Teaching with Rewards and Punishments: Reinforcement or Communication?

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

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

Fiery Cushman (cushman@fas.harvard.edu)
Department of Psychology, 1484 William James Hall, 33 Kirkland St.
Cambridge, MA 02138 USA

Joseph L. Austerweil (joseph_austerweil@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, 190 Thayer St.
Providence, RI 02912 USA

Abstract

Teaching with evaluative feedback involves expectations about how a learner will interpret rewards and punishments. We formalize two hypotheses of how a teacher implicitly expects a learner to interpret feedback – a *reward-maximizing model* based on standard reinforcement learning and an *action-feedback model* based on research on communicative intent – and describe a virtual animal-training task that distinguishes the two. The results of two experiments in which people gave learners feedback for isolated actions (Exp. 1) or while learning over time (Exp. 2) support the action-feedback model over the reward-maximizing model.

Keywords: pedagogy; reward; punishment; reinforcement learning; feedback; evaluative feedback; communication

Introduction

Imagine Eve, a 4-year-old toddler, who uses the toilet for the first time. Her proud parents might give her a hug and some stickers for the accomplishment. Or consider Fido, the chocolate labrador puppy, who ignores the paved walkway leading to the house and tramples over a freshly planted bed of flowers. Fido's owner, who spent the last two months tending to his lawn, scolds Fido harshly in a firm and imposing voice. In both cases, a teacher (Eve's parents, Fido's owner) attempts to modify another agent's behavior (a child or a dog) using valenced stimuli (stickers, scolding). This type of interaction – teaching via evaluative feedback – occurs frequently between parents and their children (Owen et al., 2012) as well as between humans and other species such as dogs.

Here, we explore how teachers implicitly expect a learner to interpret rewards and punishments intended to modify behavior. That is, we examine how teachers provide evaluative feedback in response to the actions of a learner.

Reinforcement learning models of human and animal learning based on operant conditioning (Sutton & Barto, 1998; Dayan & Niv, 2008) assume that learners are *reward-maximizing*. Teachers might share this assumption, namely, that evaluative feedback will be treated as face value rewards and punishments. Positive responses are pleasurable, rewarding outcomes of behavior to be maximized, while negative responses are painful, punishing outcomes to be minimized. The learner is expected to interpret feedback like any other valenced stimulus that results from acting on the environment, such as a ripe apple having fallen from a shaken branch or a burnt finger having touched a hot stove. On this view, when Eve's parents want to teach her about using the toilet and not the living room, they intend the sticker to serve as an incentive. "Eve loves stickers," they reason, "so she will want to use the toilet again".

In contrast, peoples' understanding of communicative intent when learning new concepts from examples (Sperber & Wilson, 1986; Csibra & Gergely, 2009; Shafto et al., 2014) suggests an *action-feedback* model in which teachers expect learners to treat responses communicatively or as commentary about an action. Rewards signal to the learner that the action just 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. From this perspective, when toilet training Eve, her parents intend the sticker to serve as a signal that she is doing the right thing. "Eve knows we don't give stickers out for nothing," they might reason, "so she'll learn that she should be using the toilet."

In this paper, we first formalize these two hypotheses of teaching via evaluative feedback in the framework of Markov Decision Processes. Teachers who teach according to a *reward-maximizing model* expect learners to treat positive feedback as desirable rewards to be maximized and negative feedback as undesirable punishments to be minimized. In contrast, those who teach based on an *action-feedback model* expect learners to treat positive and negative stimuli as signals for correct or incorrect behavior. We then describe a novel teaching task that qualitatively distinguishes these two models. Finally, we present results from two experiments that show the majority of people do not teach in accord with a reward-maximizing account, but instead broadly follow the predictions of the action-feedback model.

Model

We first describe an *interaction model* of the teacher-learner dynamics during teaching with evaluative feedback. Second, we propose two *learner models* (reward-maximizing and action-feedback) that capture a teacher's expectation of how the learner interprets feedback. Finally, we show how the models can be distinguished in a novel sequential-teaching paradigm.

Interaction Model

To model the interaction between a teacher and a learner, we use the Markov Decision Process formalism (Bellman, 1957). On each timestep t, the teacher observes a learner interacting with an environment composed of states ($s_t \in \mathcal{S}$) by performing any one of the actions available in a state ($a_t \in \mathcal{A}(s_t)$). Each action is generated from a learner's behavioral repertoire at a given time step, represented as a policy, π_t, where π_t is a mapping from states to available actions ($\pi: s \rightarrow a \in \mathcal{A}(s)$).[1]

After observing the agent's current state, action, and subsequent state (s_t, a_t, s_{t+1}), the teacher responds to the learner with a positive or negative feedback signal of a finite magnitude ($f_t \in [-1,1]$). The function that takes an observation of the learner and returns feedback we call the *feedback* function:

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

The pattern of rewards and punishments that constitute this feedback function is determined by the *target policy* (π^*) that the teacher wants the learner to acquire. The interaction then continues into the next timestep (Figure 1).

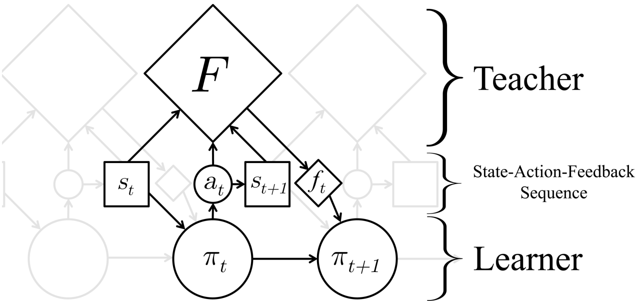

Figure 1: Teacher-learner interaction model. The learner's current policy, π_t, takes in the current state s_t and returns an action a_t, leading to the next state s_{t+1}. A *feedback function, F,* observes this and gives feedback f_t to the learner, resulting in the modified policy π_{t+1}.

Learner Models

Learning consists of changes in an agent's policy over time, ($\pi_0, \pi_1, \ldots, \pi_{T-1}, \pi_T$), resulting in a *learned policy*, π_T. When teaching by evaluative feedback, a teacher expects a learner to learn from the feedback signal produced by F. Thus, each of our models characterizes the functional relationship between a learned policy, π_T, and feedback function, F.

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:

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

Importantly, future rewards may be treated as less rewarding than immediate ones, so we include a discount parameter $0 \leq \gamma \leq 1$. 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:

$$\pi_{\mathrm{RM}}(s) = \operatorname*{argmax}_{a \in \mathcal{A}(s)} \max_{\pi} q_\pi(s,a)$$

for all $s \in \mathcal{S}$. Note that the model is agnostic about the precise learning mechanism that updates π_t as long as it converges on the reward-maximizing policy π_{RM}. For instance, π_{RM} could be learned via explicit planning or trial-and-error learning – e.g. model-based learning vs. model-free Q-learning (Dayan & Niv, 2008).

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. 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:

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

Then for all $s \in \mathcal{S}$, the action-feedback agent will learn the policy:

$$\pi_{\mathrm{AF}}(s) = \operatorname*{argmax}_{a \in \mathcal{A}(s)} j(s,a).$$

[1] For simplicity, we assume that state transitions and policies are deterministic (e.g. $P(s_{t+1}|s_t, a_t) = 1$), however, this can be generalized to stochastic environments and policies.

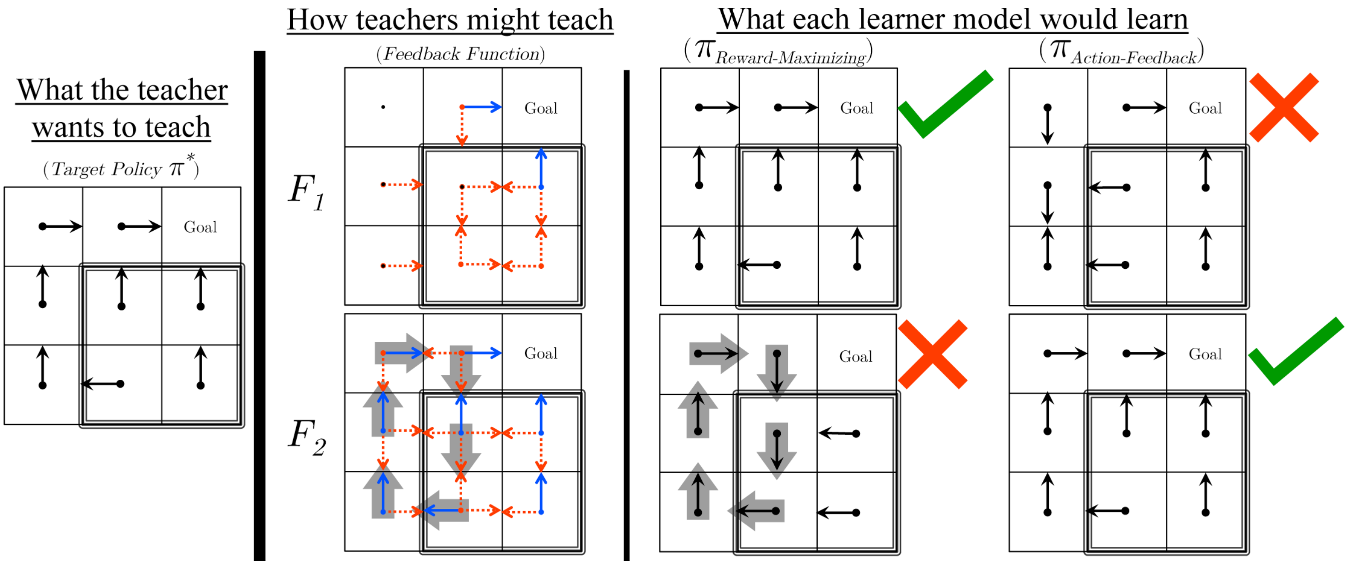
What the teacher wants to teach
(Target Policy π^*)
How teachers might teach
(Feedback Function)
What each learner model would learn
($\pi_{Reward\text{-}Maximizing}$)
($\pi_{Action\text{-}Feedback}$)
F_1
F_2
Goal

Figure 2: 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.

How do teachers teach?

When do the two models diverge? That is, when does $\pi_{RM} \neq \pi_{\text{AF}}$ 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 $\pi_{\text{RM}} = \pi_{\text{AF}}$. 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 2 illustrates the predicament of Fido's owner in a simplified gridworld.

We designed a dog-training paradigm, the Garden-Path task, reminiscent of the one faced by Fido's owner (Figure 3) 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). Experiment 1 investigated peoples' teaching patterns for isolated actions taken by a learner. Experiment 2 investigated how people teach a single learning agent over time.

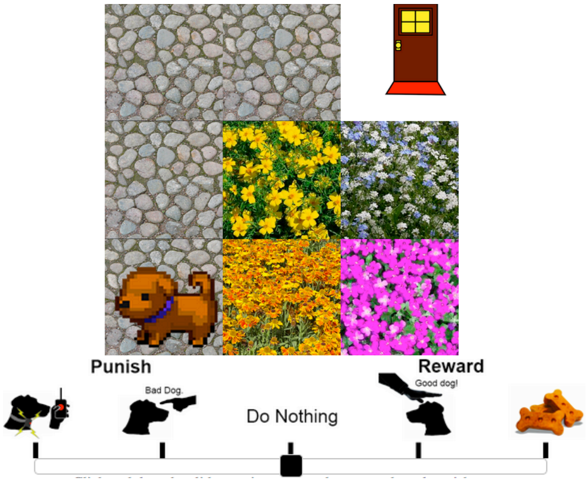

Figure 3: Fido's Garden-Path task. On each trial, a dog moves and then participants give their feedback. (Demo at: http://research.clps.brown.edu/mkho/gardenpath/task.html)

Experiment 1: Teaching Isolated Actions

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

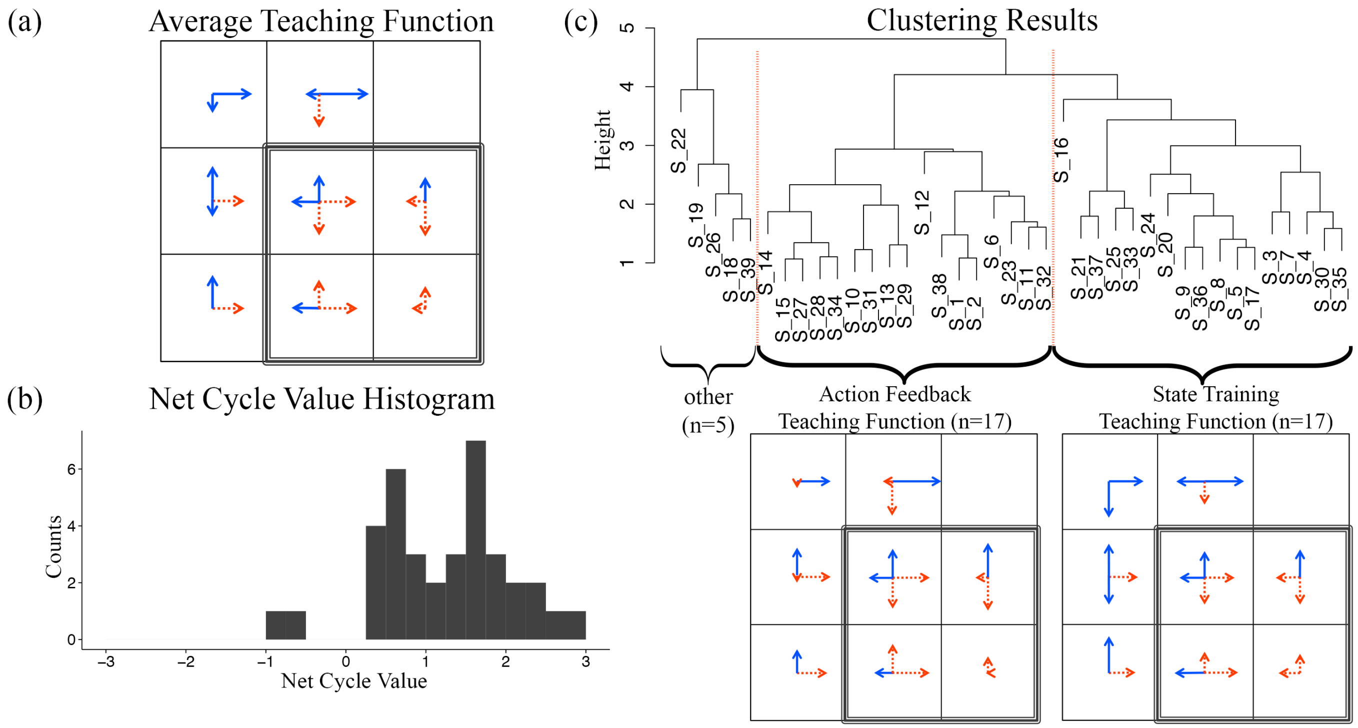

Figure 4: Results of Experiment 1. (a) Average teaching function of all participants. (b) Net value of responses on cycle trials by individual. (c) Results of hierarchical clustering of participants' responses with the average teaching function of the two largest clusters. These correspond to an action-feedback function and a "state-training" function (see text).

Method

Participants and materials Thirty-nine people from MTurk participated. 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 We told participants 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. We told participants 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.

To evaluate participants' perceptions of whether the dogs value future rewards, we asked several questions about dog preferences after the task. For example, one question asked whether a dog would prefer 2 scoldings followed by 4 praises or prefer receiving nothing at all.

Results

Positive Cycles We first analyzed whether participants' stationary feedback functions had positive cycles that could be discovered by a reward-maximizing learner. Figure 4a 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, we 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 4c) 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 of feedback.

Feedback Value 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. Additional questions confirmed that the scale itself was interpreted symmetrically, however, we will not discuss them due to space limitations.

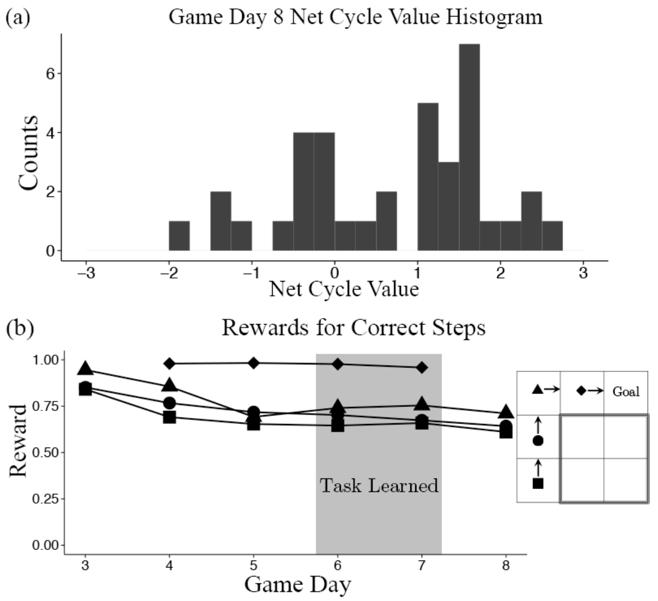

Figure 5: Experiment 2 (a) Responses on final game day still have positive cycles (b) Rewards for correct actions decrease but stay positive even when the learner is trained (days 6-7).

Experiment 2: Teaching Over Time

Experiment 1 examined responses to isolated actions. Experiment 2 had participants teach a single dog over time. This allows us to test whether positive cycles still arise and whether teachers can properly track a learner's policy.

Method

Participants and materials The same interface was used. Thirty-seven people trained 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. We showed all participants the same pre-determined set of actions.

Procedure We told participants that they would train a single dog over the course of 8 game days and that at the end of the experiment, we would test the dog, on its own, 3 times at the beginning of the path. A bonus was contingent on the dog's performance (but everyone won). 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). After the task, we asked participants about the responsiveness of the dog to feedback on a 1-5 scale. All other details of the task, including post-task questions, were otherwise the same as in Experiment 1.

Results

All participants rated dog responsiveness above 1 = not responsive at all (mean=3.45, SE=.11). Additionally, preference judgments were similar to those 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).

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 our interaction model assumes.

Discussion

Teachers often use reward and punishment to modify the behavior of other agents such as children and animals. In two experiments, we examined how teachers expect learners to interpret evaluative feedback. Our results demonstrate that when giving feedback for isolated actions (Exp. 1) and when training a single learner over time (Exp. 2), people's patterns of reward and punishment produce *positive cycles*. That is, people deliver rewards in a manner that a reward-maximizing agent (of the variety found in standard RL) would discover and capitalize on (Dayan & Niv, 2008; Sutton & Barto, 1998).

These results can be explained by the action-feedback model, which is based on work on communicative intent (Sperber & Wilson, 1986; Csibra & Gergely, 2009; Shafto et al. 2014).[2] On this view, teacher feedback is not just a face value reward but instead a signal about an action's correctness. This allows an action-feedback learner to learn the desired policy even in the presence of positive cycles.

At the same time, our current action-feedback model does not completely account for teachers' communicative use of rewards and punishments. For instance, it does not entirely explain state-training teaching functions (Exp. 1) or diminishing rewards (Exp. 2). State training could reflect teachers' attempts to teach intermediate policies, while diminishing rewards may involve consideration of the history of the learner. Our model could be extended to include teacher's inferences about what the learner has learned so far.

Relatedly, the domains considered here are simple gridworlds, and we do not assume that learners generalize information learned about one tile token to another of the same type (e.g. two path tiles). The presence of state-training suggests teachers may not make this assumption and instead expect shared knowledge of state types.

Additionally, we mainly looked at teacher expectations in light of learning *outcomes* and bracketed the question of how specific learning *mechanisms* interact with patterns of feedback online. Since these studies deliberately hold learner behavior constant, we will compare how different teaching strategies interact with different algorithms (e.g. model-based RL, model-free RL, or online action-feedback) during learning.

More broadly, if human teachers do not naturally expect to interact with reward-maximizers but rather something akin to an action-feedback learner, one question is whether human learners (or other agents) meet those expectations. If so, this may suggest that rewards and punishments delivered communicatively from another agent are processed distinctly from those delivered otherwise or from the environment. Future work should investigate mechanisms of teaching with *and* learning from reward and punishment, as well as their interaction.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-1058262).

References

Bellman, R. E. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics, 6*, 679–684.

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.

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. *Proceedings of the 28th AAAI conference on artificial intelligence.*

Ng, A. Y., Harada, D., & Russell, S. J. (1999). Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. *Proceedings of the 16th international conference on machine learning* (pp. 278-287).

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.

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

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

[2] Note that whereas Shafto et al. (2014) look at using examples to teach concepts, we examine using feedback to teach behavior.