Teaching Agents with Evaluative Feedback:
Communication versus Reward

Mark K Ho

Advisor: Michael L. Littman

People often use rewards and punishments to modify the behavior of other agents like children and pets. Consider Eve, a 4-year-old who loves to build dirt castles. Whenever she comes back into the house, much to the chagrin of her parents, she tracks traces of her muddy creations all over the living room carpet. Eve’s parents want her to learn how to behave in the house, if not for their sake then her own as a future functioning member of society.

Merely explaining that mud in the house is bad is probably insufficient to accomplish this, rather, some combination of rewards and punishments will be needed to motivate Eve to change her behavior. Indeed, the use of praise, positive nonverbal responses, reprimand, and negative nonverbal responses to children who follow parental directives has been shown to increase compliance (Owen, Slep, Heyman, 2012). Furthermore, carrots and sticks can be used not only with children, but also with animals (Breland and Breland, 1961) and artificial agents (Sutton & Barto, 1998).

This thesis explores how people use rewards and punishments in order to modify the behavior of another organism. I will first discuss two theories of how people use evaluative feedback and previous research related to each. One theory characterizes teaching with evaluative feedback as an attempt to incentivize behavior. The other characterizes it as a communicative act to be mutually recognized. Then I will discuss formal models of these two theories based on work in reinforcement learning and social robotics. Finally, I describe experiments and results that speak to the plausibility of each theory.

From a psychological perspective, understanding how people represent other agents during ongoing interaction reveals important mechanisms that operate during social cognition that have implications for human pedagogy. For computer scientists and roboticists, characterizing user expectations during social interaction provides a framework for designing robots that integrate seamlessly with the natural human tendency to teach and communicate.

Evaluative Feedback as Incentives

Since the behaviorists (Skinner, 1938), psychologists have been interested in how organisms learn from positive and negative experiences with the world. More recently, computer scientists and computationally oriented psychologists have characterized learning mechanisms in terms of reward maximization (Sutton & Barto, 1998; Dayan & Niv, 2008). During teaching, humans could also conceive of their feedback as face value rewards and punishments for the learner. That is, a teacher might reason that from a learner's perspective, positive responses are pleasurable, rewarding outcomes while negative responses are painful, punishing ones. A capable learner should then learn the behavior that maximizes 'rewards' while minimizing 'punishments'. On this view, behavioral modification simply involves constructing the appropriate reward schedule for a task.

At first glance, there are a number of reasons why one might suppose that people use rewards and punishments to incentivize behavior. First, to a large extent, people themselves are reward-maximizing agents that attempt to find the most efficient way to get the most rewards from their environments. Most people experience pain and pleasure, and do what they can to avoid the former while getting the most of the latter. Teachers may simply be projecting their own experiences with pain and pleasure onto others while giving evaluative feedback.

Second, it is well established that people utilize "theory of mind" or adopt an "intentional stance" while intepreting the behavior of other agents (Dennett, 1989; Malle, 2005). In particular, people often explain the behavior of others in terms of beliefs – representations of the world – and desires – evaluations of states in the world. From early infancy onwards, humans expect agents to behave “efficiently” to accomplish goals determined by their desires (Gergely and Csibra, 2003). Furthermore, adult judgments about what agents think and want closely match the estimates of inverse reinforcement-learning models that infer beliefs and desires based on agent behavior (Baker, Saxe & Tenenbaum, 2009). At least within relatively familiar and transparent domains, people can accurately represent how beliefs, desires, and actions interrelate. Using rewards and punishments to incentivize the behavior of another agent would simply be another straightforward application of this pre-existing capacity.

Evaluative Feedback as Communicative Acts

But while teacher-delivered rewards and punishments can serve as incentives, they can also be characterized as uniquely social acts separate from the non-social environment. A stimulus that is presented with the intent to teach or communicate can convey different information than one that merely appears asocially (Sperber & Wilson, 1986). Research into the properties of pedagogy suggests that people can recognize teachers’ communicative intent, and use this to draw stronger inferences about stimuli than would otherwise be warrented. Young children are especially responsive to pedagogical intent (Csibra & Gergely, 2009).

Recognition of communicative or pedagogical intent has been shown to influence learning in a number of domains such as using examples to teach categories or during word learning (Shafto et al., 2014; Frank & Goodman, 2012). As an example, imagine two children, one who happens to come across a picture of a zebra and another that you have just given a picture of a zebra to. Compared to the first case, if you told the second child that the animal in front of her was a zebra, she would be more likely to infer that zebras are uniquely characterized by their stripes. This is because a teacher who gives informative or representative examples to illustrate a concept would only have used this strange striped horse as an example of a zebra if being a strange striped horse was characteristic of being a zebra. This assumption is not present when the example is a random, unintentional sample that just happens to be a strange striped horse.

Along these lines, people may conceive of using rewards and punishments to teach as involving a presupposition of communicative intent on the learner's part. In contrast to teaching a reward-maximizing learner, an action-feedback learner expects responses to be communicative or commentary about an action. Rewards signal to the learner that the action performed was correct given the circumstances, whereas punishments signal that the action was wrong or incorrect. Teachers further expect such a learner to be motivated to perform correct actions and avoid incorrect ones in a given state.

Models

First, I will describe an interaction model of the teacher-learner dynamics during teaching with evaluative feedback. Second, I will propose two learner models (reward-maximizing and action-feedback) that capture how teachers use rewards and punishments to modify behavior.

Interaction Model

The interaction between a teacher and a learner can be modeled as a Markov Decision Process (Bellman, 1957). A teacher and learner have a shared representation of the environment consisting of states ($s \in S$), actions ($a \in A(s)\ \forall s$), and transitions ($T: (s, a) \rightarrow s$). The learner performs actions in states, which results in transitioning to a new state. In this paper, the teacher is effectively assumed to be omniscient and placeless with respect to the MDP that the learner inhabits.

On each timestep t, the learner generates an action from a behavioral repertoire, represented as a policy π_t, which is a mapping from states to available actions ($\pi\colon s \to a \in A(s)$). After observing the learner's current state, action, and subsequent state(s_t, a_t, s_{t+1}), the teacher responds to the learner with a positively or negatively valenced feedback signal of a finite magnitude ($f_t \in [-1,1]$). The function that takes an observation of the learner and returns feedback called the *feedback* function:

$$F(s_t, a_t, s_{t+1}) = f_t$$

The pattern of reward and punishments that constitute this feedback function is determined by the target policy, π^*, that the teacher wants the learner to acquire. In other words, the teacher is modeled as having an intention to teach a specific behavior using rewards and punishments as a response to actions of the agent. This formulation suggests that the feedback function is a static mapping from learner transitions to rewards and punishments, however, F could in princple change dynamically over the course of the teaching/learning history.

For simplicity, I assume that state transitions, learners' learned policies, and teachers' feedback functions are deterministic. A more comprehensive model of teacher-learner dynamics that takes the noise induced by the environment, learner, and teacher into account is conceptually straightforward. However, it may be that introducing uncertainty may have unexpected results since this interaction involves recognizing the intent behind and not simply the consequence of agents' actions. Future work should explore this possibility.

Modeling Learning

Learning consists of changes in how an agent responds to stimuli. This can be modeled as changes to a learner's policy over time $(\pi_0, \pi_1, \ldots, \pi_{T-1}, \pi_T)$ and as convergence to a final *learned policy*, π_T. A teacher using evaluative feedback can be thought of as steering the learning trajectory through policy space such that it settles on the desired set of behaviors. Thus, each model must characterize how the feedback function, F, relates to a learned policy, π_T, and the mechanism that modifies a policy at each time step.

Modeling Reward Maximization The reward-maximizing agent treats teacher-feedback from a feedback function as a face value reward to be maximized over the long term – exactly like the reward signal found in standard reinforcement learning (RL) (Sutton & Barto, 1998). That is, a reward-maximizing agent calculates the cumulative long-term value of each available action a in the current state s_t, under the current policy π_t. Call this the *action-value, $q_\pi(s,a)$,* from a state with a policy:

$$\begin{aligned} q_{\pi_t}(s_t, a_t) &= f_t + \sum_{k=1}^{\infty} \gamma^k f_{t+k} \\ &= F(s_t, a_t, s_{t+1}) + \sum_{k=1}^{\infty} \gamma^k F(s_{t+k}, \pi_t(s_{t+k}), s_{t+k+1}) \,. \end{aligned}$$

Importantly, future rewards may be treated as less rewarding than immediate ones, so a discount parameter $0 \leq \gamma \leq 1$ is included. As its name suggests, the reward-maximizing agent is interested in eventually learning a policy, π_{RM}, that maximizes the action-value in all states. Thus, such an agent learns the policy:

$$\pi_{\text{Reward-Maximization}}(s) = \underset{a \in \mathcal{A}(s)}{\operatorname{argmax}} \max_{\pi} q_{\pi}(s, a)$$

for all $s \in \mathcal{S}$.

Reinforcement learning (Sutton & Barto, 1998) describes a number of algorithms that converge on π_{RM} given a reward signal to be maximized. Here, I will focus primarily on two classes of algorithms that have this property: model-free Q-learning, and model-based learning. One motivation for limiting myself to these two is that they each capture two very broad types of reward-maximizing learning characterized in humans and animals (Dayan & Niv, 2008). Q-learning aligns closely with associative learning theories that have their roots in behaviorist accounts of learning from reward and punishment. Conversely, model-based learning mechanisms have been implicated in multistep planning during learning.

The standard Q-learning algorithm estimates action-values, $q(s, a)$, under the current policy, π_t, while exploring the world and exploiting the rewards within it, eventually converging on the true action-values (Watkins & Dayan, 1992). The agent does not have an explicit representation of transitions in the world or rewards, but rather uses its own experience of transitioning in the world to evaluate the value of various actions. Formally, Q-learning updates its current action-values according to the following rule (where $0 < \alpha < 1$ controls the learning rate):

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha \left[f_{t+1} + \gamma \max_{a} q(s_{t+1}, a) - q(s_t, a_t) \right].$$

In contrast, model-based learning algorithms maintain a representation of state transitions and reward functions in the world. This allows the learner to deduce the optimal policy given what is known about the world. Algorithms like Rmax have the learner simultaneously learn the transition model of the world and the world reward function (Brafman & Tennenholtz, 2002). Here, however, I will assume that the learner has a complete and accurate model of state transitions and is mainly concerned with learning and exploiting an unknown feedback function.

Modeling Action-Feedback The action-feedback agent treats feedback as a direct signal for the correctness or incorrrectness of an action. A positive teacher response indicates that the action matches the corresponding action in the target policy, while a negative response indicates it does not match (Loftin et al., 2014). Thus, teacher responses map directly onto whether an action should or should not be done, and we can define the *action-correctness*, $j(s,a)$, from the present state as:

$$j(s_t, a_t) = f_t = F(s_t, a_t, s_{t+1}).$$

For all states and actions, j is initialized to 0. One intuitive model of how a teacher determines what is correct and incorrect could be as follows: an action is correct if it maches the corresponding action in the teacher's desired policy, and it is incorrect if it does not match (suppose also that there is some small possibility that the action is neither correct nor incorrect). These relationships among the target policy, a learner's action in a state, and the correctness of an action can be captured in the form of a likelihood ($\lambda > 0$):

$$\mathrm{P}(j(s,a) > 0 | \pi^* = \pi) = \begin{cases} 1-\lambda & \text{if } \pi(s) = a \\ 0 & \text{if } \pi(s) \neq a \end{cases}$$
$$\mathrm{P}(j(s,a) = 0 | \pi^* = \pi) = \lambda$$

$$\mathrm{P}(j(s,a) < 0 | \pi^* = \pi) = \begin{cases} 0 & \text{if } \pi(s) = a \\ 1 - \lambda & \text{if } \pi(s) \neq a \end{cases}$$

To infer the target policy being communicated by the teacher, the action-feedback learner can use these likelihoods to calculate a distribution over possible policies given action-correctness according to Bayes rule:

$$\mathrm{P}(\pi^* = \pi | \, j\,) = \frac{\prod_{s,a} \mathrm{P}(\, j(s,a) | \pi^*) \mathrm{P}(\pi^* = \pi)}{\sum_{\pi} \prod_{s,a} \mathrm{P}(\, j(s,a) | \pi^*) \mathrm{P}(\pi^* = \pi)}$$

Finally, the learner will adopt the policy with actions most likely to be correct for all $s \in \mathcal{S}$:

$$\pi_{Action-Feedback}(s) = \underset{a \in \mathcal{A}(s)}{\operatorname{argmax}} \mathrm{P}(\pi^*(s) = a | \, j\,)$$

Note that an important part of this model is the prior over policies that the agent has. In the experiments below I discuss some possible priors that learners may have.

Distinguishing Reward-Maximization from Action-Feedback

How could a person trying to teach a reward-maximizing learner be distinguished from one trying to teach an action-feedback learner? That is, when does $\pi_{RM} \neq \pi_{\mathrm{AF}}$ for a feedback function F? Furthermore, when does a reward-maximizing learner or an action-feedback learner acquire the target policy, π^*?

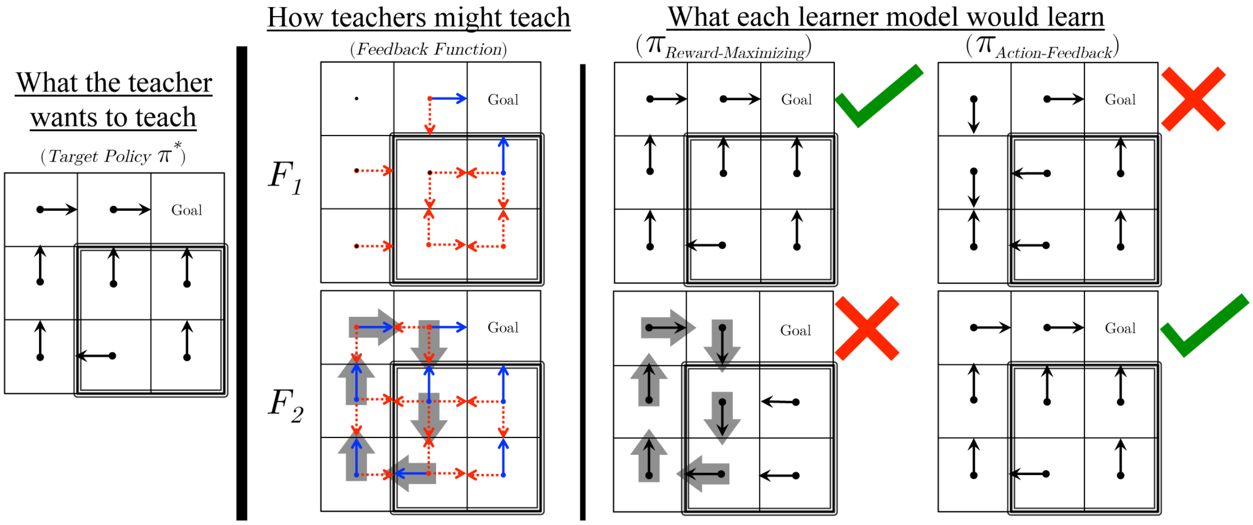
For the learner models, the reward-maximizing discount parameter, γ, must be sufficiently large. Otherwise, the learner's estimate of an action's *correctness* and its *value* coincide $q_{\pi_{RM}}(s,a) = j(s,a)$ for all s, a, and $\pi_{\mathrm{RM}} = \pi_{\mathrm{AF}}$. This means the two can only diverge when the reward-maximizing learner cares about future feedback.

For feedback functions, the learned policies of the models can diverge given *positive cycles*: state-action-feedback sequences where the learner returns to an initial state,

$(s_0, a_0, s_1, a_1, \ldots s_n, a_n, s_0, \ldots)$, but receives a net positive reward, $F(s_0, a_0, s_1) + \gamma F(s_1, a_1, s_2) + \cdots + \gamma^n F(s_n, a_n, s_0) > 0$ (Ng, Harada, & Russell, 1999).

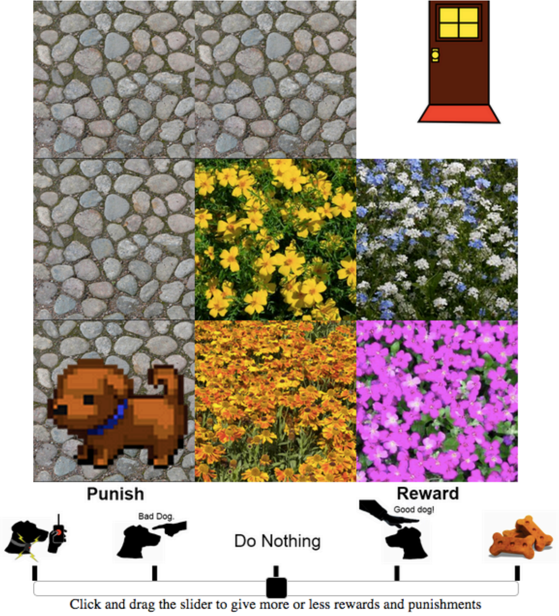
For example, consider what happens if Fido is punished for going into the garden but rewarded for getting on the path or heading towards the house. Suppose Fido heads towards the house along the path, gains rewards, and stops at the door. At this point, Fido could enter the house and get a final, perhaps large, reward. But, if Fido is a reward-maximizing learner who values future rewards, he could double back through the garden, take the punishments, follow the path to the house again, and gain even more rewards. If the tradeoff between punishments and rewards is a net gain, this is a positive cycle. Figure 1 illustrates the predicament of Fido's owner in a simplified gridworld.

Figure 1: The task faced by Fido's owner. Tiles enclosed by double lines are the garden; unenclosed tiles are the path. The owner wants to teach Fido π^*. The two rows show two possible *feedback functions* F_1 and F_2 (solid blue arrows are rewards, dotted red arrows are punishments) as well as the policies learned by the two models. A reward-maximizing learner will not learn the target policy under F_2 because of the *positive cycle* (big grey arrows). Note that a feedback function may not yield a unique an action-feedback policy.


What the teacher wants to teach
(Target Policy π^*)
Goal
How teachers might teach
(Feedback Function)
F_1
F_2
Goal
Goal
What each learner model would learn
($\pi_{Reward\text{-}Maximizing}$)
($\pi_{Action\text{-}Feedback}$)
Goal
Goal
Goal
Goal

I designed a dog-training paradigm, the Garden-Path task, reminiscent of the one faced by Fido's owner (Figure 2) to determine whether people produce positive cycles, the presence of which would indicate that people expect action-feedback but not reward-maximizing learners. Dogs were chosen because people are unlikely to attribute sophisticated cognitive capacities to them (unlike with human children) but are likely to be familiar with them (unlike robots).

Figure 3: Garden-Path task. On each trial, a dog moves and then participants give their feedback.


Punish
Reward
Bad Dog.
Good dog!
Do Nothing
Click and drag the slider to give more or less rewards and punishments

Experimental Overview

The models and paradigm described above outline several possible theories of how people use rewards and punishments to teach and a way to empirically distinguish them. The following section will describe 3 sets of experiments run with human participants that test these accounts.

Experiment 1 investigated peoples' teaching patterns for isolated actions taken by a learner to determine peoples' "stationary" feedback functions. Experiments 2a and 2b investigated teaching a single learning agent over time when that agent is improving. This tests how teachers give feedback over time and once the agent successfully learns a task. Finally, Experiments 3a and 3b examine how people teach responsive learners that incorporate their feedback to learn a task. By having participants train implementations of the models described above, I can determine what default teaching strategies people use, and how flexibly they switch between strategies.

The broad goal of these studies is to determine how people conceptualize teaching with rewards and punishments. That is, do they use feedback as incentives for behavior or with the presumption of communicative intent? A secondary goal is to characterize other qualtiative features of teaching. For instance, these methods can test how people behave once the task has been learned and whether people change strategies. Additionally, I discuss individual differences and their impact on task behavior.

Experiment 1

In Experiment 1, participants provided feedback to learners who performed isolated actions in the Garden-Path task. This enables me to 'map out' their feedback functions over the entire state-action space.

Method

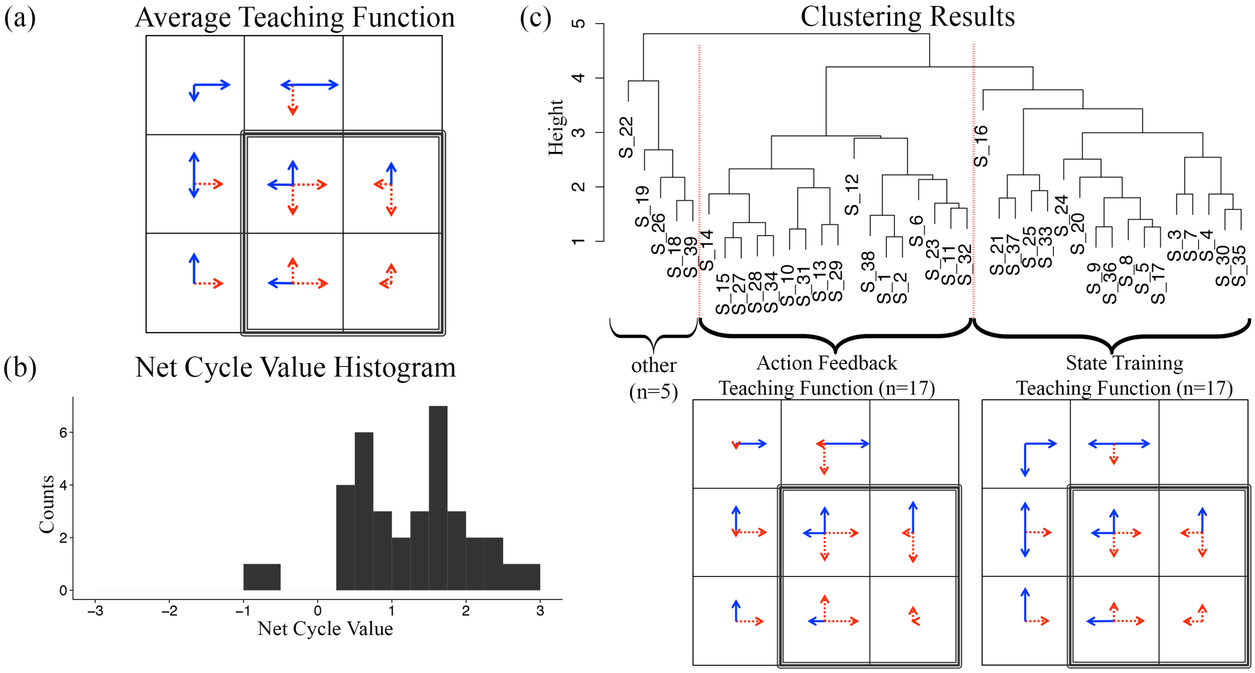
Participants and materials Forty people from Amazon Mechanical Turk participated, and one was excluded due to a technical error (18 female). On each trial the dog starts at a tile, rotates to face one of the four cardinal directions, and then walks onto the adjacent tile (3000ms). After viewing the dog's movement, participants provide feedback ranging continuously from highly negative to highly positive: "a mild but uncomfortable shock" to scolding the dog ("Bad Dog") to "doing nothing" to praising the dog ("Good dog!") to "a few delicious treats". The instructions explicitly stated that the scale should be seen as ‘balanced’ such that distances from the midpoint of the scale were equivalently positive or negative.

Procedure Participants were told that they would help train a school of 24 distinct dogs to “go into the house by staying along the path and staying out of the garden” and that the goal of training is for each dog to be able to do this independently. The entire task consisted of 24 trials that covered all possible initial locations, actions, and final locations. Trial order was randomized under the constraint that no trial began where the previous trial had ended. Participants were asked to imagine they had placed the dog in that location at the beginning of the trial. They had to answer several comprehension questions completely correctly to start the task.

After completing the main task, participants were asked several questions about their training and background. Questions included “Would you say you tended to use punishments or rewards to train?”, “Did you change the amount you punished or rewarded over the course of the task?”, “How effective do you think you training strategy would be with a real dog?”, “Do you currently have a dog?”, “How much life experience have you had with dogs?”, “Do you currently have young children?”, “How much life experience have you had with young children?”, and a Cognitive Reflection Test proxy that asks if the participant would prefer a 15% chance of winning $1,000,000 or a 100% chance of winning $500.

Additionally, questions were asked about the dogs’ preferences with respect to the response scale. For eight sequences of punishments and rewards, participants answered whether they thought the dog would prefer the sequence, nothing, or both equally. The sequences tested were: (1) scolding twice followed by praising twice, (2) two scoldings followed by three praises, (3) two scoldings followed by four praises, (4) one shock followed by one biscuit, (5) one shock followed by two biscuits, (6) one shock followed by two praises, (7) one shock followed by three praises, (8) two scoldings followed by one biscuit. The final page asked several demographic questions and for feedback.

Figure 3: Results of Experiment 1. (a) Average teaching function of all participants. (b) Net value of responses on cycle trials by individual. (c) Results of hierarchical clustering of participants' responses with the average teaching function of the two largest clusters. These correspond to an action-feedback function and a "state-training" function (see text).



Results

Positive Cycles I first analyzed whether participants' stationary feedback functions had positive cycles that could be discovered by a reward-maximizing learner. Figure 3a graphs the average feedback function, where the response scale was coded as between -1 and +1. The aggregated pattern of feedback reveals that starting from the lower left-hand corner and performing the action sequence <up, up, right, down, down, left> yields a net positive feedback. This *positive cycle* had an average value of +1.20, SE=0.20 ($t(38) = 5.99$, $p < .001$). Furthermore, individual-level responses had positive cycles. Figure 4b is a histogram of net cycle values and clearly demonstrates that 36 out of 39 participants delivered a net positive reward along this route.

Feedback Function Types Previous work has shown that people adopt different 'training strategies' when giving RL agents rewards and punishments (Loftin et al., 2014). To identify individual differences in feedback functions, I performed a hierarchical clustering analysis. Individual feedback functions were represented as 22-dimensional vectors of responses between -1 and 1 (actions from the terminal state were not included), and we calculated a Euclidean-distance dissimilarity matrix. Clusters were identified using a complete linkage method.

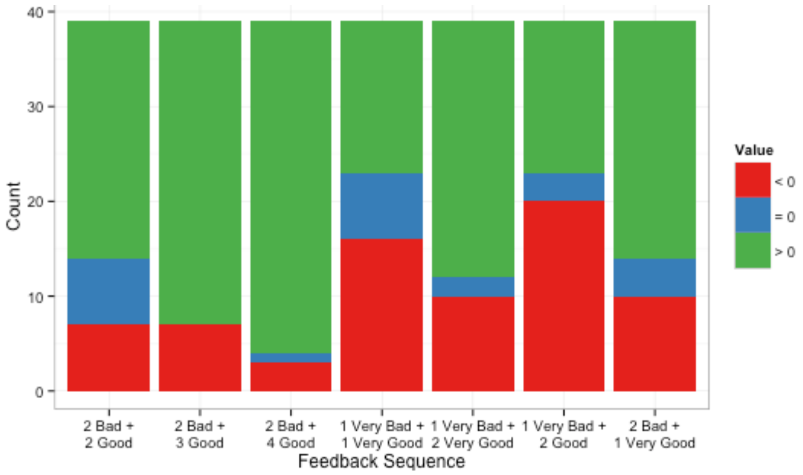
Results (Figure 3c) reveal two large, homogeneous clusters (n=17 each) and a single small, heterogenous cluster (n=5). The first large cluster (left) closely matches the action-feedback model that rewards correct actions and punishes incorrect ones. The two subclusters in this cluster reflect response magnitude differences. The second large cluster (right), reveals a feedback pattern distinct from either the reward-maximizing or action-feedback model. Participants gave rewarding responses based on the general permissibility/impermissibility of state-types, even if they were not correct for the specific task being trained. For example, if the dog stayed on the path but walked away from the door, a "state training" teacher would still give a reward. This leads to even worse positive cycles that could be exploited by a reward-maximizing agent who simply walks back and forth along the path. Importantly, only 5 of the 17 state training participants did not mention 'going to the house' in a pre-task free-response question, suggesting it is not due to a misunderstanding of the task. Noticably, only one participant (found in the small 'other' cluster) showed a 'reward-maximizing' pattern.

Self-Report Strategy Questions When asked whether they used rewards and/or punishments to teach, 87% of participants answered that they used a combination of punishments and rewards to teach, 10% answered that they mainly used rewards, and 3% (1 participant) indicated punishments. When asked whether they changed the amount they rewarded or punished over the course of the task, 46% responded they changed, 48% responded they did not change, and 6% responded they were not sure.

Response Scale Values and Dog Preferences Figure 4 plots the proportion of responses for each of the 8 dog preference questions asked following the end of the task. 79% of participants believed that two scoldings followed by two praises was equal to or greater than nothing, suggesting that the scale was perceived symmetrically (or reward biased) at moderate ranges. 85% of participants responded that one shock followed by one biscuit was equal to or greater than nothing, which corresponds to the scale being symmetric or reward biased at extreme values.

Additionally, participants perceived that the dog would assign a positive net value to the future expected rewards in the positive cycle (i.e. γ is sufficiently large). 92% of participants responded that the dog would prefer 2 scoldings (-.5 twice) followed by 4 praisings (.5 four times) to nothing (0), indicating that $(-.5) + \gamma(-.5) + \gamma^2(.5) + \gamma^3(.5) + \gamma^4(.5) + \gamma^5(.5) > 0$ (i.e. $\gamma > .79$). Most participants (85%) used rewards greater than or equal to punishments on cycle trials, indicating that most would expect a reward-maximizing agent to prefer the identified cycle at measured values.

Figure 4: Judgment counts for dog preference questions



Individual Differences An analysis based on gender, experience with dogs, experience with kids, and the CRT gambling proxy showed no behavioral differences. Table 1 shows feedback strategy counts for gender, dog ownership, parenthood, and response to the CRT proxy. There were no differences in behavior due to gender. People who identified as male versus female showed no difference in net cycle value (t(37.0) = -0.57, p = 0.56) nor were they more or less likely to show action-feedback or state-training patterns ($\chi^2(1, N = 39)$ = 0.12, p = 0.73).

If different teaching behaviors (in particular, producing or avoiding positive cycles) tend to be learned, then first-hand experience with dogs or children should make a difference for cycle values or training strategies (e.g. action-feedback or state-training). Otherwise, the absence of any difference would suggest people have an innate strategy for teaching with rewards and punishments.

Participants reported having a fair amount of experience with dogs: 33% reported owning a dog, experience with dogs on a 5 point scale was high (Mean = 3.25, SE = 0.17), and confidence that their strategy would work with a real dog on a 7 point scale was high (Mean = 5.46, SE = 0.13). However, none of these measures predicted teaching behaviors. For example, there was no relation between having a dog and using action-feedback or state-based teaching strategies ($\chi^2(1, N = 39)$ = 0.14, p = 0.71). Similarly, there was no difference in the net cycle value between those with dogs and those without (*t*(32.8) = 1.15, *p* = 0.26). Contrary to the expectation that those with dogs would be better at training reward-maximizing learners, life experience with dogs had a slight *positive* correlation with positive cycle values (*r*(37) = .33, *p* = .04). Peoples' judgments of how effective their feedback would be with a real dog was unrelated to cycle values (Pearson correlation: *r*(37) = -.18, *p* = 0.28).

Next, I looked at whether self-reports of child experience related to teaching behavior. 33% of participants reported having children, and on a 5-point scale, they reported a mean child life experience of 3.10 (SE=.26). Paralleling the results with dog experience, there was no detectable difference by parenthood. There was no relation between parenthood and strategy ($\chi^2(1, N = 39)$ = 0.13, *p* = 0.72). Nor was there any difference in cycle values between parents and non-parents (t(18.0) = -0.51, p = 0.62). Finally, the separate continuous measure of child experience did not correlate with cycle values (r(37) = .18, p = 0.27).

Finally, an analysis of the CRT proxy questions did not reveal any differences in terms of final cycle values or feedback strategy. A test for independence between feedback strategy and CRT proxy showed no difference ($\chi^2(1, N = 39)$ = 1.16, *p* = 0.28). Similarly, a comparison of cycle values between those who gave the low CRT answer (a 100% chance of winning $500) and those who gave the high CRT anser (a 15% chance of winning $1,000,000) showed no difference (*t*(19.0) = 1.219, *p* = 0.24).

Table 1: Experiment 1 Individual Difference measures and Feedback strategy

	Gender		Parenthood		Dog Ownership		CRT proxy	
Feedback Strategy	Male	Female	Parent	Non-Parent	Dog	No Dog	$500	$1,000,000
Action-Feedback	10	7	5	12	4	13	13	4
State-Training	8	9	7	10	6	11	9	8
Note: no comparisons were significant								

Discussion

In this first experiment, participants trained individual actions performed by different agents. Several key results emerge from this preliminary study. First, teachers readily produce positive cycles that a reward-maximizing agent would learn to exploit. Second, peoples' feedback cluster into two general types: action-feedback and state-training. This first type would be effective at training the action-feedback learner model, which assumes that rewards and punishments *signal* the correctness and incorrectness of actions respectively.

The state-training pattern of feedback was not originally predicted by either the reward-maximizing or action-feedback learner models. It may be that such teachers attempt to teach intermediate policies (e.g. "stay on the path") before teaching the complete policy. Alternatively, teachers may assume that the learner has a state-type representation of path- and garden-tiles and attempt to leverage this during teaching. Additional studies should test in what conditions people engage in state-training over action-training.

This study is also methodologically valuable since people quickly and easily understood how to use rewards and punishments to train another agent in a virtual environment. In particular, the response scale developed for this paradigm provides an intuitive interface for giving rewards and punishments. The analysis of dog preferences with respect to the scale shows that it is also straightforward to interpret experimentally.

The analysis of individual differences showed that teaching behavior does not relate in a systematic way with experience teaching dogs or children. This surprising result suggests that teaching as if learners are not reward-maximizing is an innate tendency unaffected by experience.[1]

The design of the experiment abstracts away from any history that a teacher may have with a learner. This permits comparison of different participants' reponses to the same learner

[1] An alternative possibility is that experience simply confirms the teaching strategy that people have innately.

action independent of previous or future actions. The downside of this approach is that teaching, by its very nature, is interactive, ongoing, and contextual, and involves a teacher responding to a learner's entire policy (and not just individual actions).

Experiments 2a and 2b

Experiment 1 examined responses to isolated actions. In Experiments 2a and 2b, participants teach a single dog over time. This tests whether teachers can properly track a learner's policy, whether positive cycles arise during online interaction, and whether any systematic patterns emerge when training over time. In particular, teacher-delivered rewards and punishments might track the current estimate of the learner's ability on the task – e.g. how close their current policy is to the target policy. Alternatively, feedback might be responsive to *changes* in the learner's ability on the task or their apparent improvement. If teachers primarily reward for improvements and punish for stagnation, this would lead them to decrease net rewards once the learner appears to have acquired the target policy.

Experiment 2a

Method

Participants and materials The same interface as Experiment 1 was used. Forty people participated, but 3 were excluded for technical reasons (16 female). Participants were told they would train a single dog over 8 game days. Each day, the dog began in the lower left corner and movements on each day were predetermined. Apparent performance improved over the course of the first 5 days, were optimal on the 6th and 7th days, and on the 8th day the dog proceeded on the positive cycle steps identified in Experiment 1. Except for the final day, the dog's behavior on days 1 through 7 was generated by choosing the optimal action in a given state with a probability

$1 - \epsilon$ or any of the actions with a probability $\epsilon/(\# \text{ actions} - 1)$. ϵ was 1.0, 1.0, 0.45, 0.1, 0.1, 0.0, and 0.0 for days 1 to 7 respectively. Unless the dog made it to the door, at which point that day ended, each day was 6 steps long. All participants were shown the same pre-determined set of actions.

Procedure I told participants that they would train a single dog over the course of 8 game days and that at the end of the experiment, the dog would be tested, on its own, 3 times at the beginning of the path. A bonus was contingent on the dog's performance (but everyone won the full bonus). Between each game day, participants answered questions regarding the dog's current ability and its improvement since the last day (only after days 2-8).

Following completion of the task, participants answered the same questions as in Experiment 1, including the dog preference questions. Additionally, they were asked "How responsive did you feel the dog was to your feedback?", "Overall, how good do you think you were at training the dog in this task?", and "Do you have experience training dogs?"

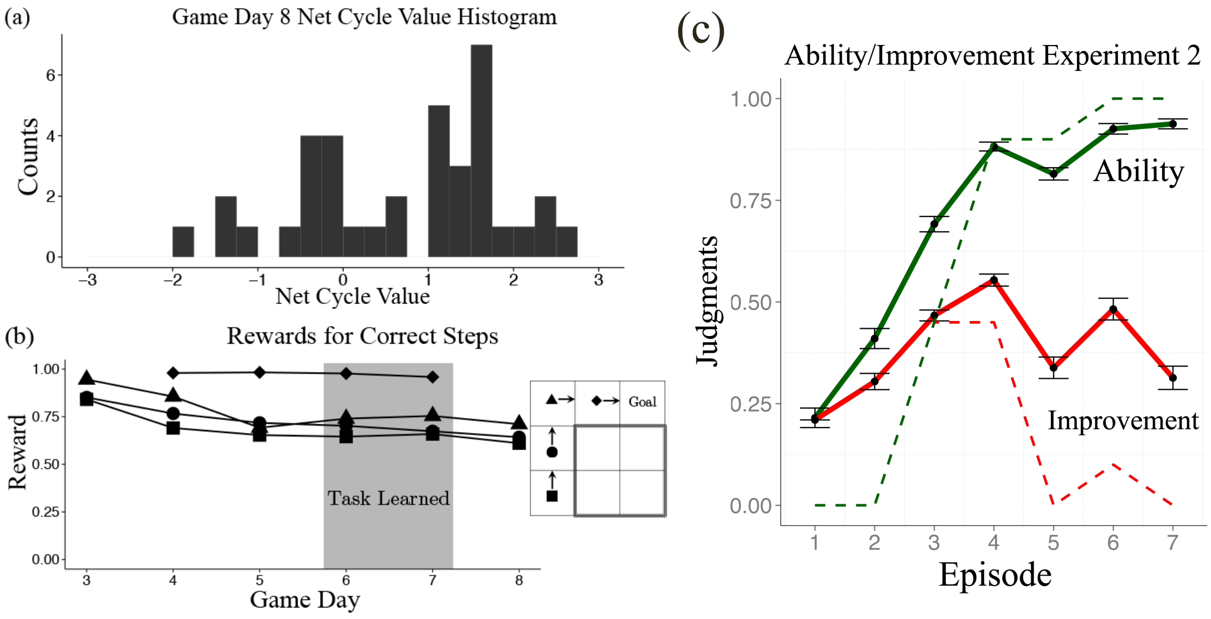
Results

Perception of Task Participants believed that they were teaching the dog effectively. All responses to a 5-point question about dog responsiveness were above 1 = not responsive at all (mean=3.45, SE = .11). Furthermore, all responses to the question about how good they were at training were above 3 on a 7-point scale (mean=5.48, SE = .12). Additionally, dog preference judgments were nearly identical to those reported in Experiment 1.

Positive Cycles and Diminishing Rewards When teaching a single learner over time, most participants' feedback functions showed positive cycles. The final day in the dog training task had the dog take the 6 steps corresponding to the positive cycle identified in Experiment 1. Although smaller, the average total reward for these 6 steps was still a positive value: +0.67, SE=0.19 (one-sided t-test: t(36)=3.53, p < .001). As compared to Experiment 1, however, fewer participants had a net positive cycle value on day 8 (24 out of 37, Figure 5a).

Consistent with smaller and fewer positive cycle values on the final day, rewards for correct steps declined but remained positive over days 3 to 8. A repeated measures ANOVA of responses with Day and Action as factors showed both main effects (Day: F(1,36) = 15.69, p < 0.001; Action: F(3, 108) = 47.0, p < 0.001) and an interaction (Day x Action: F(3, 108) = 4.78, p < 0.01). This suggests that although people do produce positive cycles consistent with action-feedback expectations, some teachers attempt to 'wean' the learner off of rewards (Figure 5b).

Figure 5: Experiment 2 results. (a) Histogram of final game day net cycle values. (b) Average rewards for each of the 4 correct steps on each day. (c) Average ability and improvement judgments over the 8 game days (solid lines) along with the 'true' ability and improvements in terms of $1 - \epsilon$ (dotted lines).



Tracking Learner Ability and Improvement Participants only have access to the learner's interactions with the environment, and so can only infer its policy indirectly. Despite this, judgments of the dog's ability at the task following each day tracked the value of $1 - \epsilon$

extremely closely (mean Pearson correlation = 0.93, SE=0.008; t(36)=119.67, p < .001). Similarly, judgments of the dog's improvement tracked day-to-day changes in ϵ (mean Pearson correlation = 0.85, SE=0.014; *t*(36) = 62.39, *p* < 0.001). Thus, when teaching via evaluative feedback, teachers infer the current state of the learner's policy and track changes to that policy over time as the interaction model assumes. (Figure 5c)

Individual Differences Final day cycle values were analyzed based on participants' responses to questions regarding dog experience, child experience, CRT proxy, gender, and political ideology. Means and comparisons between groups are summarized in Table 2.

More experience with dogs was associated with greater postive cycle values in the final day. For instance, people who reported owning a dog had significantly higher average positive cycle values than those who did not (t(32.0) = 2.11, p = 0.04). Similarly, people who reported having had at least some experience training dogs had higher values, though this was not significant (t(29.1) = 1.17, p = .25). Further corroborating this pattern of results, there is a positive but non-significant correlation between reported dog life experience (Mean = 3.0 on a 5 point scale; SE = 0.19) and final day positive cycle value (r(35) = .17, p = 0.30).

In contrast to dog experience, experience with children revealed a non-significant negative correlation with final cycle value (r(35) = -.27, p = 0.10). Only 4 out of 37 participants reported having children, so I did not include an analysis based on this question.

As in the first experiment, CRT proxy question responses and gender did not reveal any important patterns with respect to cycle values. Those who chose the $500 option had higher cycle values, but this difference was not significant (t(34.5) = 0.64, p = .52). Men had non-significantly lower net cycle values (t(33.8) = -.30, p = .76).

An analysis based on political ideology, however, did reveal systematic differences. Self-reported Conservatives had lower final cycle values than Moderates who had lower ones than Liberals (F(2,34) = 5.133, p = 0.01). This suggests that attitudes towards using rewards and punishments to modify behavior may be closely related to general social and political attitudes.

Table 2: Experiment 2a Final Day Cycle Costs by Individual Differences

Individual Difference		n	Final Day Mean Cycle Cost		
Dog Ownership	Dog	12	1.16 (0.23)	*t*(32.0) = 2.11	*p* = 0.04
	No Dog	25	0.44 (0.25)		
Dog Training Experience	At least some	10	0.96 (0.23)	*t*(29.1) = 1.17	*p* = 0.25
	None	27	0.57 (0.25)		
CRT Proxy	$500	17	0.81 (0.28)	*t*(34.5) = 0.64	*p* = 0.52
	$1,000,000	20	0.56 (0.27)		
Gender	Male	21	0.62 (0.26)	*t*(33.8) = -0.30	*p* = 0.76
	Female	16	0.74 (0.28)		
Political Ideology	Conservative	7	0.01 (0.42)	*F*(2, 34) = 5.13	*p* = 0.01
	Moderate	11	0.17 (0.36)		
	Liberal	19	1.20 (0.21)		

Note: Standard Errors in parentheses

Experiment 2b

Experiment 2a showed that people only slightly decrease their rewards over time. However, the learner only performed the perfect sequence of actions two days in a row. Experiment 2b sought to test whether teachers would completely remove rewards given a longer series of perfect days.

Method

Participants and materials Forty-one Amazon Mechanical Turk workers participated, with no exclusions (13 female). The general structure of the experiment was the same as

Experiment 2a. As before, the agents were deterministically preprogrammed with actions derived using different values of ϵ. The only difference was that participants trained learners that improved over days 1 to 4, performed perfectly over days 5 to 11, and regressed on days 12 and 13. On day 12, the learner performed the identified positive cycle.

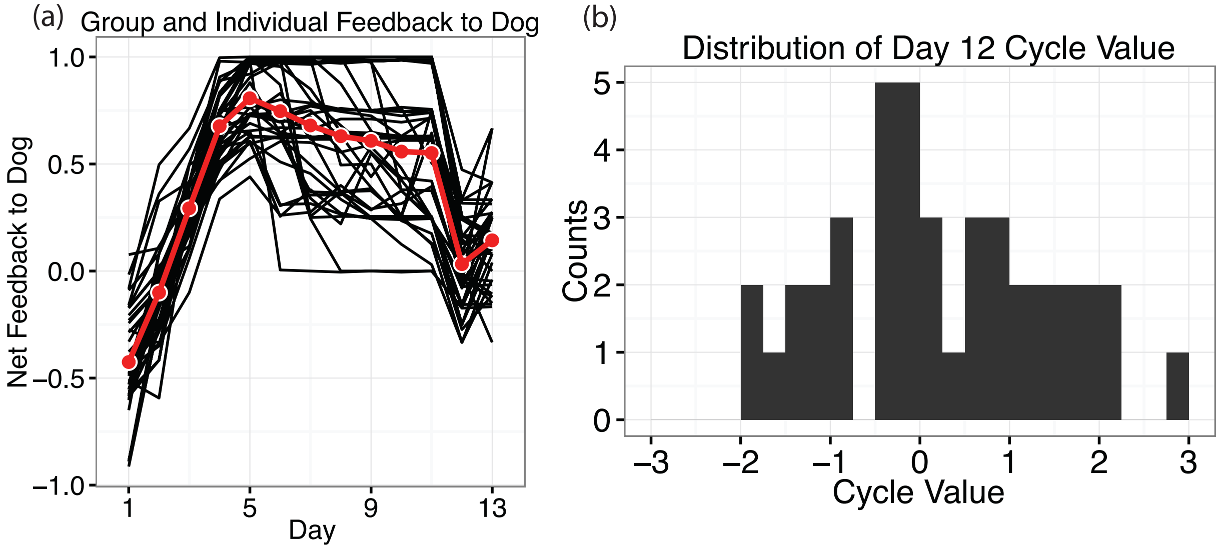
Procedure Participants were given the same instructions as in the previous experiment.

Results

Diminishing Rewards and Final Positive Cycle From days 5 to 11, during which the dog performed the desired behavior perfectly, participants' rewards diminished without disappearing completely. Net participant reward and day over these days were significantly negatively correlated (Average Correlation: -0.44; SE = 0.08; two-sided t-test: $t(40) = -5.32$, $p < .001$). However, an analysis of individual feedback behavior over time reveals start differences between participants. Figure 6a shows both the average net feedback per day and individual net feedback patterns over time. Out of the 41 participants, only 17 had significantly negative correlations ($p < .05$). Furthermore, as the graph indicates, only 3 of these 'decreasers' completely removed rewards over these 6 trials, although more may have done so if the learner had not regressed in trials 12 and 13.

Figure 6b shows the distribution of net cycle values on day 12, when the learner performed the 6 step cycle that was originally identified in Experiment 1. However, this time after 12 days, the average cycle value is no longer significantly positive (Mean Net Cycle Value = .19; SE = .19; one-sided t-test: $t(40) = 1.03$, $p = 0.15$). Nonetheless, around half of the participants in this study produced net positive cycles: 21 out of 41.

Figure 6: Experiment 2b results. (a) Net cycle values on Day 12. (b) Net feedback to dog by day (participant mean in red, individual participants in black).



Individual Differences Following the previous studies, I examined how previous dog experience, child experience, CRT proxy, gender, and political ideology interacted with two variables: decreasing rewards once the task had been learned, and day 12 cycle values. Means and comparisons between groups are summarized in Tables 3 and 4. The main factors that warrant discussion are dog experience and political ideology. Child experience, CRT proxy, and gender showed no significant patterns.

More dog experience was associated with using fewer rewards over time since life experience with dogs correlated with decreasing rewards (*r*(39) = -0.31, p = 0.05) and those with dog training experience decreased rewards more (Table 4). However, dog ownership did not impact either day 12 cycle values or tendency to decrease rewards.

Similar to Experiment 2a, a pattern emerged based on political ideology. Conservatives had fewer day 12 cycle values than Moderates or Liberals, and they also decreased their rewards the most (Tables 3 and 4). These two patterns were non-significant, however, they are consistent with earlier findings that suggest Conservatives attempt to use fewer rewards during teaching.

Table 3: Experiment 2b Individual Difference Comparisons of Day 12 Cycle Value

			Day 12		
Individual Difference		n	Cycle Value		
Dog Ownership	Dog	17	0.51 (.26)	*t*(37.3) = 1.52	*p* = 0.13
	No Dog	24	-0.04 (0.25)		
Dog Training Experience	At least some	16	-0.07 (0.23)	*t*(38.9) = -1.23	*p* = 0.23
	None	25	0.36 (0.27)		
Parenthood	Parent	7	0.37 (0.47)	*t*(8.45) = 0.42	*p* = 0.68
	Non-parent	34	0.15 (0.21)		
CRT proxy	$500	23	0.16 (0.23)	*t*(31.9) = -0.19	*p* = 0.85
	$1,000,000	18	0.23 (0.32)		
Gender	Female	13	0.51 (0.37)	*t*(20.18) = 1.1	*p* = 0.29
	Male	28	0.05 (0.21)		
Political Ideology	Conservative	10	0.06 (0.36)	*F*(2,38) = 0.12	*p* = 0.89
	Moderate	13	0.31 (0.40)		
	Liberal	18	0.18 (0.26)		

Table 4: Experiment 2b Individual Difference Comparisons of Decreasing Reward Measure (correlation between day and feedback for days 4 to 11)

Individual Difference		n	Reward Decrease Correlation		
Dog Ownership	Dog	17	-0.42 (0.13)	*t*(33.5) = 0.23	*p* = 0.81
	No Dog	24	-0.46 (0.11)		
Dog Training Experience	At least some	16	-0.74 (0.06)	*t*(34.7) = -3.67	*p* < .001
	None	25	-0.25 (0.12)		
Parenthood	Parent	7	-0.38 (0.19)	*t*(9.0) = 0.33	*p* = 0.75
	Non-parent	34	-0.45 (0.09)		
CRT proxy	$500	23	-0.47 (0.10)	*t*(33.9) = -0.43	*p* = 0.67
	$1,000,000	18	-0.40 (0.14)		
Gender	Female	13	-0.57 (0.12)	*t*(30.24) = -1.20	*p* = 0.24
	Male	28	-0.38 (0.11)		
Political Ideology	Conservative	10	-0.60 (0.15)	*F*(2,38) = 0.66	*p* = 0.52
	Moderate	13	-0.43 (0.14)		
	Liberal	18	-0.36 (0.14)		

Discussion

To reliably measure how people use rewards and punishments during and after successful teaching, participants in Experiments 2a and 2b were given learners that improved over time. This experimental set up revealed a number of important aspects about teaching using rewards and punishments. First, participants use rewards to teach the task and continue to reward learners long after the behavior has been mastered. Although some participants wean learners off of rewards as the experiment progresses, this occurs only after many trials, and many continue to produce net positive reward cycles that a reward-maximizing agent would exploit. Thus, the tendency to produce positive cycles is a robust phenomenon not limited to the case of giving feedback for isolated actions.

Second, people can reliably track the state of a learner's policy with respect to a target policy over time since ability and improvement judgments closely correlated with the probability of the learner selecting the correct action. This is an important component of the interaction model stated earlier.

Finally, the individual difference measures across the two studies suggest a possible pattern for dog owners and those with dog experience. Experiment 2a show that they produce large positive cycles once a task has been mastered, while Experiment 2b shows that they are more likely to decrease their rewards and then have fewer positive reward cycles once the dog has learned the task. Relatedly, political ideology appeared to show a consistent pattern of results: Conservatives give fewer rewards and decrease them once the task has been learned. This suggests that teaching behavior results from general social dispositions and are not specific to teaching. Definitive conclusions about individual differences, however, are difficult to draw since many comparisons were not found to be significant.

Experiments 3a b

Experiments 2a and 2b only test how teaching proceeds when everything goes as expected. A learner may regress for various reasons, such as misunderstanding the teacher's intention or simply being ill-suited to the teacher's strategy. When this occurs, teachers may or may not adapt their teaching strategy to the particular learner. For example, when faced with a reward-maximizing learner, people could either relentlessly continue using an action-feedback strategy or switch to the more appropriate one. Strategy switching will depend on if alternate strategies are spontaneously considered as well as if learning is transparent to the teacher.

Experiments 3a and 3b investigated how people trained *responsive learners* that learned using either reward-maximizing or action-feedback type algorithms. In the first experiment, learners were programmed to perform the current learned policy 80% of the time and explore 20% of the time. This tested whether people would spontaneously update their expectation to teach an action-feedback learner and switch to reward-maximization in a *noisy learning setting*. The second experiment tested whether people would change strategies in a *deterministic learning setting* where the agent's actions perfectly reflect their current policy.

Experiment 3a

Method

Participants and materials One-hundred and twenty Amazon Mechanical Turk workers participated in the experiment, but two were excluded due to missing data (61 Female, 1 Other, 55 Male). The same interface was used as in the previous experiments.

Participants were placed into one of 4 conditions: Q-learning, Model-based, Uniform Action-Feedback, or State-Reward Action-Feedback. All the algorithms were implemented in the webpage and operate as described in the first modeling section. The Uniform Action-Feedback model assumes a uniform distribution over all target policies π^*. The State-Reward Action-Feedback model assumes a uniform distribution over possible reward functions that assign +1, 0, or -1 to entering a given tile on the gridworld. Each of these reward functions has a unique optimal policy that maximizes that reward function, so the uniform distribution over reward functions produces a distribution over *policies* that the learner considers.

Procedure Participants were told they would train a dog over the course of 10 game days, each of which ends after 10 steps or once the dog gets to the door of the house. They were told the dog would learn the task as they gave it feedback and that the dog would appear at the beginning of the path on each day. The instructions indicated that a bonus was contingent on how well the dog performed on its own following the task (although everyone received the same bonus).

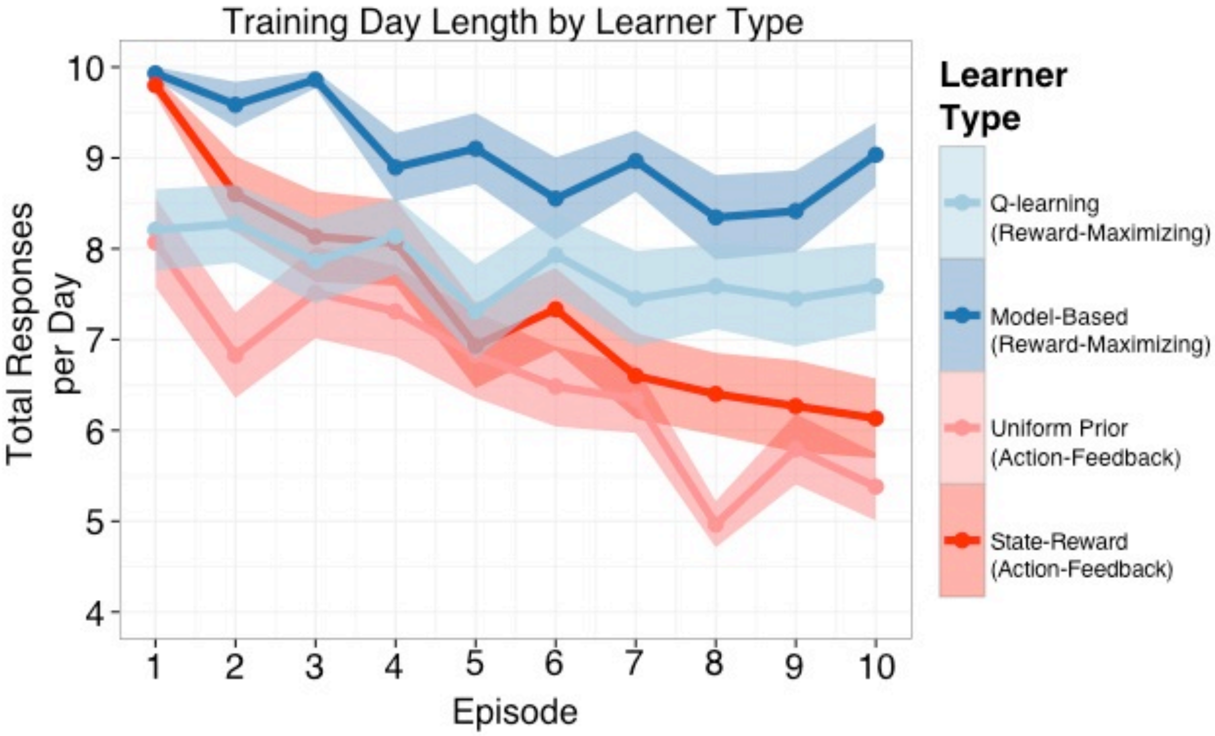
On each trial, the learner performed the action dictated by their current policy with a probability of .8 and chose one of the other actions with a probaiblity of .2/(# of actions – 1). Participants gave feedback consisting of one of 5 discrete options similar to the markings in the previous experiments (shocking, scolding, doing nothing, praising, and biscuits). Learners then updated their policies with the participant's response before performing the next action.

Following the task, several questions were asked, and participants were shown how their dog performed on the task by itself.

Results

Learner Performance over Time The total steps per day by day provides a coarse-level view of how teachers in the different conditioned performed. Figure 7 displays these results. Of particular interest is that although the reward-maximizing algorithms do improve somewhat, the action-feedback learners do significantly better by the end of the task. The steps per day on the last three days were non-normally distributed, so a Kruskal-Wallis rank sum test over the average length of the last three days was used. There was a significant difference between the conditions ($\chi^2(5) = 57.75$, $p < .001$). Bonferroni corrected pairwise comparisons using a Wilcox rank sum test revealed a significant difference between the Uniform Prior Action-Feedback and Model-based learners ($p < .001$), between the State-Reward Action-Feedback and Model-based learners ($p < .001$), and between the Q-learner and Uniform Prior Action-Feedback learner ($p < .001$).

Figure 7: Experiment 3a results. Each line represents the average number of steps in each day by the four learner type conditions.



Learner Behavior and Teacher Responses The previous analysis showed that people are better at training action-feedback learners than reward-maximizing learners. Further analysis of the learning dynamics reveals that this is due to participants' persistent use of an action-feedback strategy even when training reward-maximizing learners.

Over the course of the task, participants in all four conditions gave consistent levels of rewards and punishments for the lower-left, middle-left, upper-left, and upper-center actions (Figure 8). However, this tends to produce positive cycles for reward-maximizing learners. By design, the model-based learners learn the reward-maximizing policy on each turn and so 'chase after' the rewards given by the teachers in a goal-driven manner. The Q-learners, on the other hand, can only learn about positive cycles through trial and error by exploring certain portions of the grid. This results in some Q-learners performing the teacher's 4-step policy and others exploiting discovered positive cycles.

Positive cycle exploitation by learners can be seen in Figure 9, particularly in the graphs corresponding to the top-center and mid-center states. From the top-center state, both of the action-feedback learners enter the house state with the highest probability (80%) by the end of the task. In contrast, this action never comes to dominate the behavior of model-based learners, while Q-learners only perform this action around 50% of the time across the whole task. Some of these learners stay on the path but go left and then go right again. Others go down and then left onto the path. Both of these behaviors allow them to exploit the feedback given by the teacher.

Figure 8: Experiment 3a results. Teacher feedback by state (column) and action (colors) over time for the four algorithms (rows). The colors of the arrows roughly correspond to the 'optimality' of each action in each state.

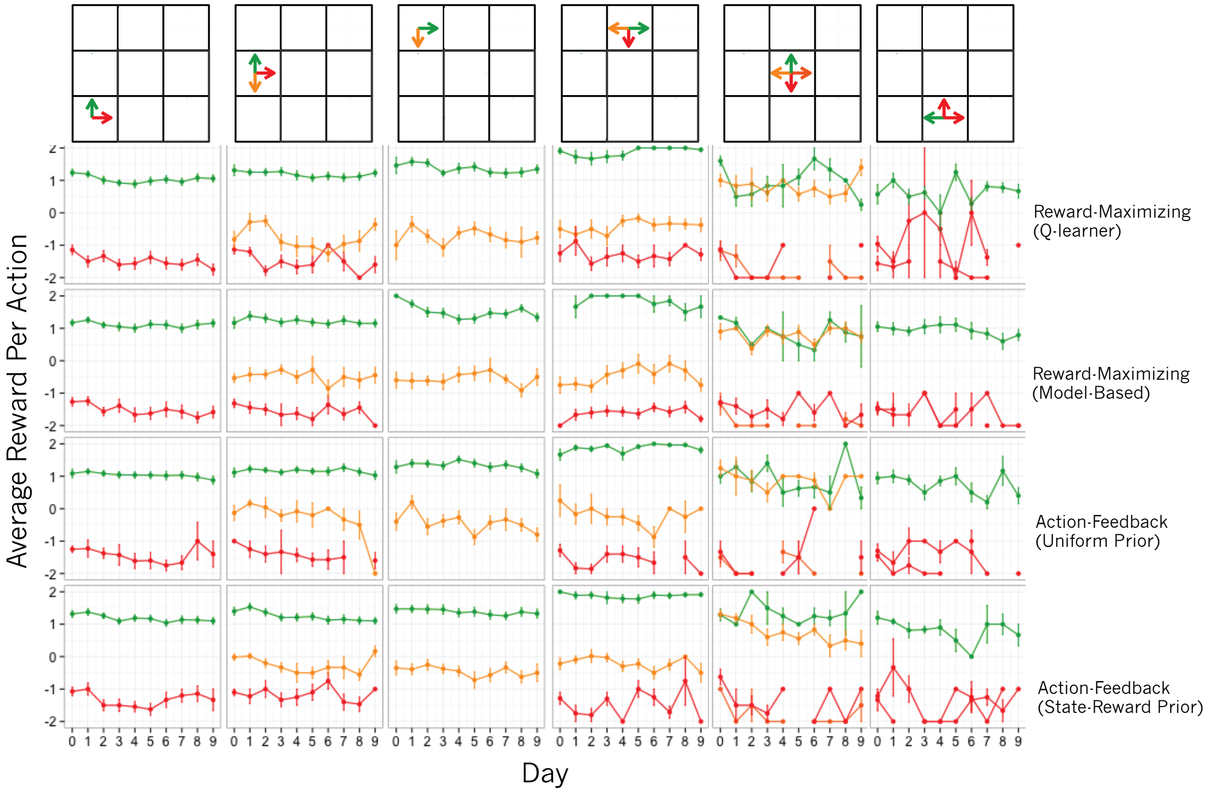
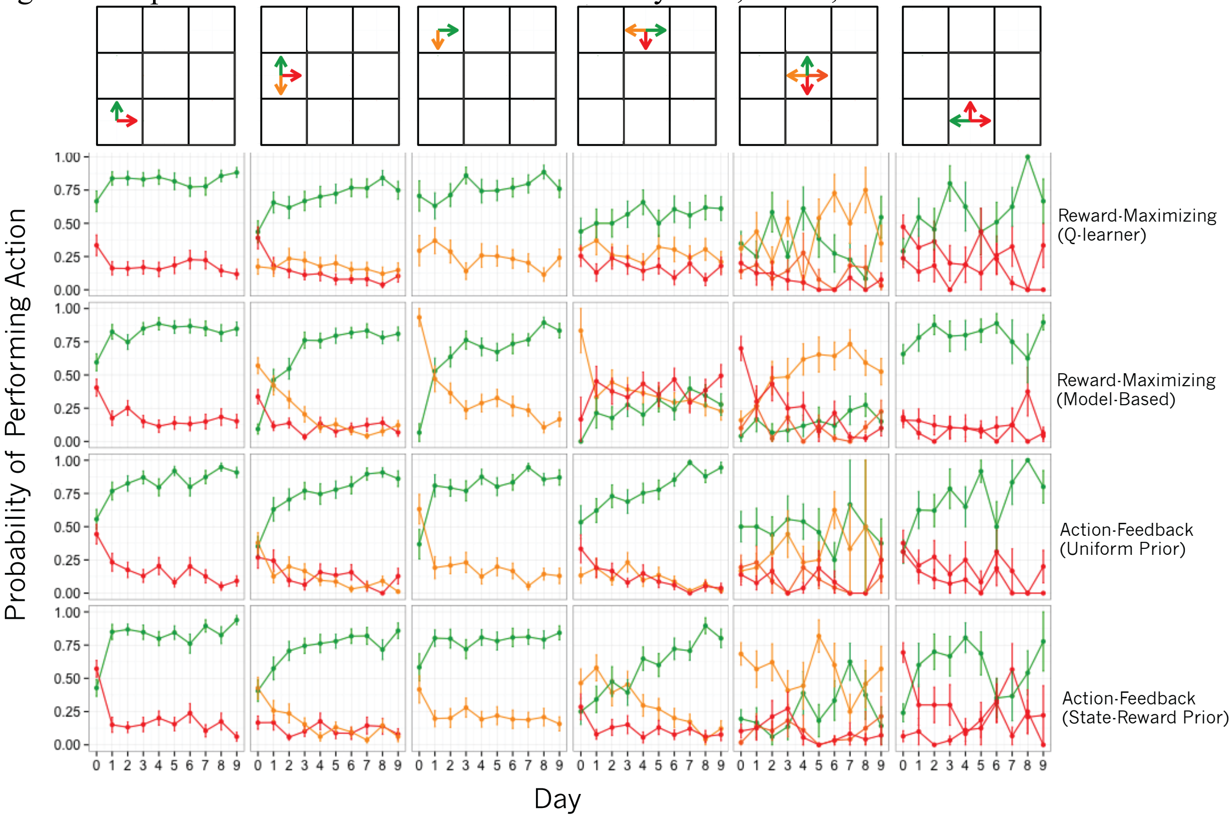


Figure 9: Experiment 3a results. Learner behavior by state, action, and condition over time.



Experiment 3b

Methods

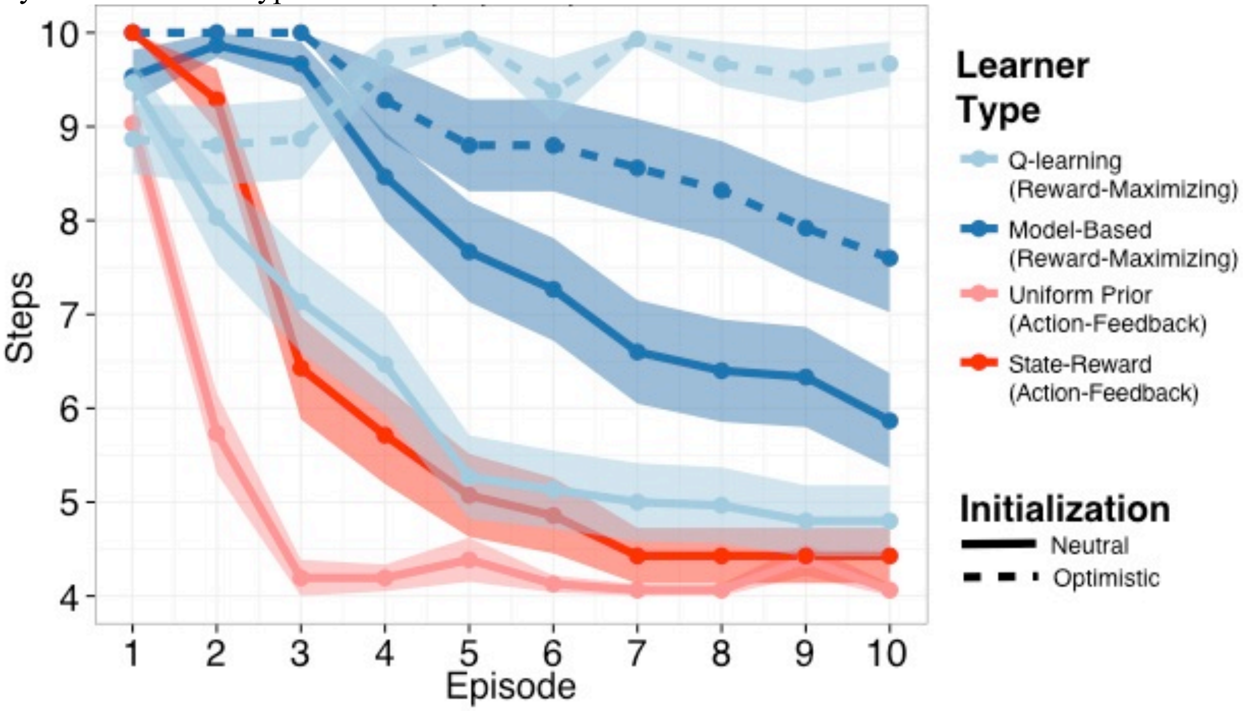
Participants and materials One-hundred and eighty Amazon Mechanical Turk workers participated in the study. Five were excluded due to technical errors. The materials used in this study were identical to the first, except in addition to the four learners used previously, two additional 'optimistic' reward-maximizing learners were included. The Optimistic Q-learner had all of its Q-values initialized to +2.5, while the optimistic model-based learner had its estimate of the reward function initialized to .5 for all actions. Additionally, agents in this study never deviated from their current learned policy and explored their environment.

Procedure The procedure was the same for this study as Experiment 3a. Participants read the instructions, answered comprehension questions, took the task, and answered several follow up questions.

Results

Learner Performance over Time As in the previous study, the Action-Feedback learners tended to perform much better than their reward-maximizing counterparts. Figure 10 shows the number of steps taken by the learner on each day of the task for each of the 6 conditions. People teaching the Uniform Action-Feedback learner succeed in the task the fastest since nearly all participants show perfect performance by the second day. The Neutral Q-learner and State-Reward Action-Feedback learner show the next fastest performance and begin to approach perfect performance by the last few days. Meanwhile, the two Model-based implementations show gradual improvement, with the Optimistic version doing the slowest. Finally, the Optimistic Q-learner shows almost no improvement. A Kruskal-Wallis rank sum test showed a significant difference in average steps the last 3 days ($\chi^2(5)$ = 87.27, p < .001). Pairwise comparisons showed that Optimistic Q-learners differed from Neutral Q-learners, Optimistic Model-based learners, and both Action-Feedback learners (all p < .001); and that Optimistic Model-based learners differed from Neutral Q-learners, and both Action-Feedback learners.

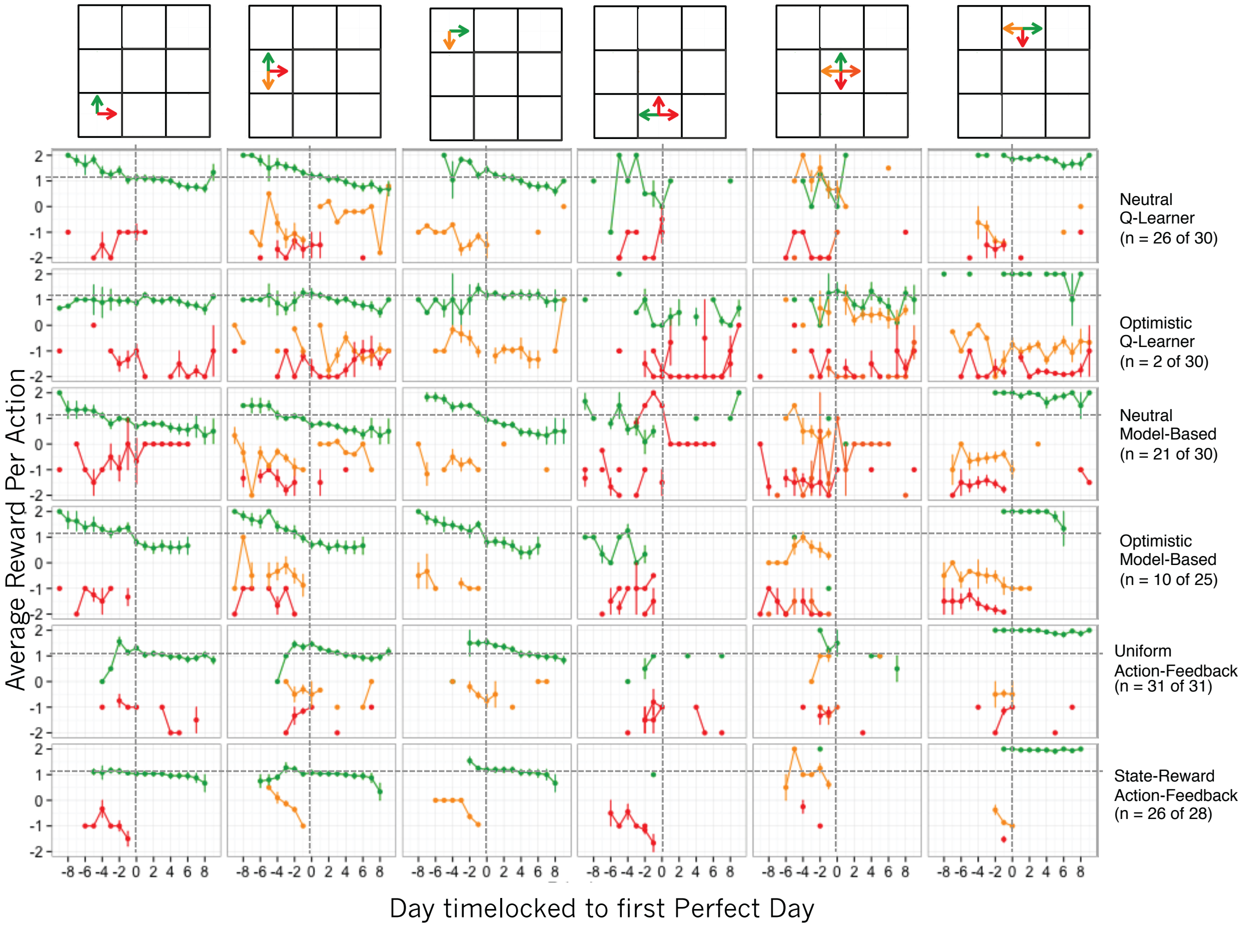
Figure 10: Experiment 3b results. Each line represents the average number of steps in each day by the four learner type conditions.



Changing Teacher Responses Improving model-based learner performance and participants' post-task feedback suggest that the teachers changed their feedback strategies during the task. A number of teachers explicitly mentioned that they had realized that the dog was trying to get more rewards from them, and so modified the magnitude of their rewards. To confirm that an improvement in performance was preceded by a change in teacher feedback strategy for reward-maximizing learners, I took participants who completed at least one day within 4 steps and looked at their responses timelocked to their first perfect trial. Figure 11 shows graphs for the different state-action pairs on days relative to the first time that the teacher completed the day in 4 steps.

As the graphs indicate, those training reward-maximizing agents (Q-learners and Model-based learners) decreased the rewards they gave for all the intermediate steps on the path. Average rewards for the left tiles fall enough that net positive cycles disappear from the feedback schedule, and the learner learns the task entirely. Thus, when learners behave purely deterministically and their policy is transparent to the teacher, teachers recognize that feedback is being treated as a reward signal to be maximized and switch strategies.

Figure 11: Experiment 3b results. Average rewards for each state-action pair over time. Days are timelocked to the first perfect trial that occurs in a participant's results. Only participants who had at least one perfect day are included.



Cognitive Style and Changing Teaching Strategies One individual difference measure that predicted whether people would switch strategies was the CRT proxy question. A χ^2-test for independence showed that the number of people who picked the Gamble option were more likely to train the reward-maximizing dogs ($\chi^2(1) = 5.56$, p = 0.018). This suggests that switching strategies results from a deliberative and active process of evaluating the action-feedback strategy.

Discussion

In two experiments, I have shown that people using rewards and punishments to teach responsive agents have a default expectation to be teaching Action-Feedback learners who recognize communicative intent. Furthermore, unless the behavior of a reward-maximizing agent involves no exploration and perfect responsiveness to the teacher, teachers will persist in their default strategy. This is important because it suggests that teaching reward-maximizing learners in natural settings will be difficult for human adults like the participants in these studies. It is not implausible to think dogs are likely to perform random behaviors and/or view the world optimistically. Alternatively, consider a child who is not only processing feedback from her parent, but also exploring her environment and figuring out whether the cookie jar is empty or what household objects break.

General Discussion

The three sets of studies discussed here cover teaching by evaluative feedback when responding to individual actions (Experiment 1), responding to a single agent that improves over time (Experiments 2a and 2b), and interacting with an agent that learns from rewards (Experiments 3a and 3b). The results of these studies clearly suggest that people are naturally inclined to expect a learner to recognize that rewards and punishments have communicative intent. That is, people deliver feedback as if they are training an action-feedback learner rather than incentivizing behavior for a reward-maximizing learner. This important distinction results in teachers providing patterns of feedback that could be exploited by reward-maximizing learners as shown in Experiment 3a.

These results also shed light on teaching before, during, and after a learner has acquired the desired behavior. When teaching, people are reluctant to withdraw rewards before the task is learned (Experiment 3a), unless the learner clearly indicates that they are exploiting positive cycles (Experiment 3b). Even once an agent has learned a task, people often persist in giving some rewards for an extended amount of time and rarely remove rewards all together (Experiment 2b). Thus, while in the default mode of teaching action-feedback learners, people are fairly consistent in the feedback that they will give to a learner.

Communicative intent has been studied in the context of concept learning (Gergely & Csibra, 2009; Shafto et al., 2014) as well as with language pragmatics (Frank & Goodman, 2012). But to my knowledge, there has been limited research on this topic as it relates to learning from and teaching with rewards and punishments. Given that rewards and punishments, by definition, are stimuli in the environment that an agent places some intrinsic value upon, they must play an essential role in social interactions like teaching. Its hard to imagine a society of organisms that interact without causing or intending good or bad things to happen to one another. Similarly, it is hard to imagine a parent who never uses good and bad consequences to discipline their child. The work here starts to fill this gap in psychological research between learning from valenced stimuli and teaching with mutually recognized communicative intent.

Additionally, these results reveal important dimensions of variability in peoples' teaching behavior by looking at the individual level. Experiment 1, for instance, showed that people have distinct training strategies corresponding to action-feedback and state-training. Meanwhile, individual feedback patterns in Experiment 2b reveal that some people systematically decrease the rewards that they give over time once the task is learned, while others maintain a constant level of reward. Finally, some participants in Experiment 3b recognized that they were training a reward-maximizing agent, while others persisted in giving feedback to a non-existent action-feedback learner. Some of these teaching behaviors are potentially related to more stable features of a teacher (e.g. decreasing rewards and dog experience, recognizing reward-maximization and cognitive style, avoiding positive cycles and political conservativism), but many of them simply reflect idiosyncraies of the teacher.

Previous work on developing artificial agents that learn from feedback has shown that a sensitivity teacher strategies improves performance. For example, Loftin et al. (2014) developed an algorithm that estimated a person's tendency to reward or not reward correctness and punish or not punish incorrectness. These agents simultaneously learned a task while learning how the teacher communicated. Along similar lines, the dimensions of variability identified in these studies could be used to develop more user-friendly and user-tailored systems that learn from human-delivered feedback. Researchers could design systems that adopt a learning style customized to a person's teaching style, rather than use a one-size-fits-all approach.

Additionally, the three different paradigms developed here can be easily modified to search for additional dimensions of teaching behavior and interaction. For example, how will people detect and respond to systematic behavioral errors like overgeneralization (e.g. washing dishes in the bathroom sink)? How will the average user respond? Will users tend to cluster into different types of responses? These dimensions are not limited to giving rewards or punishments either. For instance, a user might pick up the robot and move it to the kitchen. An agent that treated this as a regular transition might learn that washing dishes in the bathroom sink teleports it to the kitchen; an agent that interpreted it as a pedagogical state transition would learn that using the bathroom sink was probably a mistake.

This thesis synthesized insights from psychology and reinforcement learning to study how people use rewards and punishments in teaching. It is clear that people have strong default biases to train with the presumption of communicative intent, but that they are also adaptive and vary in specific ways. Future research should explore how people use rewards and punishments to better understand this aspect of social interaction and design artificial agents.

Acknowledgements

The author would like to thank Michael L. Littman, Fiery Cushman, and Joseph Austerweil for their valuable insight and assistance while collaborating on this project.

Citations

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bellman, R. (1957). *A Markovian decision process*. DTIC Document.

Brafman, R. I., & Tennenholtz, M. (2003). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, *3*, 213–231.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*(11), 681.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Loftin, R., MacGlashan, J., Peng, B., Taylor, M. E., Littman, M. L., Huang, J., & Roberts, D. L. (2014). A strategy-aware technique for learning behaviors from discrete human feedback. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI-2014)*.

Malle, B. F. (2005). Folk theory of mind: Conceptual foundations of human social cognition. *The New Unconscious*, 225–255.

Ng, A. Y., Harada, D., & Russell, S. J. (1999). Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 278–287). Morgan Kaufmann Publishers Inc.

Owen, D. J., Slep, A. M., & Heyman, R. E. (2012). The effect of praise, positive nonverbal response, reprimand, and negative nonverbal response on child compliance: a systematic review. *Clinical Child and Family Psychology Review*, *15*(4), 364–385.

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.

Skinner, B.F. (1938). *The behavior of organisms: an experimental analysis*. Oxford, England: Appleton-Century.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Cambridge, Massachusetts: Harvard University Press.

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.