher is initially more likely to move the agent closer to the end of a target trajectory, but as the agent shows they can perform the target action in a state the utility of moving the agent to that state decreases. Meanwhile, the relative utility of placing the agent in a slightly earlier stage in the trajectory increases.

A teacher using an *undoing* strategy has a target policy $\pi^* : S \to A$ that it is attempting to teach. On each time step, if an agent's action $a_t \neq \pi^*(s_t)$, then $v_t = s_t$. That is, when an agent takes an incorrect action, that action is undone by the teacher and the agent is placed back in the state she took the incorrect action. A teacher using a *correcting* strategy also has a target policy π^* that it is attempting to teach. However, if an agent's action $a_t \neq \pi^*(s_t)$, then $v_t = \arg\max_s T(s \mid s_t, \pi^*(s_t))$. That is, the teacher will move the agent to the state it would have been in had the agent taken the target action.

Simulations

To understand the interaction of teaching strategy and learner interpretation, we simulated the performance of a RL agent for each combination in a gridworld task.



Teacher's Reward Function

* -10						* +10
	-1	-1		-1	-1	
	-1	-1		-1	-1	
			Start			

Learner's Reward Function

* +10						* +10
			Start			

Experiment Interface

Figure 2: Task used for simulations and experiment. Asterisks (*) indicate absorbing states, both providing reward to the learner, whereas the teacher received reward if the learner entered the right door, but was punished if the learner entered the left door. The teacher received a mild punishment whenever the learner entered a garden tile.

Task

The task we used is shown in Figure 2. It consists of a 7 × 4 gridworld where the learning agent always starts a round in the center tile of the first row. At any given location, a subset of the four cardinal directions is available to the learning agent (e.g. at the bottom edge, "down" is not available as an action). On each episode, the learning agent starts in the bottom-middle tile and the upper-right and upper-left corners of the gridworld are absorbing states.

In our task, the teacher and learner have different rewards for the learner's actions in the MDP. In particular, the two absorbing states (goals) both have a +10 reward for the learner, but for the teacher, only one has +10 while the other has −10. Additionally, there are several non-absorbing tiles that give the teacher −1 if the learner enters them. These features of the task are visualized in Figure 2.

All simulations used a Q-learning agent with a tabular representation of states ($Q_0(s, a) = 0 \forall s, a$, $\alpha = .9$, and $\gamma = .95$). Each simulated teacher interacted with the learner for 12

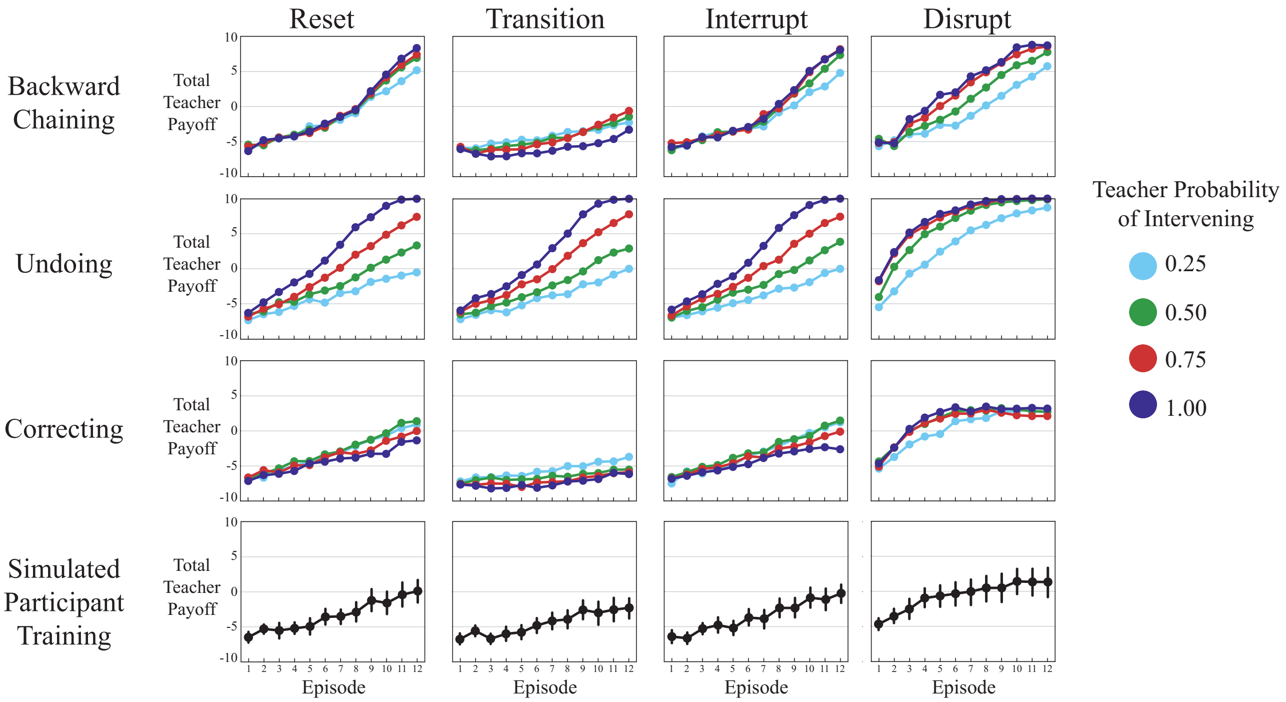


Figure 3: Simulated *backward chaining*, *undoing*, and *correcting* results or different intervention interpretations and intervention probabilities (top three rows). Results of learners trained using participant responses on task (bottom). Total teacher payoff is the net reward of the learner's behavior based on the teacher's reward function during the evaluation phase of each episode.

episodes. Each episode was divided into two phases: a teaching phase and an evaluation phase. During the teaching phase, the simulated teacher interacted with the Q-learner, which selected actions using a softmax rule ($\lambda_Q = .1$) and engaged in learning. During the evaluation phase, the learner performed the task *without* teacher interaction or learning and used a greedy policy. Additionally, the performance was measured with respect to the *teacher's* payoffs based on her reward function. Each episode phase ended after 25 time steps.

Teaching Strategies and Interpretations

We tested all combinations of teaching strategy and intervention interpretation ($\{$*backwards chaining*, *undoing*, *correcting*$\} \times \{$*resetting*, *transitioning*, *interrupting*, *disrupting*$\}$). In natural situations, it is not likely that teachers intervene every time a learner takes an incorrect action. Thus, we tested the performance of the models given different probabilities of intervening given that the learner performed an incorrect action: $0.25, 0.50, 0.75, 1.0$. This allowed us to evaluate the robustness of different teaching method and intervention interpretation combinations when feedback is imperfect. Each combination of teaching strategy, intervention type, and intervention probability were simulated 1000 times and teaching performance was based on the evaluation phase.

Results and Discussion

Simulation results are plotted in Figure 3. When interaction probability is high, *undoing* is most effective. This is because interventions act as impassable obstacles to the learning agent, which, combined with a discount rate, makes taking incorrect actions less beneficial than alternative actions that change the state and lead to reward. However, an exception is when the learner interprets interventions as *disrupting*, where the average performance of the *undoing* teaching strategy decreases quickly as intervention probability drops. This is because the teacher is less likely to serve as an obstacle, which makes it less likely that the agent will learn that incorrect actions are less efficacious. Across all interpretations, *undoing* outperforms *correcting* because *undoing* implicitly teaches the learner that the garden tiles are negative, whereas, *correcting* does not. *Undoing* also leads to more learning experience because *correcting* allows the agent to progress on the task without actually taking target actions.

When the probability of intervention is high ($1.0 - 0.75$), the *backward chaining* strategy performs as well as or worse than the *undoing* strategy. Unlike *undoing*, a global strategy like *backward chaining*'s efficacy is robust to less frequent interventions. This is because these interventions ensure that the learning agent has mastered a subset of states and acquired an accurate value representation as opposed to acting as a constraint on transitions in the environment.

The different intervention types also interacted with the teaching strategies in important ways. First, *undoing* shows identical patterns regardless of whether the intervention type is *resetting*, *transitioning*, or *interrupting*. When it is *disrupting*, learners reach maximum performance even more

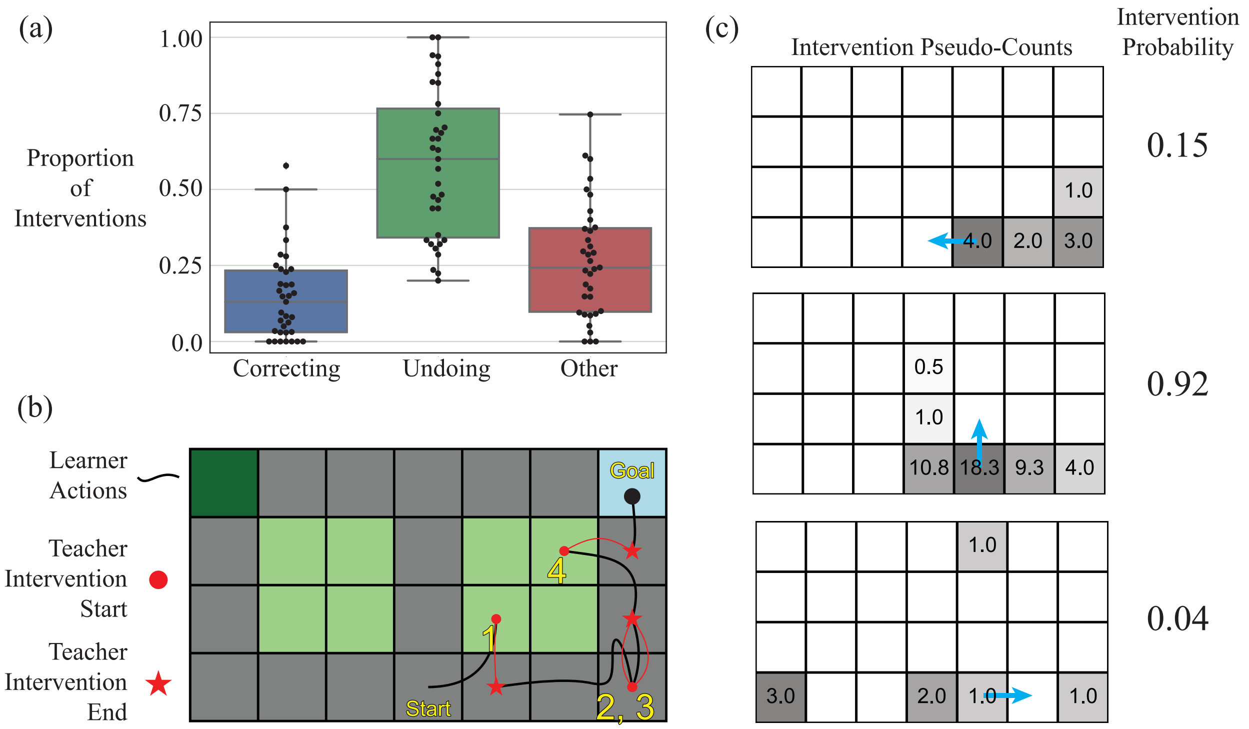


Figure 4: Experimental results. (a) Boxplot of proportion of correcting, undoing, and other interventions performed by individual participants. For many participants, the majority of their interventions were to undo the learner's action. (b) Graphical visualization of teacher-learner interaction during an episode ($\varepsilon = 0.8$) illustrating local interventions. Yellow numbers indicate order of interventions. (c) Graphical visualization of participant interventions for actions taken from the same state. For each episode, each participant has one pseudo-count that is divided among all of their interventions in that episode. The number in each tile represents the sum of these pseudo-counts over participants. The intervention probability is the proportion of times that action was subsequently intervened upon.

quickly. Second, for the *backward chaining* strategy, all strategies but *transitioning* led to learners acquiring policies that approached the target behavior. This is likely because the *transitioning* interpretation results in learners using the teacher's interventions as a way to "teleport" to a desirable location on the grid and not properly learn the task.

Experiment

How do people teach using interventions? Do they use a global strategy like *backwards chaining* or a local one like *undoing* or *correcting*? Our simulations suggest that *undoing* is the best teaching strategy if teachers intervene when the learner makes a mistake with high probability. However, *backwards chaining* works better when the teacher intervenes infrequently. Alternatively, it seems intuitive to intervene such that the learner is shown the correct state she should have gone to, and human teachers might use this strategy despite its sub-optimality with Q-learners. To explore these possibilities, we had human teachers interact with agents that were pre-programmed to improve over time. This gave us the opportunity to view how people would teach by intervention independent of the learning mechanism.

Experimental Design

Participants and materials Thirty-five MTurk participants took a dog training study that used the interface shown in Figure 2. On each trial, the dog would start at a tile and then walk to an adjacent tile. If the participant did not click on the dog at any point during its movement or within 1s of the dog entering the next tile, the next trial would start. If the participant clicked on the dog, then the dog "paused" and they could drag it to any tile on the gridworld and drop it. The dog then "unpaused" and the subsequent trial would then start at that tile. When the dog reached either "dog bowl," an animated dog treat would appear to indicate that the dog had experienced a reward. Entering either dog bowl tile ended an episode.

Procedure Before the main task, participants completed training trials that taught them how to intervene on the dog's behavior by picking it up. For the main experiment, they were told that they were trying to train a dog to perform a task on its own. The task was for the dog to *only* go to its own dog bowl, located in the upper-right tile, while avoiding their neighbor's dog bowl, located in the upper-left tile, and also avoiding the two lawns. Thus, the participants' goal in the task maps onto the teacher reward function shown in Figure 2. They had 12 "days" (i.e. episodes) in which they could train the dog, and they were told that each day ended once the dog became tired after 25 steps or became satiated by eating a dog treat. Each trial, the dog was programmed to execute the target policy with a probability of $1-\varepsilon$ and a random action otherwise. ε started at 1.0 for the first episode and then decreased by 0.1 each subsequent episode until $\varepsilon = 0.0$. This gave the impression that the dog was improving over time regardless of the intervention strategy used.

After the task was completed, participants were asked to answer several questions regarding their strategy, how well the dog responded, task difficulty, expected training efficacy, expected efficacy with a real dog, dog ownership, dog training experience, and several demographic questions.

Results

Intervening People make relatively sparse, local interventions that match the *undoing* model. Participants intervened on learners' behavior more when the learner performed a non-target action than when they performed a target action (non-target: M = 0.66, S.D. = 0.22; target: M = 0.06, S.D. = 0.10; paired t-test: $t(34) = 13.77, p < .001$). Additionally, the proportion of non-target actions that were intervened upon was between 0.5 and 0.75, the regime where *backward chaining* and *undoing* perform comparably. Interventions were also fairly local and close to the final state that resulted in the learner's action (Average Manhattan Distance between next state and intervention: M = 1.64, S.D. = 0.49). This indicates that *backwards chaining* was not often used as a strategy since that strategy requires making more global interventions. Finally, as Figure 4a reveals, many participants performed *undoing* interventions in which an agent that took a non-target action was placed back into its original position (Correcting: M = 0.15, S.D. = 0.14; Undoing: M = 0.59, S.D. = 0.24; Other: M = 0.27, S.D. = 0.19; $\chi^2(2) = 335.89, p < .001$).

Teaching Q-learners To compare human and model strategies, we used participants' responses to train Q-learners in the same task. We approximated how participants would have taught real learners by sampling from their responses to a learner's action in the task whenever a simulated learner took the same action. If a particular participant never observed an agent's take a simulated action, the default response was to not intervene. These results are plotted in Figure 3 for comparison with the simulation results.

Discussion

Our simulations revealed important interactions among teaching strategy, intervention interpretation, and intervention probability. In particular, *undoing*, which involves local changes to an agent's state, is an especially effective strategy only when interventions are frequent, while *backward chaining*, which involves state-changes reflecting the global structure of the task, is moderately effective regardless of intervention frequency. Incidentally, when people teach by intervention, they typically engage in *undoing*, but they do it less often than they should to train Q-learners (66%). Generally, people make moderately frequent local interventions.

As this is a preliminary investigation into teaching by intervention, this work has limitations. We use Q-learning as the learner, but other RL algorithms may respond better to human interventions. And given previous work showing that people often teach with communicative intent (Shafto, Goodman, & Griffiths, 2014; Ho et al., 2015), it may be that the standard RL framework is inadequate for capturing peoples' relatively sparse, local interventions. Future work will also need to test a wider range of MDP tasks. Nonetheless, these simulations and models are a first step towards understanding the everyday phenomenon of teaching by intervention in humans.

Acknowledgments

MKH is supported by the NSF GRFP (DGE-1058262).

References

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Brugger, A., Lariviere, L. A., Mumme, D. L., & Bushnell, E. W. (2007). Doing the right thing: Infants' selection of actions to imitate from observed event sequences. *Child Development*, *78*(3), 802–824.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340.

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*(2), 213–233.

Ho, M. K., Littman, M. L., Cushman, F., & Austerweil, J. L. (2015). Teaching with rewards and punishments: Reinforcement or Communication. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 920–925). Austin, TX: Conference Science Society.

Király, I., Csibra, G., & Gergely, G. (2013). Beyond rational imitation: Learning arbitrary means actions from communicative demonstrations. *Journal of Experimental Child Psychology*, *116*(2), 471–486.

Knox, W. B., & Stone, P. (2015). Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. *Artificial Intelligence*, *225*, 24–50.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

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

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279–292.