American Psychological Association

Journal of Experimental Psychology: General

0096-3445/19/$12.00

2019, Vol. 148, No. 3, 520–549
http://dx.doi.org/10.1037/xge0000569

People Teach With Rewards and Punishments as Communication, Not Reinforcements

Mark K. Ho
Brown University

Fiery Cushman
Harvard University

Michael L. Littman
Brown University

Joseph L. Austerweil
University of Wisconsin-Madison

Carrots and sticks motivate behavior, and people can teach new behaviors to other organisms, such as children or nonhuman animals, by tapping into their reward learning mechanisms. But how people teach with reward and punishment depends on their expectations about the learner. We examine how people teach using reward and punishment by contrasting two hypotheses. The first is evaluative feedback as reinforcement, where rewards and punishments are used to shape learner behavior through reinforcement learning mechanisms. The second is evaluative feedback as communication, where rewards and punishments are used to signal target behavior to a learning agent reasoning about a teacher's pedagogical goals. We present formalizations of learning from these 2 teaching strategies based on computational frameworks for reinforcement learning. Our analysis based on these models motivates a simple interactive teaching paradigm that distinguishes between the two teaching hypotheses. Across 3 sets of experiments, we find that people are strongly biased to use evaluative feedback communicatively rather than as reinforcement.

Keywords: pedagogy, reward, punishment, reinforcement learning, communication

Consider Alex, a 2-year-old who uses the toilet for the first time. His parents are thrilled and reward him with a hug and some candy. Or imagine Fido, a dog whose owner enjoys gardening. One day Fido tramples on his owner's flowerbeds. To prevent this from happening again, Fido's owner buys an electric collar that allows him to deliver a mild shock whenever Fido heads toward the garden. These situations involve teaching with evaluative feedback: instances in which people use rewards and punishments (e.g., candy and electric shocks) to teach another agent (e.g., a child or a dog) a behavior (e.g., using a toilet or staying out of the garden).

Mark K. Ho, Department of Cognitive, Linguistics, and Psychological Sciences, Brown University; Fiery Cushman, Department of Psychology, Harvard University; Michael L. Littman, Department of Computer Science, Brown University; Joseph L. Austerweil, Department of Psychology, University of Wisconsin-Madison.

Early versions of some data and ideas contained in this article were presented at the 37th Annual Conference of the Cognitive Science Society in 2015 as well as the 2nd Reinforcement Learning and Decision-Making Conference in 2015.

Mark K. Ho was supported by the NSF GRF under Grant DGE-1058262. Fiery Cushman was supported by grant N00014-19-1-2025 from the Office of Naval Research and grant 61061 from the John Templeton Foundation. Michael L. Littman received support from the NSF. Joseph L. Austerweil was supported by the Office of the CVGRE at UW-Madison with funding from the WARF.

Correspondence concerning this article should be addressed to Mark K. Ho, who is now at the Department of Psychology, Princeton University, Princeton, NJ 08540. E-mail: mho@princeton.edu

People readily use rewards and punishments to teach children (Owen, Slep, & Heyman, 2012), adults (Fehr & Gächter, 2002), pets (Hiby, Rooney, & Bradshaw, 2004), and, more recently, robots (Isbell, Shelton, Kearns, Singh, & Stone, 2001; Knox & Stone, 2015; Loftin et al., 2014). But how do people teach with evaluative feedback?

One possibility is that people treat teaching with rewards and punishments simply as the inverse of learning from rewards and punishments. Reward learning has been studied for more than a century, building on foundational accounts such as the law of effect and operant conditioning (Rotter, 1966; Skinner, 1938; Thorndike, 1898). Contemporary theories of causal learning, cognitive control, and reward-based decision-making all begin with the basic idea that organisms learn to take actions that make good outcomes more likely and bad outcomes less likely (Collins & Frank, 2013; Dayan & Niv, 2008; Gershman, Blei, & Niv, 2010). Put somewhat differently, agents learn by adapting their thoughts and actions to maximize environmental rewards while minimizing punishments (Sutton & Barto, 1998). Possibly, then, teaching is designed to provide a set of primary rewards and punishments that, when maximized by a reward-learning agent, shape their behavior to match some desired target.

This view is widespread in psychology. Classic theories of socialization draw on this idea (Aronfreed, 1968; Grusec & Kuczynski, 1997; Maccoby, 1992; Sears, Whiting, Nowlis, & Sears, 1953) and continue to inform our understanding of how parents teach children (Owen et al., 2012). Moreover, behavioral and neuroscientific accounts of social rewards are based on related assumptions about reinforcement and utility maximization in

games (Fehr & Gächter, 2002; Izuma, Saito, & Sadato, 2008; Jones et al., 2011; Lin, Adolphs, & Rangel, 2012) as are predictions of teaching in nonhuman animals (Caro & Hauser, 1992; Clutton-Brock & Parker, 1995; Kline, 2015). Given this account's roots in theories of operant conditioning, we call this hypothesis *evaluative feedback as reinforcement*.

At the same time, many researchers over the past few decades have focused on the centrality of mental state inference and recognition of pedagogical intent when learning from social others. For instance, theory of mind representations underlie children's abilities to successfully learn about hidden causal structure from behavior (Lyons, Young, & Keil, 2007), imitate intentions based on failed actions (Meltzoff, 1995), and integrate information about an actor's constraints (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015). Moreover, across a range of domains, it has been shown that social learning is enhanced when teachers signal their intention to teach via ostensive or pedagogical cues (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Butler & Markman, 2012, 2014; Csibra & Gergely, 2009; Sage & Baldwin, 2011; Shafto, Goodman, & Griffiths, 2014). These results provide clear evidence of deep connections between theory of mind, communicative intent, and social learning.

Could a learner's theory of mind also play a role in teaching with evaluative feedback? Specifically, might teachers employ rewards and punishments designed not as a quantity to be maximized, but instead as signs to be interpreted in light of teaching goals? If so, they would expect learners to infer the communicative intent underlying the set of rewards and punishments that teachers provide. For example, when Alex's parents give him candy for using the toilet, they could expect him to reason "Using the toilet must be correct since my parents gave me candy" rather than "Using the toilet is a good way to get candy (i.e., reward)." In other words, rather than shape behavior through reinforcement, feedback could signal the correctness or incorrectness of behaviors with respect to a teacher's target behavior. We call this alternative account *evaluative feedback as communication*.

Our goal is to ask whether people use evaluative feedback as reinforcement or as communication. To make this problem empirically tractable we exploit a phenomenon called *positive reward cycles* (Ng, Harada, & Russell, 1999). In this context, positive reward cycles are patterns of rewards and punishment that can signal the correctness of individual actions to a learner who interprets them as *signs*, and yet would lead a *reward-maximizing* agent astray. If we find that human teachers spontaneously generate positive reward cycles, this would imply that their evaluative feedback is designed for communication. If we find that they avoid such cycles, this would be more consistent with their evaluative feedback being designed for reinforcement.

To illustrate the basic form of a positive reward cycle, suppose Fido's owner wants to teach Fido to go into the house by walking along a path while avoiding flowerbeds. To do so, she gives Fido a biscuit every time he moves toward the house along the desired path, hoping to signal that these are the right actions to perform. If Fido learns from rewards in this way—inferring that they are intended to communicate the right actions to perform—eventually he will learn the complete task. But if Fido simply wants to maximize rewards, he will instead learn to "cycle" back and forth along the path endlessly, never entering the house. This is the reward maximizing response because the nonrewarded actions away from the door position him to gain new rewarded actions when moving back toward the door. This repetitive behavior by a reward-maximizing learner is a consequence of positive reward cycles in a teacher's pattern of feedback. If Fido learns this way, then the owner's best recourse is to break the cycle, either by rewarding Fido only when he makes it in the house, or by punishing him for running backward to a sufficient degree that it offsets the rewards of moving forward.

To characterize whether teachers generate positive reward cycles, we develop formal models of the two learning methods in question: learning via reinforcement and learning via communication. These models are described in the General Methods section. To be clear, these are not computational models of our human participants, who act as teachers in our experiments. Rather, they are models that formalize how two different classes of learners would respond to the rewards and punishments that our human participants generate. Formally specifying learning allows us to pin down the assumptions and constraints of learning from reinforcement and learning from communication. This, in turn, allows us to ask whether human teaching behavior succeeds or fails to meet the qualitative demands of each class of learners (Palminteri, Wyart, & Koechlin, 2017). Moreover, formalizing these learning processes motivates a novel teaching paradigm that generates diverging predictions for evaluative feedback as reinforcement versus communication.

Using our paradigm, we conducted three sets of human experiments. Experiment 1 investigated whether people produced positive cycles when delivering feedback for individual actions taken by virtual dogs. The studies in Experiment 2 looked at how participants teach a single dog preprogrammed to improve over time. This enables us to determine whether people produce positive cycles across multiple stages of successful training. The studies in Experiment 3 had participants interactively teach different learning algorithms that treated feedback as either reinforcement or communicative signals. This allowed us to test the effectiveness of people's teaching strategies as well as whether they adapted their strategies to different learning mechanisms. Across all of these studies, results indicate that people have a strong bias to use rewards and punishments as communicative signals rather than as reinforcement.

General Methods

In this section, we formalize teacher–learner interaction dynamics, learning from reinforcement (the reward-maximizing model), and learning from communication (the action-signaling model). In addition, we note that our goal is to characterize the assumptions of two broad classes of models and not necessarily to provide fine-grained predictions of teacher behavior. By formally analyzing the general properties of the reward-maximizing and action-signaling models, we can isolate specific qualitative predictions that rule out one or the other class of models. In particular, we show how certain patterns of feedback give rise to positive reward cycles, which should not occur if people are teaching with reinforcement but may occur if people are teaching with communication. This analysis motivates the design of the Path-Following teaching paradigm that distinguishes between the two teaching accounts.

Modeling Teacher–Learner Dynamics

The interaction between a teacher and learner can be modeled as a modified Markov game (Littman, 1994) that includes a set of shared world states, $\mathcal{S}$; a set of learner actions, $\mathcal{A}^{\mathcal{L}}$; a set of teacher actions, $\mathcal{A}^{\mathcal{T}}$; and a transition function that maps previous states and joint actions to next states, T: $(s, a^{\mathcal{L}}, a^{\mathcal{T}}) \mapsto s'$. For example, the world state could be Fido's location in the yard, represented as x and y coordinates. His available actions could be walking one tile up, down, left, or right. The teacher's actions could be giving different rewards as feedback, $f \in \mathcal{A}_{\mathcal{T}} = [-1, 1]$. Negative numbers correspond to punishment and positive numbers correspond to rewards. The absolute value of f reflects the degree of the feedback. Thus praise for Fido would be positive and moderate (e.g., $f \approx +0.5$) while an electric shock would be negative and large (e.g., $f \approx -1.0$). Finally, the transitions would move Fido to new locations in the yard based on the previous state and the action that he took.

At each timestep, t, the learner takes an action from a state, the teacher responds to this action, and then they transition to a new state. Thus, after t timesteps, there has been a history of interaction between the two, $h_t = (s_0, a_0, f_0, s_1, a_1, f_1, \ldots, s_t, a_t, f_t)$.[1]

Modeling Teaching and Learning

A learner's behavior results from a history of interaction with the environment and the teacher. Formally, a learning strategy, $\mathcal{L}$: $h_t \mapsto \pi_t$, specifies how a history of interaction produces a behavioral policy that maps states to actions, π_t: $s \mapsto a^{\mathcal{L}}$. Meanwhile, a teacher who has a target policy, π^*, will have a teaching strategy, T: $h_t \mapsto F_t$, that provides a feedback policy based on a history of interaction. A feedback policy is how the teacher will provide feedback in response to an action that a learner takes from a state, F_t: $(s_t, a_t^{\mathcal{L}}) \mapsto f_t$. Given a learning strategy $\mathcal{L}$ and a target policy π^*, a teaching strategy can be better or worse depending on the learning sequence $\pi_{0:H} = (\pi_0, \pi_1, \pi_2, \ldots, \pi_H)$ that it induces. In particular, we assume that from a teacher's perspective, it is better when more of the target policy is learned earlier.

Our goal here is to understand how people teach with evaluative feedback, which means characterizing what learner models people's strategies can teach. To do this, we can consider different classes of learning models, $\mathcal{L}$, that could be taught. Specifically, we characterize two learning models: reward-maximizing (i.e., reinforcement) and action-signaling (i.e., communication).

Reward maximization. A learning agent that treats feedback as a quantity to maximize is equivalent to algorithms studied in reinforcement learning (RL; Sutton & Barto, 1998). Formally, we can define a reward-maximizing learning strategy, $\mathcal{L}^{\mathrm{RM}}$. The key assumption of a reward-maximizing learner is that it treats feedback from the teacher as a reward to maximize. This means it represents the value (i.e., maximum expected cumulative discounted reward) of taking actions from states:

$$Q(s_t, a_t) = f_t + \max_{\pi} \mathbb{E}_{\pi}[\gamma f_{t+1} + \gamma^2 f_{t+2} + \ldots | s_t, a_t] \quad (1)$$

where $\gamma \in [0, 1]$ is the learner's discount rate. A reward-maximizing policy is simply one that maximizes the action-value in every state:

$$\pi^{\mathrm{RM}}(s) = \arg\max_{a \in \mathcal{A}(s)} Q(s, a), \quad (2)$$

for all $s \in S$.

RL studies different algorithms that calculate or estimate Q based on histories of environment transitions and rewards. Because the problem of positive reward cycles applies to any reward-maximizing learning algorithm, we set aside the question of which specific algorithm a learner implements in Experiments 1 and 2. In Experiment 3, we consider two well-known classes of reward-maximizing algorithms: model-free and model-based learning. One motivation for investigating these two classes of models is that they each capture two broad types of reward-maximizing learning characterized in humans and animals. Model-free learning aligns closely with theories of operant conditioning, whereas model-based learning corresponds to goal-oriented planning based on a model of rewards and transitions (Dayan & Niv, 2008). Implementation details are included in Appendix B.

Action-signaling. A learning agent who treats feedback as a signal of the correctness of an action can be modeled as performing inference over possible target policies given the teacher's feedback. Rather than being treated as a quantity to maximize, feedback is diagnostic of an unknown variable: the teacher's target policy, π^*. Thus, we can draw on Bayesian models of social cognition and language to model these inferences (Baker, Saxe, & Tenenbaum, 2009; Loftin et al., 2014).

An action-signaling model, $\mathcal{L}^{\mathrm{AS}}$, treats evaluative feedback as signals. To estimate what the teacher's target policy is, an action-signaling learner has a generative model consisting of two components: (a) the teacher's target policy and (b) the history of interaction, including the feedback. The probability of a policy π^* given a history of interaction can be formally expressed as Bayesian inference:

$$P(\pi^* | h_t) \propto \prod_{i=0}^{t} P(f_i | s_i, a_i, \pi^*)\, P(\pi^*). \quad (3)$$

In particular, $P(f_i | s_i, a_i, \pi^*)$ is how the learner expects the teacher to give feedback as signals if π^* were the target behavior. Minimally, this could encode a feedback strategy in which the teacher rewards when the action matches the target action in a state (i.e., the action is correct) and punishes when it does not match the target action (i.e., the action is incorrect).[2] The prior term, $P(\pi^*)$, represents a learner's belief that the teacher is trying to teach policy π^* before any feedback is given. Now that we have specified the quantity computed by an action-signaling learner, we can specify how it selects actions given a history of interaction h_t:

$$\pi^{\mathrm{AS}}(s) = \arg\max_{a \in \mathcal{A}(s)} \sum_{\pi'} P(a | \pi', s) P(\pi' | h_t), \quad (4)$$

where $P(a | \pi', s) = \mathbb{1}\{\pi'(s) = a\}$ is an indicator function for whether a is the correct action for a possible target policy π'. The algorithmic details of the model are presented in Appendix B.

[1] For simplicity, we assume that state transitions and learners' learned policies 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.

[2] For the purposes of distinguishing feedback as communicative versus reinforcement, the exact form of the likelihood function is not particularly important. In simulations and pilot studies, we found similar results when piecewise linear functions or sigmoidal functions were used.

Reward-Maximization, Action-Signaling, and Positive Reward Cycles

Teaching a reward-maximizing learner requires creating and modifying a system of incentives to motivate certain behaviors. In contrast, teaching an action-signaling learner requires providing signals that an agent can use to infer the intended target policy. These two accounts make identical predictions in some situations. For example, if the behavior being taught involves only choosing one of several actions, such as pressing a specific lever once, whether feedback is being used to incentivize or to signal cannot be determined.

However, the two theories can diverge when giving feedback in multistate, multiaction situations. For instance, consider the example with Fido introduced at the beginning of this article, in which he is being taught to follow a specific path to reach the house. A simplified version of this is shown in Figure 1A. A teacher who provides feedback as signals runs the risk of creating positive reward cycles, sequences of states, actions, and feedback in which a learner returns to an initial state $(s_0, a_0, s_1, a_1, \ldots, s_n, a_n, s_0)$ and receives a net positive reward, $F(s_0, a_0, s_1) + \gamma F(s_1, a_1, s_2) + \ldots + \gamma^n F(s_n, a_n, s_0) > 0$ (Ng et al., 1999). The following example provides some intuition: Suppose Fido's owner is attempting to signal that walking along the desired path toward the house or going into the house is good, walking into the garden is bad, and exiting the garden is good. In that case, she would reward the correct actions (e.g., walking along the path to the house, exiting the garden), and punish the incorrect actions (e.g., walking into the garden). This poses a problem if Fido is trying to maximize rewards. In that case, Fido will get more rewards by ignoring the house entirely and attempting to return to an earlier part of the path, either by going back along the path, or by going through the garden and back onto an earlier position. In other words, Fido will exploit a positive reward cycle. Figure 2 illustrates how this phenomenon emerges from the interaction of different teachers and learners.

In general, if people are using feedback as signals, positive reward cycles are not inherently a problem and may even be useful for conveying information. On the other hand, if people are using evaluative feedback under the assumption that learners are reward-maximizing, then positive reward cycles should not appear or be sustained. If learners do begin cycling to maximize rewards, then the teacher would adjust their feedback to remove them. Thus, finding that people produce positive reward cycles or consistently fail to remove them would support the action-signaling account while providing strong negative evidence against the reward-maximizing account. A more formal treatment of this argument can be found in Appendix A.

When assessing the presence of positive reward cycles empirically, there are several factors to consider about the teacher's model of the learner. In particular, the learner's discount rate, planning horizon, and nonfeedback rewards will affect whether positive reward cycles are present. In addition, the teacher's strategy may be affected by their own rewards distinct from the motivation to teach. We discuss each of these factors below.

Intermediate rewards. A reward-maximizing agent will exploit states and actions that lead to the greatest reward. As such, in the Path-Following domain we consider in this article, positive reward cycles largely come down to a teacher overincentivizing actions that are an intermediate part of the target task, which will be reflected in the pattern of feedback. This also means a teacher can always remove positive reward cycles by reducing intermediate rewards. In addition, once an agent has learned the task it is no longer taking actions that would receive punishment. As a result, whether people reduce intermediate rewards is key to assessing if feedback is being used as incentives.

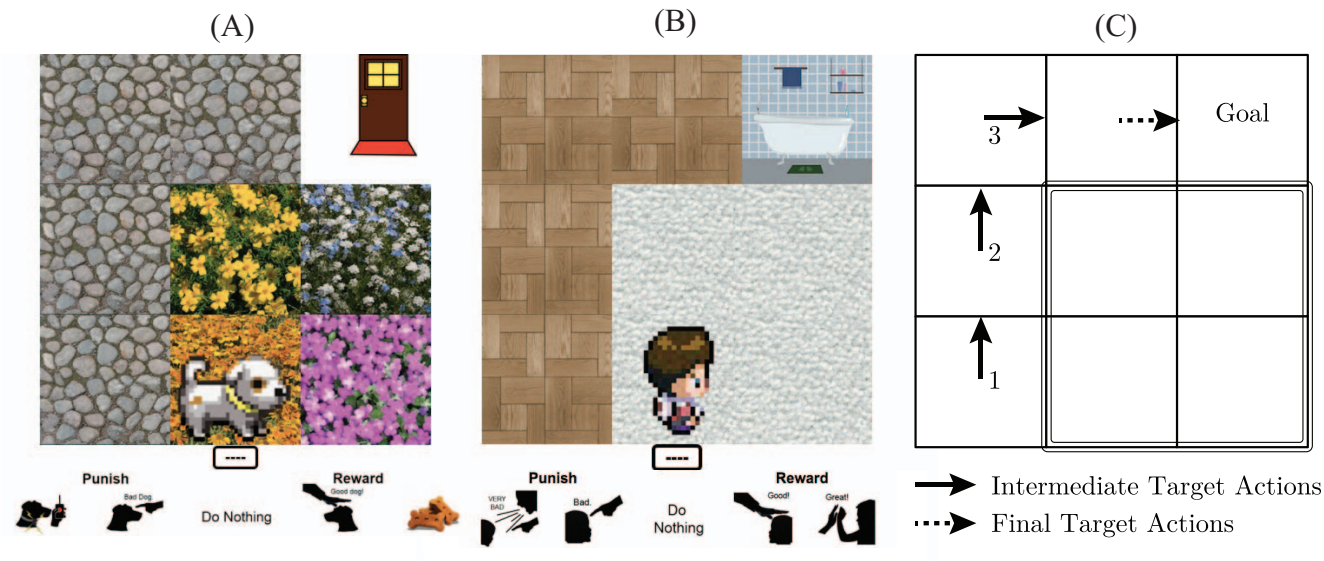

Figure 1. Path-following task. On each trial, the agent moves and then participants give their feedback. A: Dog version. B: Child version. C: Intermediate and final target actions. See the online article for the color version of this figure.

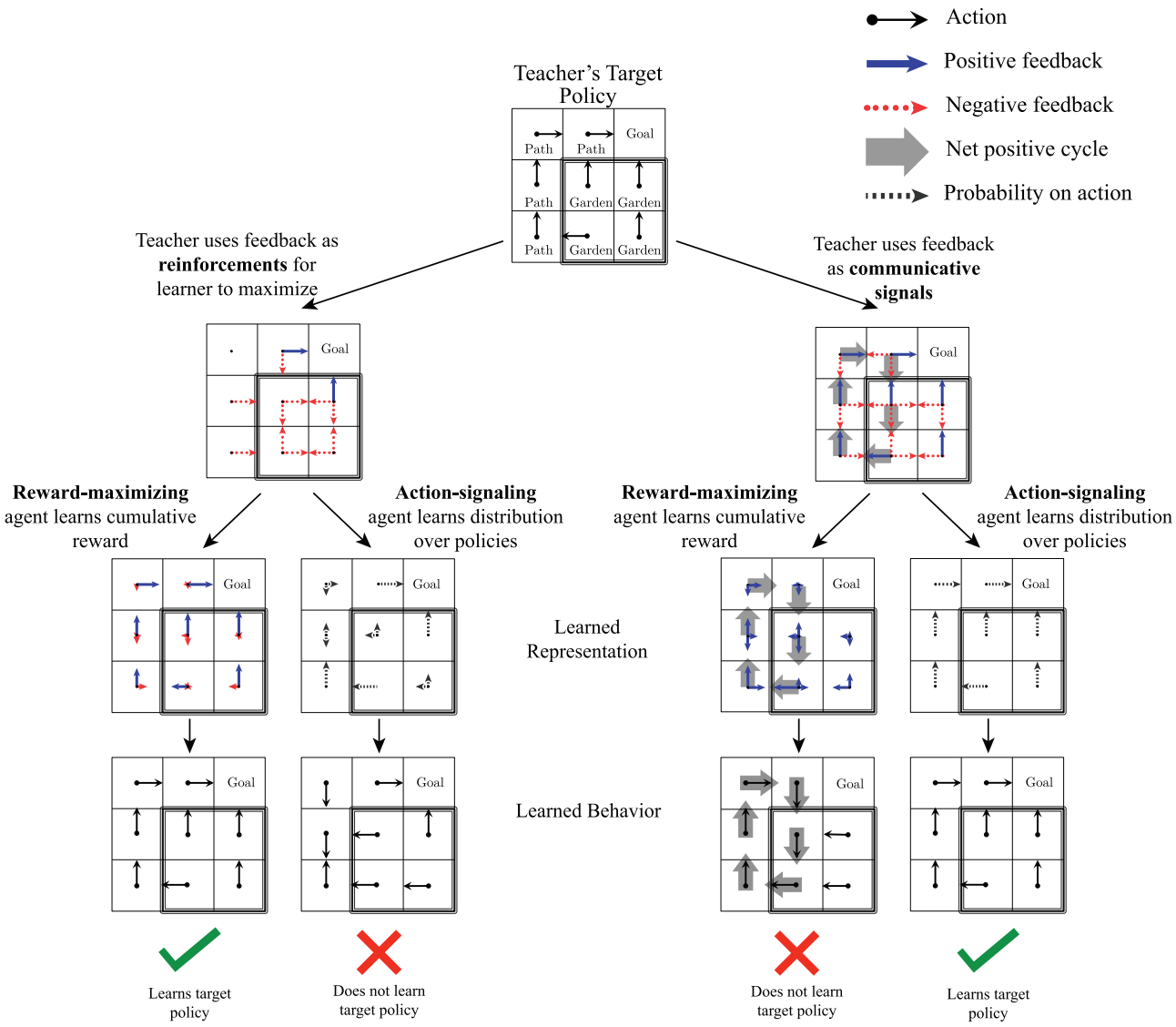

Figure 2. Feedback strategy and learner type interaction. A teacher rewards and punishes to teach a target policy. For clarity, we show static schedules of feedback, but this analysis also applies in the dynamic case (e.g., Experiment 3). Left-side: The teacher rewards only for entering the goal and punishes for entering the garden tiles. A reward-maximizing agent learns the cumulative expected reward (left branch; arrow length and color correspond to magnitude and valence, respectively); an action-signaling learner attempts to infer the target policy but is provided with insufficient information (right branch; arrow length corresponds to probability). Right-side: A teacher rewards only for correct actions and punishes for incorrect actions. The action-signaling learner can infer the target policy (right branch). But the reward-maximizing learner will want to exploit the positive reward cycle (large gray arrows) in the teacher's pattern of feedback and will not learn the target policy (left branch). The reward-maximizing learner's action value function, Q_F, is calculated with $\gamma = .9$. The action-signaling prior is uniform: $P(\pi^*) \propto 1$. See the online article for the color version of this figure.

Punishments. Whether intermediate actions are overincentivized is relative to any punishments received to continue receiving intermediate rewards. This leads to a second way for teachers to remove positive reward cycles: Increase punishments such that it is no longer worthwhile for the incentivized agent to collect intermediate rewards. This corresponds to increasing punishments for states and actions that are not part of the target policy.

Intrinsic learner rewards. The learner may have rewards other than those provided by the teacher. For instance, Fido may have a bias for efficiency and not want to take unnecessary actions, or he may actively want to play in the flower bed. In this article, we focus on the case where these nonfeedback rewards are negligible, however, future work will need to explore the interaction of teaching rewards and environmental rewards.

Learner discount rate. Because a positive reward cycle results from rewards and punishments that are integrated over time, whether and how an agent trades off temporally close and distant events is important. The learner's discount rate must be

high enough such that future rewards offset any earlier punishments.

Fixed temporal horizon. Positive reward cycles can still pose a problem even when an agent has a fixed number of steps in a time period (i.e., teaching episode). For example, stepping through a positive reward cycle and only then completing a task would yield more rewards than simply completing a task and ending a teaching episode.

Teaching costs. Our analyses here do not explicitly model what occurs when it is costly for a teacher to provide rewards and punishments. However, the influence of this factor on teacher behavior would be reflected in the teaching strategy $\mathcal{T}$. For instance, the teacher be less motivated to teach after many timesteps of low progress. Future work will need to examine the influence of such aspects of the teaching strategy.

Experimental Studies

Motivated by our analyses of evaluative feedback as reinforcement and communication, we designed an agent-teaching paradigm, the Path-Following task (Figure 1A and 1B), reminiscent of the situations we have described. The teacher's goal is to teach the learner to go from the start state (bottom left) to the terminal goal state (upper right) without going through the four tiles on the bottom right. As shown in Figure 1C, this results in a "target behavior" consisting of four actions in sequence. The first three actions are "intermediate actions," whereas the final action is a terminal action that ends a teaching episode.

Experiment 1 tested whether people produced positive cycles when delivering feedback for individual actions. In the studies in Experiment 2, participants taught a single dog preprogrammed to improve over time to see whether aspects of positive reward cycles would persist across all stages of successful training. For the studies in Experiment 3, participants trained agents that learned from feedback in accordance with the reward-maximizing and action-signaling models. This allowed us to assess how people adapted their teaching strategies to learner behavior. Across the studies we varied the particular algorithms (model-based, model-free, action-signaling), random exploration, and agent type (dogs and children).

Experiment 1: Rewarding and Punishing Isolated Actions

In Experiment 1, participants provided feedback to learners who performed isolated actions in the Path-Following task. This provides us with a static measure of participants' feedback strategy over the entire state-action space, providing an initial assessment of whether people produce positive reward cycles. In addition, Experiment 1b focuses on whether the findings of Experiment 1a hold when feedback is presented as different types of rewards and punishments to further validate our experimental design.

Experiment 1a

Method.

Participants and materials. Forty (18 female; 22 male) Amazon Mechanical Turk participants were paid $1.00 to participate. One was excluded due to a technical error. This number of participants was selected in advance. On each trial the dog started at a tile, rotated to face one of the four directions, and walked onto the adjacent tile (3,000 ms). Participants then provided feedback ranging continuously from highly negative to highly positive with the following anchor points: "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". Participants used a draggable slider that was initialized at the midpoint of the scale ("doing nothing") and had the five values marked equidistantly along the continuous scale. The instructions stated that the scale should be seen as balanced, such that distances from the midpoint of the scale were equivalently positive or negative. Participants could only continue to the next trial after providing a feedback response. Procedures were approved by the Harvard University Committee on the Use of Human Subjects (protocol #IRB14-2016, title: "A computational approach to human moral judgment").

Procedure. Participants were told that they would assist in training a "school of 24 distinct dogs" to "go into the house by staying along the path and staying out of the garden." The goal of training is for each dog to be able to complete this task independently. As there are nine tiles (four with two actions, four with three actions, and one with four actions), the task consisted of 24 trials covering each combination of initial location, action, and final location. Each trial showed a different color dog to emphasize that distinct learners were being trained. Trial order was randomized with the requirement that no trial began where the previous trial had ended to avoid suggesting a continuity between trials. Participants were then asked to imagine they had placed the dog in that location at the beginning of the trial. They had to answer several comprehension questions correctly to start the task. One of these included a free-response description of what the dog was to be taught.

After completing the main task, participants were asked several questions about their training responses and background. Also, to assess how participants interpreted tradeoffs between punishments and rewards, we asked them 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 (a) scolding twice followed by praising twice, (b) two scoldings followed by three praises, (c) two scoldings followed by four praises, (d) one shock followed by one biscuit, (e) one shock followed by two biscuits, (f) one shock followed by two praises, (g) one shock followed by three praises, and (h) two scoldings followed by one biscuit. The final page asked several demographic questions and for feedback about the experiment.

Results. Our main question is whether participants' feedback contains positive reward cycles that a reward-maximizing learner would exploit. We performed an analysis based on the marked values (e.g., *shock*, *scold*, *praise*, *biscuit*) and dog preference judgments, as well as a model-based analysis based on the numerical value of feedback. Both analyses reveal that people's feedback contain positive reward cycles.

In addition, we report a clustering analysis that shows people generally pursued one of two strategies. One corresponds directly to action-signaling, whereas the other reflected the general permissibility/impermissibility of tiles. Finally, we show that participants use the full range of the response scale while mainly anchoring on the marked responses.

Table 1
Experiment 1a: Dog Preference Question Responses (Counts Out of 39 Participants)

Feedback sequence	Prefers nothing	No preference	Prefers feedback
2 scold + 2 praise	7	7	25
2 scold + 3 praise	7	0	32
2 scold + 4 praise	3	1	35
1 shock + 1 biscuit	16	7	16
1 shock + 2 biscuit	10	2	27
1 shock + 2 praise	20	3	16
1 shock + 3 praise	16	1	22
2 scold + 1 biscuit	10	4	25

Analysis of positive reward cycles. State-action feedback and dog preference judgments (summarized in Table 1) allow us to determine whether a participant "knowingly" produces positive reward cycles. For example, if a participant's responses would allow the dog to walk back and forth between two adjacent tiles and receive *praise* and *biscuit* as feedback, this would constitute a positive reward cycle. Similarly, if responses led to a situation where a dog could return to a location in four steps, receive *scold, praise, praise, praise*, and the participant judged that this feedback would be perceived as a net positive, then that would also be a positive reward cycle. Figure 3A shows five cyclical trajectories that we analyzed as well as the number of participants who produced each. For the two-step cycles on the path tiles, we counted it as a positive cycle if responses to each step were both greater than zero. For the two longer cycles where garden tiles are entered, we coded each response as a *shock*, *scold*, *praise*, or *biscuit* if it was within 0.05 of −1, −0.5, 0.5, and 1.0, respectively. We then determined if the participant gave a dog preference judgment where that sequence (or one strictly less) would be positive. Importantly, this allows us to determine positive reward cycles without making assumptions about the numerical interpretation of feedback or the discount rate.[3] Overall, this analysis revealed that 36 out of 39 participants produced at least one positive reward cycle ($p < .001$; binomial test).

Model-based analysis of positive reward cycles. We also performed a model-based analysis of feedback to numerically assess individual positive reward cycles. First, in our discussion of the models, we considered a feedback strategy of rewarding for correct actions and punishing for incorrect actions (i.e., a particularly extreme version of action-signaling). This would produce a positive cycle starting from the lower left-hand corner and performing the action sequence *up, up, right, down, down, left*. The aggregated pattern of feedback revealed that on average, this series of actions yielded a positive net reward of +1.20 (bootstrap-estimated 95% CI [0.83, 1.61]; one sample t test with $\mu_0 = 0$: $t(38) = 5.99$, $p < .001$) as shown in Figure 3B.

In a more general version of this analysis, we investigated whether any positive reward cycle would be produced across a range of discount rates. To do so, we calculated the reward-maximizing policy for each participant's pattern of feedback, sampled a trajectory, and identified whether the trajectory returned to a previously visited state. For $\gamma = .99$, 38 of 39 participants produced feedback that resulted in a positive reward cycle ($p < .001$; binomial test). This is robust across a number of different values of γ as shown in Figure 3C.

Feedback function types. Previous human-machine interaction studies have shown that different people use different training strategies when giving RL agents rewards and punishments (Loftin et al., 2014). Using each individual's pattern of responses over the state-action space, we investigated the extent to which participant response patterns grouped into training strategies using a clustering analysis. Individual feedback patterns were represented as 22-dimensional vectors between −1 and +1 (actions from the terminal state were excluded), and a dissimilarity matrix was calculated using Manhattan distances. We then used a complete linkage method for the hierarchical clustering analysis.

Figure 4A first shows the average feedback across all participants, whereas Figure 4B shows how the clustering procedure breaks this down into two large, homogeneous clusters ($n = 15$ and $n = 16$) and several smaller clusters ($n = 8$). The average of one of the large clusters closely matches the pattern of responses we expected from the evaluative feedback as communication hypothesis. That is, rewards indicate an action is correct while punishments indicate an action is incorrect. In contrast, the average of the other large cluster (on the right) does not resemble our predictions for either evaluative feedback hypothesis. Rather, responses reflect the general permissibility/impermissibility of state-types. For instance, walking onto a particular path tile is always permissible even if it is not always optimal, so if a learner walks onto a path tile but in the wrong direction, it will get rewarded. Notably, this also produces salient positive cycles—for example, the learner could just walk back and forth along the path. In addition, participants' use of this state-training strategy is not attributable to a misunderstanding of the task because only five of the 16 state training participants failed to explicitly mention a goal of going to the house in a pretask free-response question.

Response scale usage. Although participants were given a continuous scale on which to give feedback, 79% of responses were within 0.05 of the five marked parts of the scale (−1.0, −0.5, 0.0, 0.5, 1.0). Only 10.5% of all responses were within 0.05 of the default slider response of 0. This suggests participants anchored on the marked portions of the scale, perhaps because they corresponded to specific concrete outcomes for the dog. In addition, participants used the full range of punishments and rewards, and they used rewards and punishments equally—all state-action responses: $M = -0.03$; median = 0.00; $SD = 0.67$; one sample t test against mean of 0: $t(935) = -1.26$, $p = .21$.

Experiment 1b

The design of Experiment 1a assumes that participants treat different types of feedback like a *shock*, *scold*, *praise*, and *biscuit* as being similar in kind. Namely, that all four types of feedback are rewards and punishments that could be used either as reinforcements or communicatively. However, a *shock* and *biscuit* are

[3] In addition, note that the identified positive reward cycles would cause problems when a reward-maximizing agent is not discounting and aware of the fixed length of a training episode. For example, a learner that had a horizon of six steps from the first state could plan to spend two additional steps on the path tiles before entering the goal. With eight steps, an agent could take the smaller loop through the garden before entering the goal (i.e., *up, up, right, down, left, up, right, right*), whereas 10 steps would allow a reward-maximizing agent to take the larger loop through the garden before entering the goal.

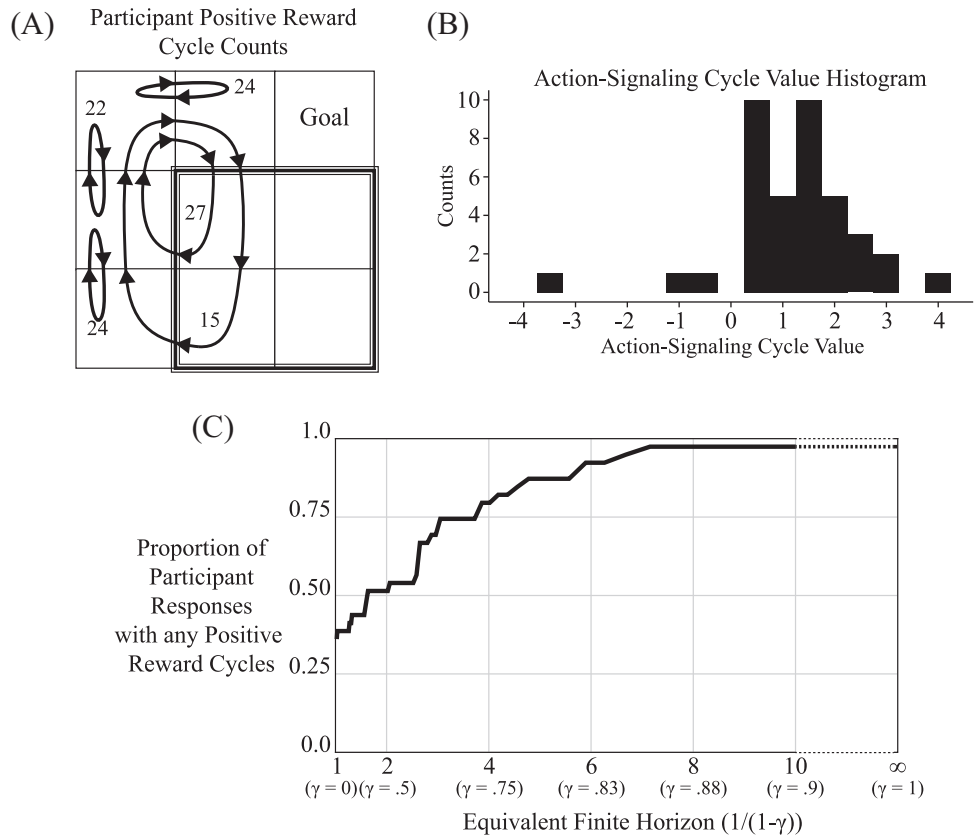

Figure 3. Experiment 1a: Positive reward cycle analysis. Our data reveal that people readily produce positive reward cycles that can teach an action-signaling agent but not a reward-maximizing agent the target task. A: Circular paths represent state-action sequences that could become positive reward cycles. Numbers indicate how many participants "knowingly" produce a positive reward cycle along that path given a qualitative coding of responses (e.g., as a *shock*, *scold*, *nothing*, *praise*, or *biscuit*) and dog preference judgments. Overall, 36 of 39 participants produced at least one of these positive reward cycles. B: Participants' responses have numerical values associated with them. This plots a histogram of the net numerical value of responses to the largest cycle shown in Panel A, which we a priori posited would be produced by a simple action-signaling teacher. C: This displays the proportion of participants whose responses produce a reward cycle for different discount rates, γ. Discount rates are converted to $1/(1-\gamma)$, which can be heuristically thought of in terms of an equivalent finite horizon since a higher γ means temporally distant rewards are less discounted (LaValle, 2006). Portions of this figure are adapted from Figure 4 in Teaching with rewards and punishments: Reinforcement or communication? *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 920–925), by M. K. Ho, M. L. Littman, F. Cushman, and J. L. Austerweil, 2015, Austin, TX: Cognitive Science Society. Copyright 2016 by Cognitive Science Society. Adapted with permission.

canonical reinforcements, whereas a *scold* and *praise* are canonical communicative actions. As a result, it is possible that participants conceptualize the scale as a mixture of communicative and reinforcing feedback, which would pose a problem for our analyses. To ensure this is not the case, we replicated Experiment 1a but with either purely reinforcement or purely communicative labels on the response scale.

Method and procedure. Eighty participants were recruited via Amazon Mechanical Turk to participate and were each paid $2.00. We used the PsiTurk experimental platform (Gureckis et al., 2016). The same procedure as Experiment 1a was used except we separated participants evenly into two conditions. In the reinforcement-scale condition, participants were given a scale with the original extreme values (*shock* and *biscuit*) but the intermediate values were replaced with *tap* and *rub*. In the communicative-scale condition, the original moderate values were used (*scold* and *praise*) and the extreme values were replaced with *harsh scold* and *strong praise*. The images used are shown in Figure 5.

For reinforcement scale, we also modified the dog preference questions in which the following sequences were compared to no feedback: 1 tap + 1 rub; 1 tap + 1 shock + 4 rub; 1 shock + 1 tap + 4 rub; 1 shock + 1 biscuit; 1 shock + 2 rub; 1 shock + 2 biscuit; 1 shock + 3 rub; 2 tap + 1 biscuit; 2 tap + 3 rub; 2 tap + 4 rub; 2 shock + 4 rub. Analogous preference questions were asked for communicative scale. Experimental procedures were approved by the University of Wisconsin-Madison Education and Social/Behavioral Science IRB (ID #20170830, title: "Exploring

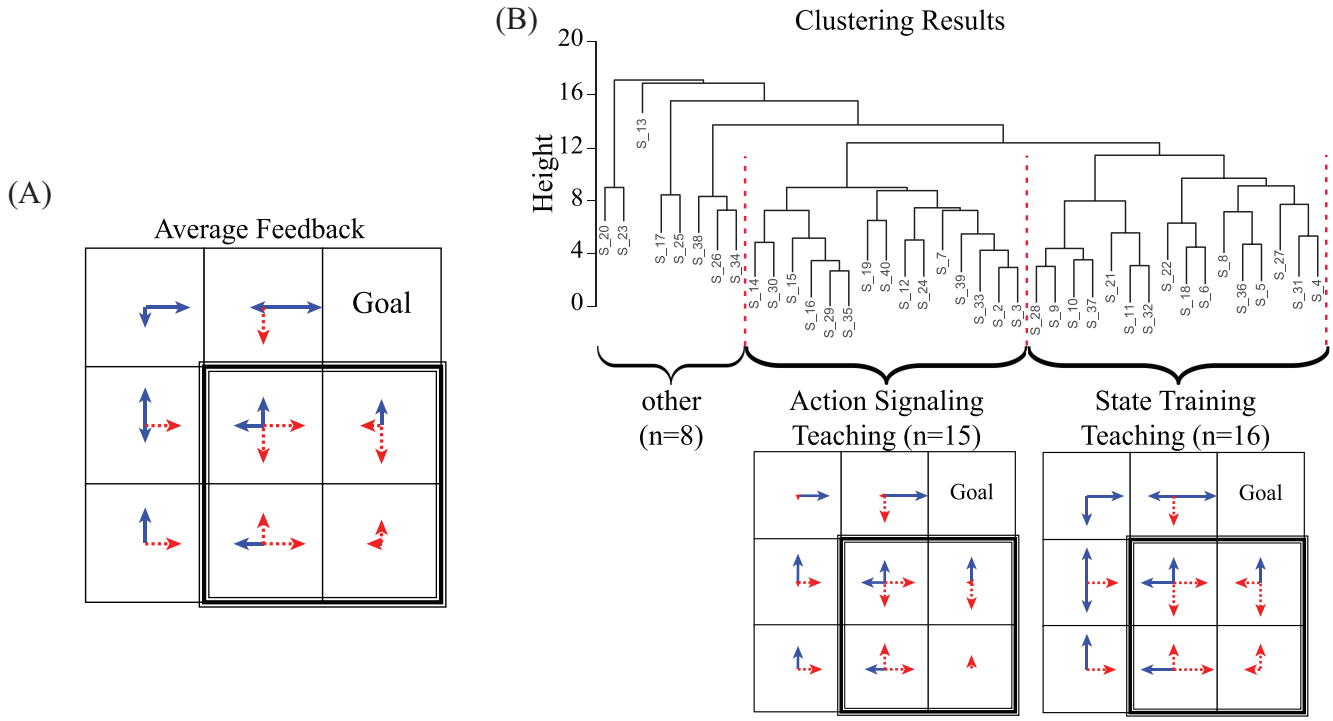

Figure 4. Experiment 1a: Feedback function analysis. Participants' global patterns of feedback accord more with action-signaling than reinforcing. A: The average feedback function over all participants. The arrow-length denotes response magnitude, whereas color denotes valence (blue/solid line = positive; red/dotted line = negative). B: Results of hierarchical clustering of participants' responses with the average teaching function of the two largest clusters. These correspond to "action-signaling" (left) and "state training" (right). Portions of this figure are adapted from Figure 4 in Teaching with rewards and punishments: Reinforcement or communication? *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 920–925), by M. K. Ho, M. L. Littman, F. Cushman, and J. L. Austerweil, 2015, Austin, TX: Cognitive Science Society. Copyright 2016 by Cognitive Science Society. Adapted with permission. See the online article for the color version of this figure.

human and machine decision-making in multi-agent environments").

Results. We found no qualitative differences between the two conditions. This was determined using the judgment-based analysis reported in Experiment 1a as well as additional quantitative analyses.

As in Experiment 1a, the state-action feedback and dog preference judgments provided by participants allow us to assess whether they "knowingly" produced positive reward cycles. Using this metric, we found no difference in the number of participants whose responses and judgments implied any positive reward cycle (reinforcement scale: 36 of 40; communication scale: 35 of 40; Fisher's exact test: odds ratio = 0.78, $p = 1.0$). There were similarly no detectable differences based on whether we looked at cycles involving only path tiles (reinforcement scale: 28 of 40; communication scale: 27 of 40; χ^2 test for independence: $\chi^2(1) = 0.00$, $p = 1.00$) or those involving path and garden tiles (reinforcement scale: 27 of 40; communication scale: 22 of 40; $\chi^2(1) = 0.84$, $p = .36$).

We also compared numerical responses (coded to range over $[-1, 1]$) for each state and action between the two conditions. A mixed-effects linear model with condition as a fixed effect and participant intercept and state-action intercept as random effects were fit. Tests with Satterthwaite's approximation show there was no significant difference in condition ($\beta = 2.9-10^{-2}$, $SE = 2.9-10^{-2}$ $t(78.0) = 1.01$, $p = .32$), indicating that participants generally did not treat the two scales differently.

Discussion. In these experiments, participants trained different learners by giving feedback for isolated actions. Several key results emerged: First, teachers incentivized intermediate successes to a degree sufficient to generate positive reward cycles. Because positive reward cycles tend to prevent reward-maximizing learners from attaining the target policy, this indicates that people do not use rewards and punishments as incentives for reward-maximizing learners, but rather as a form of communication. Second, participant feedback clustered into two general types: action-signaling and state-training. This first type reflects the predicted tendency to use rewards and punishments as signals for the correctness and incorrectness of actions. In contrast, the state-training pattern of feedback was not originally predicted by either model. It may be that these 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. Finally, in Experiment 1b, we confirmed that in this paradigm, participants use actions that are superficially reinforcing (e.g., shocks) and communicative (e.g.,

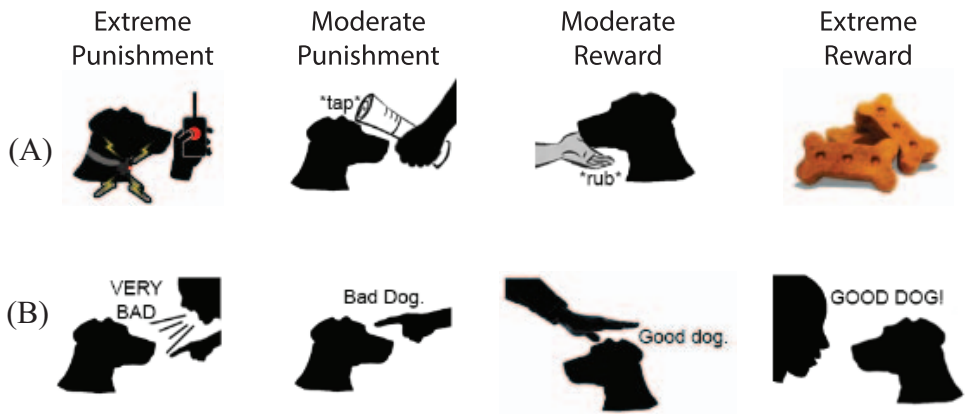

Figure 5. (A) Labels in reinforcement-scale condition and (B) communicative-scale condition. See the online article for the color version of this figure.

praise) actions similarly. This validates the use of the mixed scale in Experiment 1a, which we use in Experiments 2 and 3.

A feature of these experiments is that they eliminate any history that a teacher may have with a learner. On the one hand, this permits comparison of different participants' responses to the same learner action taken independent of previous or future actions. On the other hand, teaching is interactive, ongoing, and contextual and involves a teacher responding to a learner's entire history of previous behaviors in different situations and not just decontextualized actions. Our remaining studies investigate whether these general results hold in more realistic and interactive contexts.

Experiment 2: Teaching Improving Agents

In Experiments 2a and 2b, participants taught a single dog preprogrammed to improve over time. This allowed us to hold the interactions between a teacher and the learner constant. In particular, we were interested in whether, as the learner progressed, teachers would reduce their rewards to the point where they were no longer overincentivizing intermediate actions. In other words, would teachers eventually stop producing positive reward cycles? If, by the time the learner acquires the task, teachers' feedback constituted an incentive structure consistent with the desired task and without positive reward cycles, then the results of Experiment 1 would be weaker evidence against the reward-maximizing hypothesis. However, if participants still produce positive reward cycles even when the target task was learned, then we would have additional evidence that people use rewards and punishments as communication.

Experiment 2a

Method.

Participants and materials. We used the same interface as Experiment 1. Forty Amazon Turk participants (16 female; 24 male) were paid $1.00 and told they would receive a bonus up to $0.75 (although all received the full $0.75 bonus). Three participants' data were excluded from analysis because of technical problems. Participants were told they would train a single dog over 8 game days. Each day the dog started at the lower left tile and the day lasted six steps or until the dog entered the house tile. The dog's performance increasingly reflected the target policy over the first 5 days (regardless of the participant's feedback) and matched the target policy exactly on the 6th and 7th days. Specifically, the dog's behavior on Days 1 through 7 was ε-greedy, meaning that it chooses an optimal action in a given state with probability $1 - \varepsilon$ or any of the other (valid) actions with probability ε/(# of suboptimal valid actions). ε was 1.0, 1.0, 0.45, 0.1, 0.1, 0.0, and 0.0 for Days 1 to 7, respectively. The actions the dog took were sampled once for all participants, so all participants were shown the same specific actions. On the 8th day, the dog performed the actions of the a priori defined positive cycle from Experiment 1 (up, up, right, down, down, left).

Procedure. Participants were told they would train a single dog over the course of 8 game days. They were also told that their bonus would be contingent on the dog's solo performance, which we would test by having the dog do the task on its own three times after the experiment. At the end of each "game day," participants rated the dog's current ability using a continuous slider coded from 0 to 1, and after Days 2–8, its improvement compared to the previous day also using a continuous slider from Experiment 1a.

Following completion of the task, participants 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?" Experimental procedures were approved by the Harvard University Committee on the Use of Human Subjects (protocol #IRB14-2016, title: "A computational approach to human moral judgment").

Results.

Perception of task. Participants believed that they were teaching the dog effectively and did not suspect that their feedback had no effect on the dog. All responses to a 5-point Likert scale about dog responsiveness were above 1, which was labeled "not responsive at all" ($M = 3.45$, $SE = 0.11$). Further, all 37 responses to "Overall, how good do you think you were at training the dog in this task?" were above 4 on a 7-point Likert scale ($M = 5.48$, $SE = .12$).

Positive reward cycles and diminishing rewards. While 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 six steps corresponding to the extreme

action-signaling positive reward cycle. Although smaller, the average total reward for these six steps was still a positive value (Figure 6A; $M = +0.67$; bootstrap-estimated 95% confidence interval [CI] [0.31, 1.06]; one-sided t test: $t(36) = 3.53$, $p < .001$). Note that participants may have produced more positive reward cycles if we had tested every combination of current state, action, and next state, as in Experiment 1. However, we did not test all combinations so that we could test how people taught when interacting with a single learner.

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 (Figure 6B). Over those rounds, the learning agent performed each intermediate target action at least once. We analyzed how responses to intermediate actions (and not the final action) changed over time (e.g., as shown in Figure 1C) and across the different steps using a mixed effects linear regression with day, intermediate target action, and their interaction as fixed effects, and intercepts, day, action, and day-action interaction as random effects across participants. Tests of significance were performed using Satterthwaite's approximation. The interaction between intermediate target action and day was not significant ($\beta = -2.7 \times 10^{-3}$, $SE = 3.9 \times 10^{-3}$, $t(382.00) = -0.69$, $p = .49$). The change in reward by day was negative—$\beta = -3.4 \times 10^{-2}$, $SE = 9.0 \times 10^{-3}$, $t(39.00) = -3.71$, $p < .001$—and the change in reward by intermediate target action was positive—$\beta = 5.7 \times 10^{-2}$, $SE = 4.2 \times 10^{-2}$, $t(82.00) = 2.81$, $p < .01$. Thus, people lowered rewards over time and gave greater rewards for actions closer to the goal.

Rewards on the final day were still positive. We fit a mixed-effects linear model to only the final day intermediate target actions. Intercepts and action were random effects across participants, and action was a fixed effect. The intercept of the model was positive—$\beta = 0.65$, $SE = 4.3 \times 10^{-2}$, $t(37.00) = 15.08$, $p < .001$—indicating that although intermediate rewards decreased over time, they did not reach zero. This contributes to the presence of net positive cycles on the final day (Figure 6B). One might wonder whether rewards for intermediate actions would decrease to zero if we tested participants for more days with perfect learner performance. This is examined in Experiment 2b.

Tracking learner ability and improvement. Figure 6C depicts mean participant judgments of ability and improvement over days compared to the dog's true ability and improvement. Participants can only observe the actions, leaving the precise policy used by the dog uncertain. Despite this, dog ability judgments tracked the

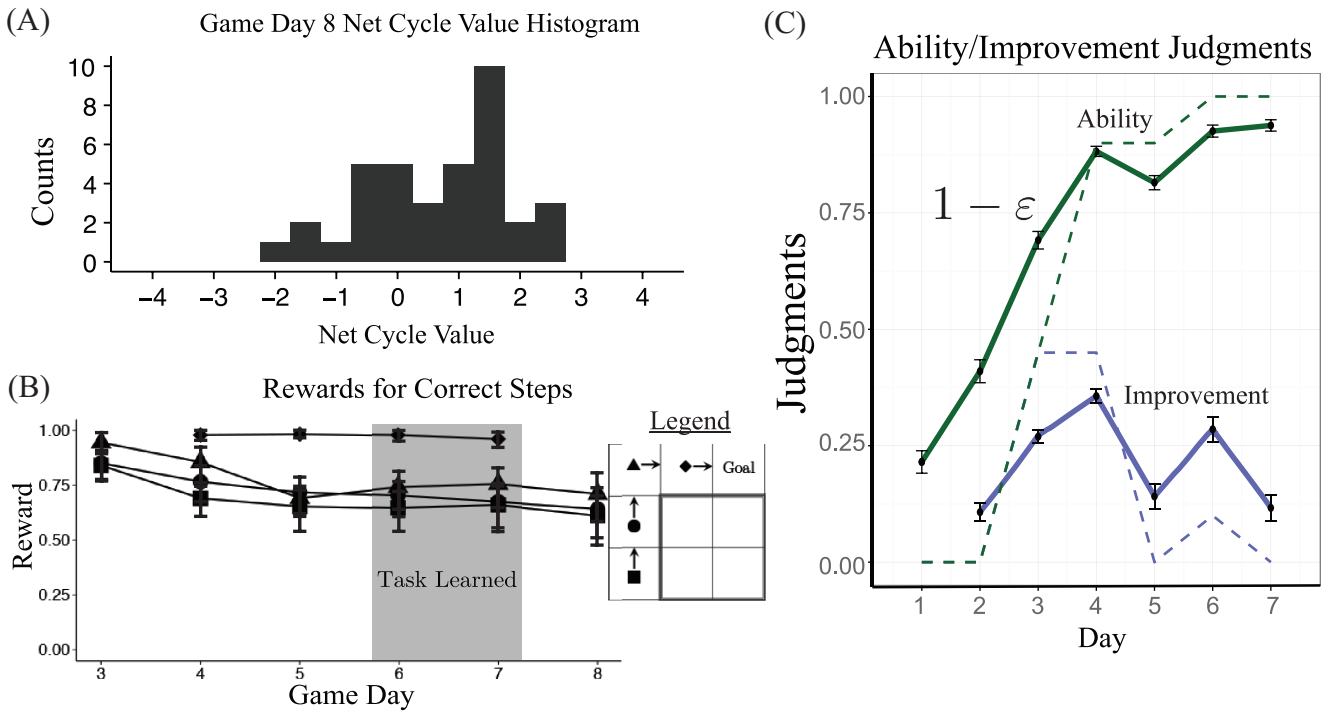

Figure 6. Experiment 2a results. We find that when providing feedback over time, participants continue to produce positive reward cycles. A: Histogram of final game day net cycle values. B: Average rewards for each of the four correct steps on each day. Different target actions are designated by different shapes. C: Average ability (green/dark grey) and improvement (blue/light grey) judgements over the eight game days (solid lines) along with the true ability and improvements in terms of $1 - \varepsilon$ (dotted lines). This indicates participants can track ability as well as changes in ability. From Figure 5 in Teaching with rewards and punishments: Reinforcement or communication? *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 920–925), by M. K. Ho, M. L. Littman, F. Cushman, and J. L. Austerweil, 2015, Austin, TX: Cognitive Science Society. Copyright 2016 by Cognitive Science Society. Adapted with permission. Error bars are bootstrapped 95% confidence intervals. See the online article for the color version of this figure.

value of $1-\varepsilon$ extremely closely—results of Fisher transformed Pearson r two-sided t test: $M\ r = .95$; 95% CI [.94, .97]; $t(36) = 22.98$, $p < .001$. Dog improvement judgments tracked day-to-day changes in ε but more weakly—results of Fisher transformed Pearson r two-sided t test: $M\ r = .44$, 95% CI [.29, .56]; $t(36) = 5.67$, $p < .001$. Thus, when teaching via evaluative feedback, teachers can track the current state of the learner's policy and changes to that policy over time (Figure 6C).

Experiment 2b

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

Method.

Participants and materials. Forty-one Amazon Mechanical Turk workers (13 female, 28 male) participated, with no exclusions. The structure of the experiment was the same as Experiment 2a but with more game days. Participants trained learners that improved over Days 1 to 4 ($\varepsilon = 1.00, 1.00, 0.55, 0.10$), performed perfectly over Days 5 to 11, performed the six-step cycle on Day 12, and then regressed on Day 13 ($\varepsilon = 0.55$).

Procedure. Participants were given the same instructions as in Experiment 2a. Experimental procedures were approved by the Harvard University Committee on the Use of Human Subjects (protocol #IRB14-2016, title: "A computational approach to human moral judgment").

Results.

Final positive reward cycles and diminishing rewards. Figure 7A shows the distribution of net cycle values on Day 12, when the learner performed the six-step cycle. After 7 days of performing the correct intermediate actions, the average net cycle value was not significantly different from 0—$M = +0.19$, bootstrap-estimated 95% CI [−0.17, 0.55]; one-sided t test: $t(40) = 1.03$, $p = .15$.

From Days 2 to 12 the dog performed intermediate actions. Over the course of these game days, participants' rewards diminished over the course of the experiment but did not disappear completely. Figure 7B shows the change in reward magnitudes for the three intermediate target actions as well at the final target action. We fit a mixed-effects linear model to feedback for intermediate target actions over these game days with the same fixed/random effects and tests as the analysis in Experiment 2a. Both the intercept, $\beta = 0.89$, $SE = 4.8 \times 10^{-2}$, $t(41.27) = 18.41$, $p < .001$, and fit for day, $\beta = -4.7 \times 10^{-2}$, $SE = 7.4 \times 10^{-3}$, $t(41.28) = -6.27$, $p < .001$, were significant, whereas action, $\beta = 2.4 \times 10^{-2}$, $SE = 1.6 \times 10^{-2}$, $t(61.98) = 1.52$, $p = .13$, and its interaction with day, $\beta = -5.1 \times 10^{-4}$, $SE = 2.0 \times 10^{-3}$, $t(249.29) = -0.26$, $p = .80$, were not.

In addition, feedback for intermediate target actions on the 12th day were still positive. We fit a mixed-effects linear model with intercepts and action across participants as random effects and action as a fixed effect. The intercept was above zero, $\beta = 0.43$, $SE = 5.4 \times 10^{-2}$, $t(41.24) = 7.78$, $p < .001$. Thus, intermediate rewards decrease steadily over time but do not reach zero even after the learning agent performs intermediate target actions perfectly for 10 game days and the complete sequence for 7 game days.

Judgments of ability and improvement. Participants tracked ability and improvement during the initial game days (1–5), during which ε was changing. Ability judgments correlated with true ability—results of Fisher transformed Pearson r two-sided t test: $M\ r = .93$, 95% CI [0.91, 0.95]; $t(40) = 23.90$, $p < .001$—and improvement judgments correlated with true improvement—Fisher transformed Pearson r two-sided t test: Mean $r = .49$, 95% CI [0.35, 0.61]; $t(40) = 6.48$, $p < .001$.

Discussion

The results of Experiments 2a and 2b further support the communicative hypothesis. Agents in these experiments improved over time as they were given feedback, yet participants continued to reward them for intermediate successes. Participants did not entirely remove rewards for intermediate target actions after 2 (Experiment 2a) or 7 days (Experiment 2b) of performing the task perfectly. Continuing to reward intermediate target actions often leads to positive reward cycles, as we discuss in the General Methods section. The one positive reward cycle that we examine

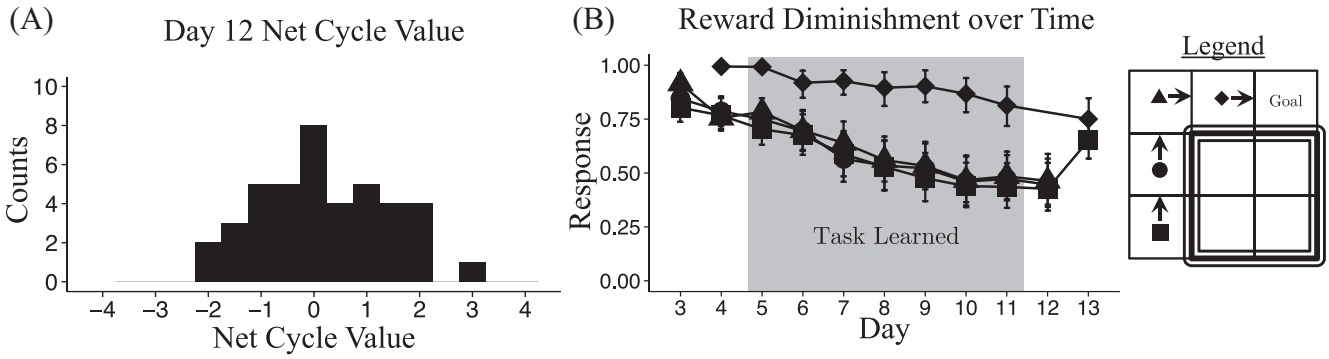

Figure 7. Experiment 2b results. Unlike in Experiment 2a, in which the target task was performed perfectly for fewer consecutive days, here it is clear that rewards eventually decrease over time. By Day 12, the net extreme cycle value is not detectably different from zero, as shown in Panel A. However, as shown in Panel B, while positive feedback to the dog becomes lower for all state-actions, it does not entirely disappear. Error bars are bootstrapped 95% confidence intervals.

directly in these experiments is the extreme, six-step cycle predicted by a strict version of the action-signaling model. We find that after 7 game days, the net reward for this cycle is no longer significantly different from zero, however, this does not rule out the possibility of other positive reward cycles. Moreover, from the perspective of using feedback as incentives, positive reward cycles can be easily avoided by quickly reducing rewards such that even minimal punishments can offset intermediate rewards. That is, if people were actively removing positive reward cycles, they could simply ensure the cycle results in a negative reward.

In addition, we find that people can reliably track a learner's policy with respect to a target policy. This supports the assumption found in much of the teaching literature that that teachers track learners' ability and changes in ability—that is, improvement (Wood, Bruner, & Ross, 1976). However, we also find that people more accurately track ability than improvement based on the same information. This may be because more observations (i.e., learner actions) are needed to update a representation of improvement as compared to ability. More broadly, previous work has used similar formalizations to model how people reason about others' mental states, such as beliefs and goals (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). The correspondence between judgments and $1 - \varepsilon$ suggest that models of social reasoning can be expanded to include how people reason about an agent's ability and improvement by modeling ability as distance to a target policy and improvement as convergence to that policy. One possibility is that ability judgments are based on the distance between the learner's policy (as inferred by the teacher) and the teacher's desired policy (improvement as the change in this value). We leave further work on how people judge the ability of other agents to future work (see Gerstenberg & Goodman, 2012, for a promising approach).

Experiment 3: Interactively Teaching with Reward and Punishment

Experiments 2a and 2b tested how teaching proceeds when everything goes as planned (i.e., when the learner learns what the teacher wants). However, a learner may not learn the desired policy for various reasons, such as misunderstanding the teacher's intention, exploiting rewards, or misinterpreting 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 might adapt their feedback such that the exploited net positive cycles are eliminated.

In this experiment we investigated how people trained learners implemented with reward-maximizing or action-signaling algorithms. This provides additional confirmation that people can train action-signaling agents, but more importantly, it allows us to determine whether participants will adapt their teaching strategy in response to positive reward cycles being exploited. To understand how people interactively teach with reward and punishment, we investigated how participants would train several classes of learning agents under different settings. We ran three different experiments with differing parameterizations and framings. In Experiment 3a, learning agents depicted as dogs performed actions according to the currently learned policy 80% of the time and otherwise explored by selecting an action at random. Although random exploration can facilitate learning, it makes it more difficult for a teacher to track the state of learning. To see if similar results would appear without random exploration, Experiment 3b also depicted learners as dogs, but the algorithms never took random exploratory actions.

Finally, all of our experiments so far have focused on people teaching dogs. One might wonder if similar results would obtain when people are teaching children with rewards and punishments. This would lend support for a more general representation for teaching with rewards and punishments. To examine this, Experiment 3c had the same design as 3a, but with learning agents depicted as preschool children.

Reward-Maximizing and Action-Signaling Conditions

How learning unfolds in real time depends crucially on the specific learning algorithm a teacher is interacting with. Thus, in all three studies, we ran six between-participants conditions with different learning algorithm implementations. Four of these were reward-maximizing algorithms and two were action-signaling algorithms. For the reward-maximizing conditions, we chose to focus on two classes of algorithms that have been widely investigated in human and animal work on decision-making and value learning: Model-based and model-free algorithms (Dayan & Niv, 2008). These two classes of learning mechanism have been used to model goal-directed and habitual learning in the human brain, respectively (Gläscher, Daw, Dayan, & O'Doherty, 2010). Model-free algorithms roughly correspond to habitual learning since they learn about reward contingencies and the future value of actions through trial and error. In contrast, model-based algorithms explicitly reason about state transitions and outcome structure of the environment, allowing them to engage in goal-oriented planning. During learning, this means that given a schedule of feedback, a model-free learner will acquire the reward-maximizing policy more slowly, as it must learn the correct "habit" at each state. A model-based learner, by contrast, can immediately update its behavior for all states and optimally plan if it detects changes to feedback.

The learning dynamics of model-free and model-based algorithms are also both sensitive to what their default expectations about the world are—that is, how they are initialized. For instance, if a reward-maximizing learner is optimistically initialized, it will assume unvisited states have high value and want to test them before exploiting visited states. In contrast, to a neutrally initialized learner, unvisited states to have a low or no value. As a result, it will "take what it knows it can get" and tend to exploit actions that it has already experienced as valuable. Previous work has shown that people tend to have an optimism bias when evaluating potential events (Weinstein, 1980) and that this can help rational decision-makers with limited resources perform better in some environments (Neumann, Rafferty, & Griffiths, 2014). In RL, optimistic initialization enables an agent to more efficiently learn the optimal reward-maximizing policy (Sutton & Barto, 1998). To determine how these dynamics affect teaching by evaluative feedback, we included neutral and optimistic initialization conditions of both the model-free and model-based algorithms.

The action-signaling algorithms implement Bayesian inference, as described in the General Methods section. Recall that the model can be expressed as update that combines the feedback likelihood and policy prior: $P(\pi^* | h) \propto P(h | \pi^*)P(\pi^*)$. The prior over policies encodes the learner's expectations over possible policies that a

teacher could be attempting to teach, and therefore plays an important role in learning. In the studies here, we tested two different policy priors. The first is a simple uniform prior over all policies, which considers all possible policies equally likely (uniform prior condition). This assumes very little about the structure of the environment or the teacher's representations. The second prior places a stronger constraint on target policies: only policies optimal for a world with any combination of +1, −1, and 0 reward-valued tiles are considered (state-reward prior condition). This second prior builds naturally on the state-action structure of the task and similar approaches have been used in related machine learning work (Loftin et al., 2014). At the end of this article, we discuss the possibility of having even richer policy priors that would permit hierarchical learning in teaching by evaluative feedback settings.

Summary

In short, we report the results of three studies, each of which has six conditions: optimistic and neutral model-free, optimistic and neutral model-based, and uniform/state-reward prior action-signaling. Figure 8 shows how these learning algorithms relate to one another, and for Experiment 3a, we report simulations to illustrate their respective learning dynamics under idealized teachers. Details of all the implementations are included in Appendix B.

Experiment 3a: Teaching Exploring Dog Learners

In Experiment 3a, participants trained learners who learned a policy in response to feedback. The agents executed the learned policy 80% of the time and otherwise explored by selecting an action at random.

Method.

Participants and materials. One-hundred and 80 Amazon Mechanical Turk workers (83 female, 96 male, one other) participated in the experiment.

Participants were placed into one of six conditions: neutral model-free, optimistic model-free, neutral model-based, optimistic model-based, uniform prior action-signaling, or state-reward prior action-signaling. In the model-free conditions, we used a standard model-free algorithm—Q-learning with replacement eligibility traces (Singh & Sutton, 1996)—with a learning rate (α) of 0.9 and an eligibility trace decay rate (λ) of 0.5. The neutral model-free algorithms were given an initial value function with all entries set to 0, while optimistic model-free learners were given one with all entries set to +3.5. These initialization values were chosen after testing their performance in simulations to ensure that such agents could be taught within the constraints of the experiment.

In the model-based conditions, learners are given the transition function of the environment and use it to speed learning by generalizing the rewards given by the participant for different actions using the world's transition function. State-action reward values were updated with a learning rate of 0.9. The neutral model-based learners were initialized with a reward function with all entries set to 0.0, whereas the optimistic model-based learners had reward entries set to +1.0. On each trial, the current value function was calculated using value iteration (Sutton & Barto, 1998). These parameter values were also chosen by finding Those That Could × Trained × Simulated Teachers.

The learning algorithms of the two action-signaling conditions used different initial priors over reward functions and updated their beliefs of the target policy as specified in Appendix B. The uniform signaling condition had a uniform distribution over all possible target policies π^*(2592 unique policies), whereas the state-reward signaling condition has a policy distribution derived from a uniform distribution over state reward functions (398 unique policies). The feedback-likelihood function used was sigmoidal with the slope parameter (κ) set to 2.

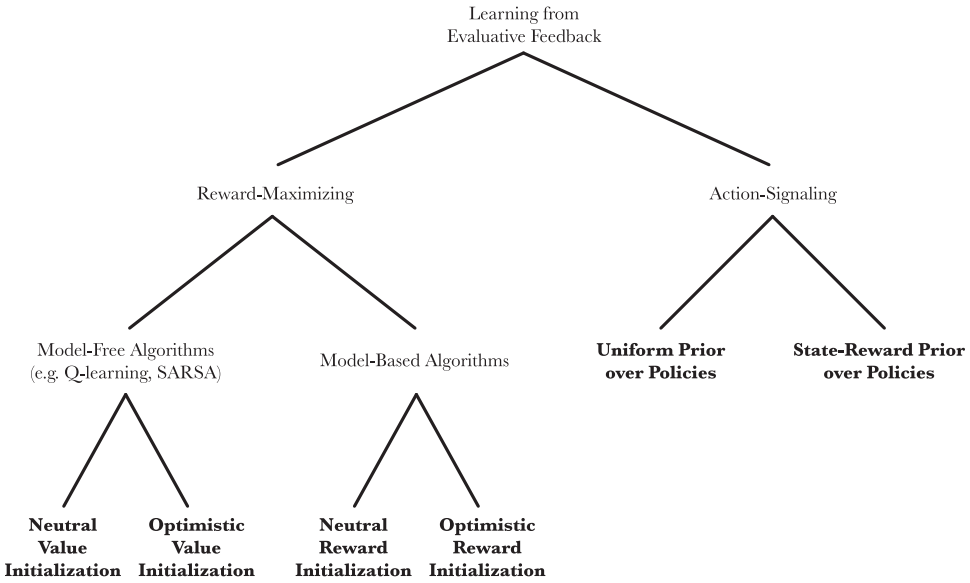

Figure 8. Types of learning from evaluative feedback that includes the conditions implemented in Experiments 3a, 3b, and 3c.

We implemented each algorithm in the browser. Because participants primarily used the anchored values in the previous experiments, the response interface was simplified from the continuous slider to five buttons corresponding to *shock*, *scold*, *do nothing*, *praise*, and *biscuit*. Otherwise, the interface was the same as the one used previously.

Procedure. Participants trained a virtual dog for 10 game days and each ended after 10 steps or once the dog reached the house. They were told the dog would learn the task as they gave it feedback and that it 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. On each trial, the learner acted according to a ε-greedy policy, selecting an optimal action as dictated by their current policy with probability 0.8 and chose a different valid action with probability 0.2/(# of suboptimal valid actions). Participants gave feedback and the algorithms then updated the learned policy with the participant's response before performing the next action. Experimental procedures were approved by Brown University's Research Protection Office (protocol #1505001248, title: "Exploring human and machine decision-making in multi-agent environments").

Simulations. To provide a baseline for interpreting the human experiment, we tested several simulations of this paradigm. For each learning algorithm, we simulated 1,000 teachers with one of two feedback policies that did not change over the course of training. The first was a characteristically incentivizing feedback strategy that punished moderately for entering the garden or walking backward on the path and rewarded highly when the goal was entered via the path. The second was a characteristically signaling feedback strategy that rewarded moderately for leaving the garden or going along the path, rewarded highly when the goal was entered via the path, and punished moderately otherwise (see Figure 9). These patterns of feedback were given to the same learning algorithms (model-based, model-free, and action-signaling) as human participants with the same probability of randomly choosing actions (0.2). This allowed us to anticipate the effectiveness of different teaching strategies with different learners, confirm that teaching is possible within the constraints of the experiment, and qualitatively analyze each model's consistency with participant teaching behavior.

In particular, we note that the incentivizing strategy is able to teach all of the different learning algorithms, including the action-signaling agents. Among the different reward-maximizing algorithms, there are differences in the speed of learning (e.g., neutrally initialized agents learn faster than optimistically initialized ones, and model-based agents typically learn faster than model-free ones). However, they all eventually learn the target policy. In contrast, the signaling teaching strategy is primarily effective for action-signaling agents. The neutrally initialized, model-free agent performs noticeably better than the other reward-maximizing

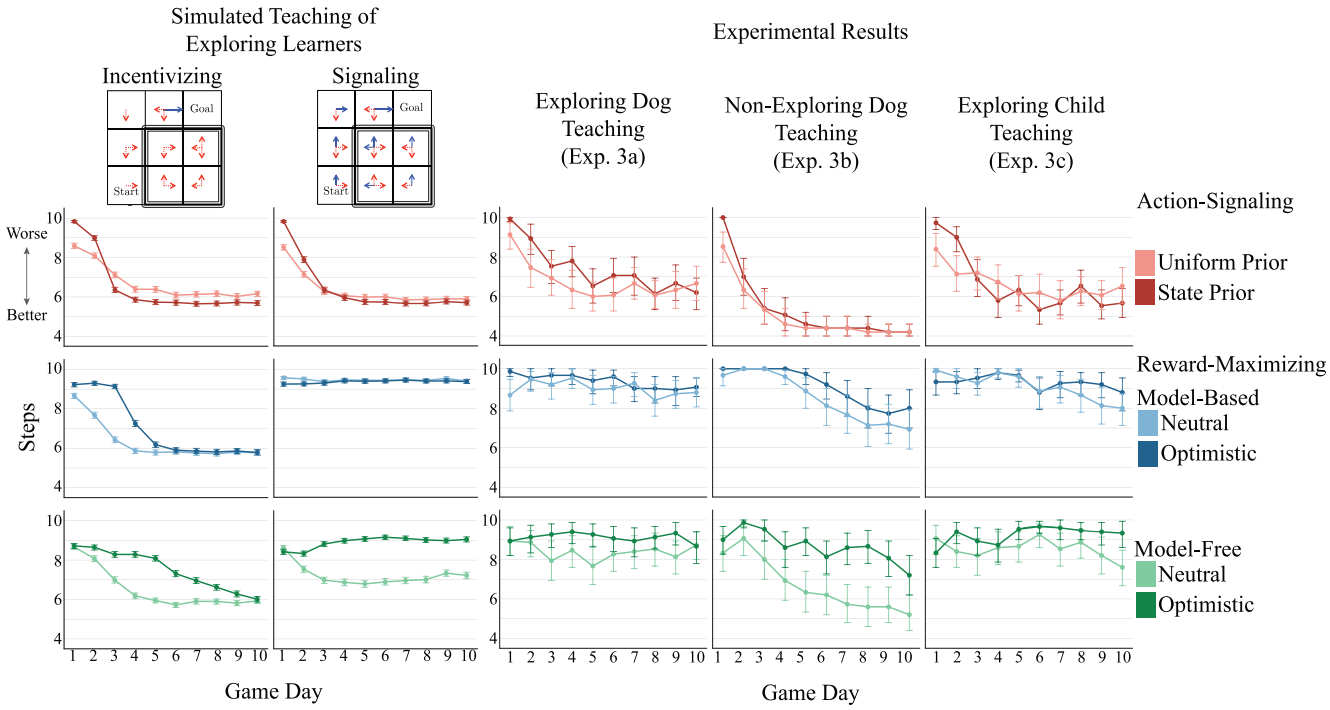

Figure 9. Steps-per-day by day of simulation (first two columns) and Experiments 3a, 3b, 3c (last three columns). Simulated teachers gave (stationary) feedback to the six types of learners in either a characteristically incentivizing or signaling manner. This demonstrates that given the right pairing of teacher and learner strategy, it is possible to successfully teach within the parameters of our task. In all three experiments, people easily taught the action-signaling agents (top row) but often struggled to successfully teach the reward maximizing agents (bottom two rows). This is consistent with their use of feedback as signals rather than simply to incentivize (column 2). See the online article for the color version of this figure.

learners because it exploits known rewards and does not explore the space as much as its counterparts.

Results.

Final trained behavior. Participants could effectively train both action-signaling learners but were unable to teach the complete task to any of the reward-maximizing learners (model-free or model-based with neutral or optimistic initializations). This was due to the persistence of positive reward cycles in participants' feedback that reward-maximizing agents were able to exploit.

Figure 10A shows two representative example trajectories produced during the ninth day of the experiment that exemplify the difference between the reward-maximizing and action-signaling conditions. Here we further analyze the differences in final trained behavior between the different conditions by looking at whether participants could teach agents the subtask of staying on the path and teach the entire task.

We first confirmed that participants in all conditions could teach learners the subtask of walking along the path (but not necessarily

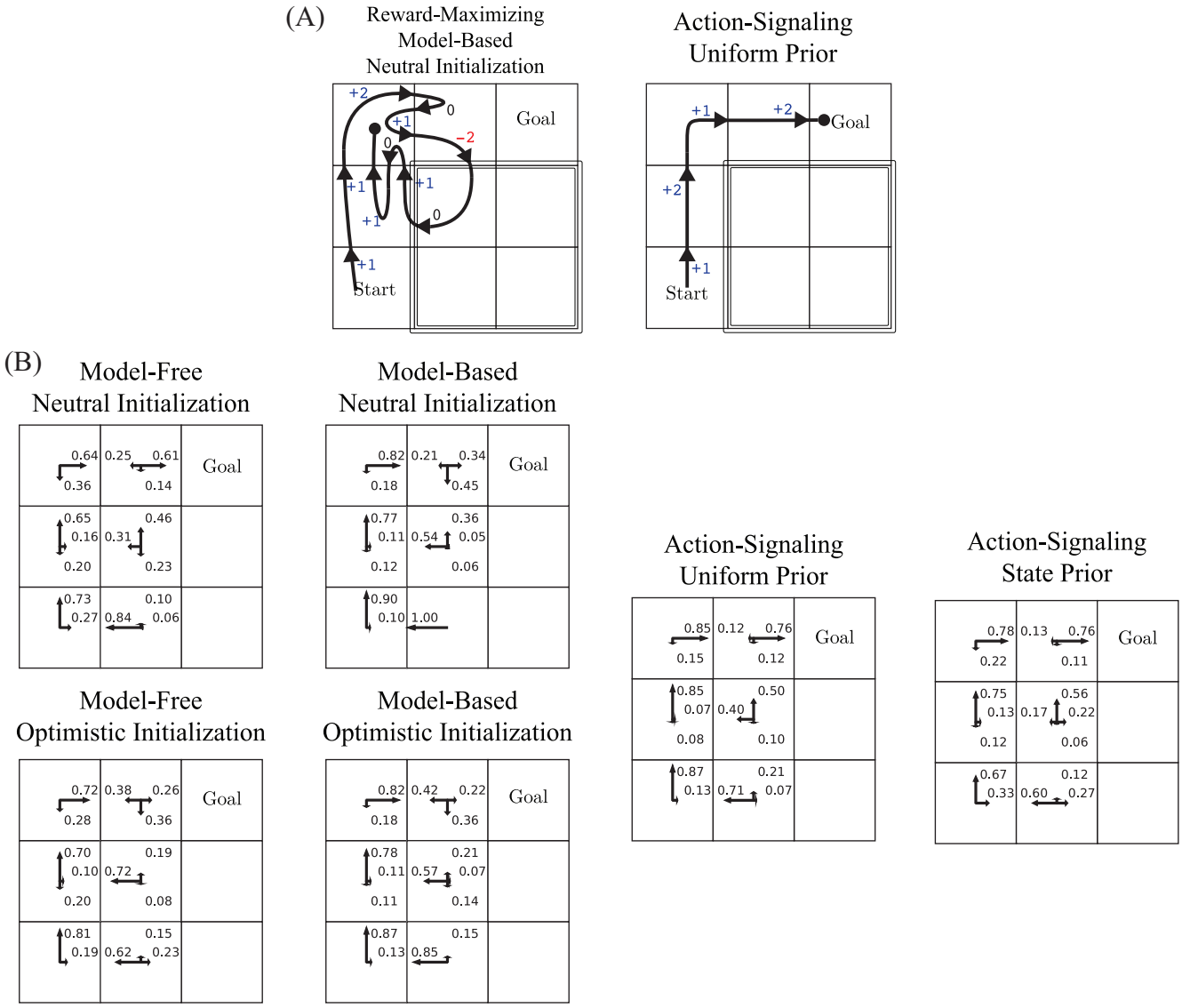

Figure 10. Experiment 3a. A: Example experimental trajectories from the ninth game day of training for one participant in the action-signaling (uniform prior) condition and one from the model-based (neutral initialization) condition. Curves with arrows represent steps taken by the learner over that day, and numbers indicate teacher feedback. Note how intermediate rewards are informative for the action-signaling learner but are exploited by the model-based learner seeking to maximize rewards. B: Proportions of learner actions from path and path-adjacent states by condition (averaged over participants) on the final day of training. By the final day, the dogs generally learn to stay on the path, but only those trained in the action-Ssgnaling conditions consistently complete the task by entering the goal (home) state. Note the different distributions of actions in the state to the left of the goal state. See the online article for the color version of this figure.

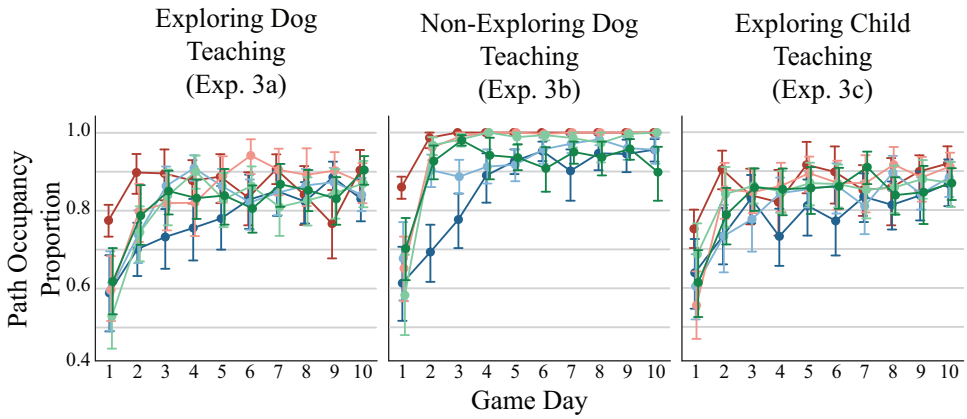

Figure 11. Path occupancy for Experiments 3a, b, and c. Proportion of path occupancy by day showing that participants in different conditions were equally successful at teaching learners to stay on the path (but not necessarily reach the goal). Error bars are bootstrapped 95% confidence intervals. See the online article for the color version of this figure.

entering the house) by the final day (see Figure 11). To do so, we investigated what proportion of states were path states (i.e., the "path occupancy") by day and condition for the last 4 days of training. We fit mixed effects logistic regression models with intercept and day as a random effects across participants, and day and condition as fixed effects. Including condition as a fixed effect did not improve the model according to a likelihood ratio test, $\chi^2(5) = 4.69$, $p = .45$. This indicates that in the final four days, success in teaching the learner to stay on the path was not affected by condition.

Given that the learner stays on the path, fewer steps-per-day is a measure of teaching success. The design of the task limits steps-per-day to being between 4 and 10, as well as even numbers due to the topology of the task. Figure 9 shows steps-per-day for each condition in Experiment 3a. To analyze differences in teaching the entire task, we operationalized training success on a day as steps-per-day being less than seven.[4] We then fit a mixed effects logistic regression model with the same random and fixed effects as in the previous path occupancy model and tested the estimated parameters with Wald Z tests. Planned contrasts were between the action-signaling and reward-maximizing conditions, between model-based and model-free reward-maximizing conditions, between neutral and optimistic initialization reward-maximizing conditions, and between uniform and state prior action-signaling conditions. These contrasts were chosen to assess the primary distinction between using rewards as signals versus incentives as well as to determine additional differences between the subtypes (see Figure 8). Participants in the action-signaling conditions were more likely to successfully teach the entire task than those in the reward-maximizing conditions ($\beta = 0.35$, $SE = 9.3 \times 10^{-2}$, Wald $Z = 3.76$, $p < .001$) and were more able to do so over time (Action-Signaling/Reward-Maximizing × Day: $\beta = 7.0 \times 10^{-2}$, $SE = 1.8 \times 10^{-2}$, Wald $Z = 3.80$, $p < .001$).

Within the reward-maximizing conditions, participants teaching model-based learners were less successful than those teaching model-free learners ($\beta = -0.48$, $SE = 0.19$, Wald $Z = -2.61$, $p < .01$), and those teaching neutral learners were more successful than those teaching optimistic learners ($\beta = 0.46$, $SE = 0.19$, Wald $Z = 2.47$, $p < .05$). Finally, between the two action-signaling conditions, those teaching learners with a uniform prior over policies were more successful than those teaching state prior learners ($\beta = 0.58$, $SE = 0.20$, Wald $Z = 2.87$, $p < .01$). Fixed effect parameter estimates are summarized in Table 2.

To summarize, an analysis of trained behavior in Experiment 3a reveals that teachers can effectively teach intermediate actions to all learners, but only those teaching action-signaling learners could teach the full task. Specifically, rather than completing the task, the reward-maximizing learners would exploit teacher's path rewards without entering the house.

Feedback for intermediate actions over time. A key question is whether participants change their feedback in response to the type of agent they are teaching. In particular, if they are sensitive to intermediate rewards being exploited, they would reduce those rewards or change their distribution over actions (see Figure 12). We compare intermediate rewards over time between conditions and find that participants do not reduce them more when teaching reward-maximizing agents. Indeed, we find that there is a slight bias to give model-based agents more rewards for intermediate actions.

To determine change in intermediate rewards, we analyzed participant responses for the first three target actions over the course of the experiment. We used a mixed effects linear model that included feedback as a dependent variable; day, intermediate action (coded 0 to 2 by index in the target action sequence not including the final action), and their interaction as random effects across participants; and day, intermediate action, day/action interaction, and condition as fixed effects. Including condition as a fixed effect improved the fit of the model according to a likelihood ratio test, $\chi^2(5) = 11.26$, $p < .05$. More complex interactions, including any interaction between condition and day, were not significant nor was the Day × Action interaction, $\chi^2(1) = 2.56$, $p = .11$, and so they were not included. Planned contrasts (the same as in the task success analysis) and tests of significance using

[4] This analysis was additionally done with the success cutoff as five and nine steps and yielded qualitatively similar results.

Table 2
Experiment 3: Estimated Fixed Effects for Logistic Regression Models of Training Success (Less Than 7 Steps in Game Day)

	Exploring dog (3a)			Nonexploring dog (3b)			Exploring child (3c)		
Coefficient	β	*SE*	Wald *Z*	β	*SE*	Wald *Z*	β	*SE*	Wald *Z*
Intercept	−1.79	.16	-11.35^{***}	−6.84	.93	-7.38^{***}	−1.62	.15	-10.91^{***}
Day	.13	2.8×10^{-2}	4.62^{***}	1.57	.23	6.70^{***}	8.8×10^{-2}	3.3×10^{-2}	2.71^{**}
AS vs. RM	.35	9.3×10^{-2}	3.76^{***}	1.71	.39	4.38^{***}	.40	8.9×10^{-2}	4.56^{***}
MB vs. MF	−.48	.18	-2.60^{**}	−5.21	1.11	-4.72^{***}	−.79	.18	-4.33^{***}
Neu. vs. Opt.	.44	.18	2.41^{*}	2.00	.99	2.01^{*}	−.36	.18	-2.00^{*}
U. vs. S. prior	.57	.20	2.81^{**}	1.38	.55	2.54^{*}	.39	.18	2.10^{*}
Condition	.23	.37	.62	3.91	1.99	1.97^{*}	−.46	.36	−1.25
Day × AS vs. RM	7.0×10^{-2}	1.8×10^{-2}	3.82^{***}	1.03	.19	5.46^{***}	.11	2.2×10^{-2}	4.86^{***}
Day × MB vs. MF	3.5×10^{-2}	3.3×10^{-2}	1.02	.19	.17	1.11	.13	3.9×10^{-2}	3.35^{***}
Day × Neu. vs. Opt.	-2.1×10^{-2}	3.3×10^{-2}	−.57	.20	.16	1.27	.15	4.0×10^{-2}	3.76^{***}
Day × U. vs. S. prior	-7.5×10^{-2}	3.9×10^{-2}	−1.85	−.41	.38	−1.08	−.10	4.7×10^{-2}	-2.09^{*}
Day × Condition	-5.2×10^{-2}	6.6×10^{-2}	−.77	−1.25	.36	-3.48^{***}	7.9×10^{-2}	7.9×10^{-2}	.09

Note. Experiment 3a: $N = 180$ participants, 10 game days (observations) per participant. Nagelkerke's $R^2 = .104$. Experiment 3b: $N = 180$ participants, 10 game days (observations) per participant. Nagelkerke's $R^2 = .260$. Experiment 3c: $N = 180$ participants, 10 game days (observations) per participant. Nagelkerke's $R^2 = .37$. AS = action-signaling; RM = reward-maximizing; MB = model-based reward-maximizing; MF = model-free reward-maximizing; Neu = neutrally initialized reward-maximizing learners; Opt = optimistically initialized reward-maximizing learners; U. = uniform prior; S. = state-based prior.
$^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Satterthwaite's approximation showed no difference between the action-signaling and reward-maximizing conditions, $\beta = 1.7 \times 10^{-3}$, $SE = 1.0 \times 10^{-2}$, $(177.58) = -0.17$, $p = .86$. Instead, it was model-based algorithms that received greater rewards than model-free algorithms, $\beta = 5.2 \times 10^{-2}$, $SE = 1.8 \times 10^{-2}$, $t(179.92) = 2.89$, $p < .01$). The full model and parameter estimates are reported in Appendix C.

In short, there is no evidence that participants lowered their rewards based on whether the agent they were teaching was designed to exploit their rewards. Indeed, we find they gave greater rewards to model-based algorithms that then even more adaptively maximize rewards.

Experiment 3b: Teaching Nonexploring Dog Learners

In Experiment 3a, the learning algorithms randomly explored with some small probability. This facilitates learning but might make teaching more difficult because it adds noise to the learner's policy. To determine if participants would show similar teaching strategies when actions more directly reflected the learned policy,

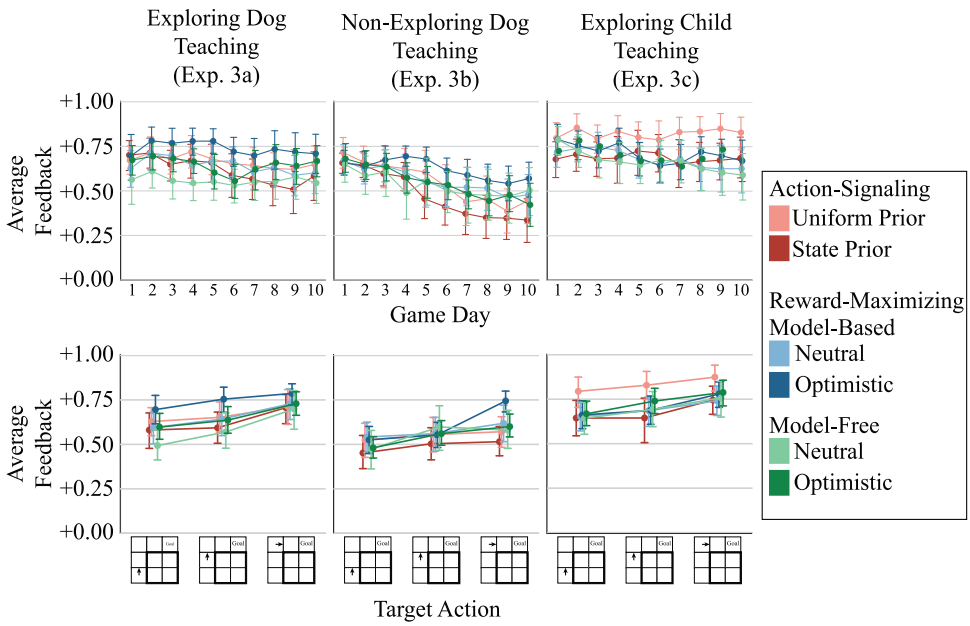

Figure 12. Experiment 3a, 3b, and 3c: Average feedback by participant for intermediate target actions over the course of the experiment (top row) and by action (bottom row). Error bars are bootstrapped 95% confidence intervals. See the online article for the color version of this figure.

Experiment 3b used the same design as 3a, but the algorithms always took the best action as dictated by the currently learned policy.

Method. One hundred and 80 (83 female, 95 male, two other) Amazon Mechanical Turk workers participated in the study with 30 participants assigned to each of the six conditions. The materials and procedure used in this study were identical to those in the previous study except agents always selected actions from the currently learned policy. Experimental procedures were approved by Brown University's Research Protection Office (protocol # 1505001248, title: "Exploring human and machine decision-making in multi-agent environments").

Results.

Final trained behavior. As in Experiment 3a, participants in all conditions were able to teach the intermediate task of staying on the path by the final 4 days, whereas those in the action-signaling conditions were more successful at training the entire task. There were additional differences that we note below.

We analyzed path occupancy for the last four episodes for the reward-maximizing conditions only. This is because in the action-signaling condition, path occupancy was at ceiling (see Figure 11). As in the analysis for Experiment 3a, we used a mixed effects logistic regression with path occupancy as a binary variable; intercepts and day as random effects across participants; and condition and day as fixed effects. The intercept value indicated that overall, path tiles were 27.4 times more likely to be occupied than not ($\beta = 3.32$, $SE = 1.64$, Wald $Z = 2.02$, $p < .05$). In addition, using the same planned contrasts as in Experiment 3a, we found neutrally initialized agents learned to stay on the path better than optimistically initialized learners ($\beta = 0.82$, $SE = 0.23$, Wald $Z = 3.12$, $p < .001$).

We analyzed training success using the same model as in Experiment 3a: *Success* was defined as the number of steps in a day being less than seven; day was a random effect across participants; condition and day were fixed effects; planned comparisons were between action-signaling versus reward-maximizing, model-based versus model-free, and uniform versus state-prior conditions; and tests of significance were performed using Satterthwaite's approximation. Parameter estimates and tests are reported in Table 2. In particular, we found that, as expected, participants in the action-signaling conditions were more successful at teaching the complete task than those in the reward-maximizing conditions over the course of the game days (Action-Signaling/Reward-Maximizing × Day: $\beta = 1.03$, $SE = 0.19$, Wald $Z = 5.46$, $p < .001$).

Feedback for intermediate actions over time. Even without random learner exploration, we found little evidence that participants gave less rewards to reward-maximizing agents. Rather, if anything, there was a slight bias to give greater rewards in the reward-maximizing conditions for intermediate actions closer to the end of the target sequence.

We fit a mixed effects linear regression model to feedback data. Likelihood ratio tests showed that for fixed effects, day, $\chi^2(1) = 381.02$, $p < .0001$, action, $\chi^2(1) = 233.38$, $p < .0001$, and a Day × Action interaction, $\chi^2(1) = 5.25$, $p < .05$, were significant. Condition, $\chi^2(5) = 32.72$, $p < .0001$, and its interaction with intermediate action, $\chi^2(5) = 23.38$, $p < .001$, were also significant, but neither the interaction between condition and day, $\chi^2(5) = 6.02$, $p = .30$, nor the full three-way interaction, $\chi^2(5) = 6.11$, $p = .30$, were significant. Our model thus included condition and its interaction with intermediate action, but no additional interactions.

Tests of significance with Satterthwaite's approximation showed no difference between action-signaling conditions and reward-maximizing conditions ($\beta = 1.2 \times 10^{-2}$, $SE = 1.3 \times 10^{-2}$, $t(185.47) = .91$, $p = .36$). However, those in the reward-maximizing conditions tended to give higher rewards for later actions (Action × Action-Signaling/Reward-Maximizing contrast: $\beta = -2.0 \times 10^{-2}$, $SE = 4.7 \times 10^{-3}$, $t[178.72] = -4.36$, $p < .001$). All parameter estimates are reported in Appendix C.

Experiment 3c: Teaching Exploring Child Learners

The previous studies all required participants to train a virtual dog. We wanted to ensure that our experimental results supporting the evaluative feedback as communication hypothesis generalizes beyond teaching dogs and to teaching other humans. Experiment 3c thus uses the same design as Experiment 3a but adapted to a child learner rather than a dog. (Of course, in both cases, participants were aware that they were training virtual dogs and children rather than actual ones).

Method. One hundred and 80 Amazon Mechanical Turk workers (80 female, 99 male, one other) participated. The same six conditions and parameter values for the learning algorithms from Experiment 3a were used.

Given that the agent was now a child, we used a different cover story for the task. Rather than having the agent be a dog, it was a 4-year-old boy named Alex. Participants were told that Alex liked to play in the mud, and so he would often track dirt into the house when he walked around inside. Analogous to the previous studies, the goal of the task was to teach Alex to go to the bathroom by walking along the wood floor and not on the white carpet, which is more difficult to clean. Finally, the options for evaluative feedback were changed to harsh scolding, mild scolding, doing nothing, mild praise, and high praise. The interface is shown in Figure 1.

We told participants they would teach Alex over 10 game days, each of which ended after Alex took 10 steps or if Alex reached the bathroom. Alex learned the task as he received feedback and would appear at the living room entrance (left-bottom) at the beginning of each day. As before, they were told that their bonus was contingent on how well Alex performed on his own after the 10 days (though all participants were paid the full bonus). Following the task, several questions were asked about the experiment and the participant's background. Experimental procedures were approved by Brown University's Research Protection Office (protocol #1505001248, title: "Exploring human and machine decision-making in multi-agent environments").

Results.

Final trained behavior. Similar to Experiments 3a and 3b, participants in all conditions were able to teach staying on the path and those teaching action-signaling learners performed better at the task than those teaching reward-maximizing learners. We also observed several significant contrasts within the set of action-signaling conditions and set of reward-maximizing conditions, as shown in Table 2. Specifically, in this final study, there was steeper learning over time in the model-based conditions compared to the model-free conditions, the neutral reward-maximizing conditions compared to the optimistic reward-maximizing conditions,

and the state-based action-signaling condition compared to the uniform action-signaling condition.

The same analysis as in Experiment 3a was used to evaluate path occupancy and training success. We fit a mixed-effects logistic regression model with intercept and day as a random effect across participants, and day, condition, and their interaction as fixed effects. Neither fixed effects for day, $\chi^2(1) = 3.63$, $p = .06$, condition, $\chi^2(5) = 5.53$, $p = .35$, nor their interaction, $\chi^2(5) = 5.96$, $p = .31$, were significant, indicating that there were no detectable differences across conditions in terms of teaching the learners to stay on the path.

Consistent with our previous results, people were less likely to successfully train reward-maximizing agents and people were even less successful at teaching model-based agents. We analyzed training success as in Experiment 3a: We fit a mixed-effects logistic regression model with steps-per day being less than seven as the binary outcome for success; random effects of intercept and day across participants; fixed effects of day, condition and their interaction; the same planned contrasts; and tests with Satterthwaite's approximation. Training success was more likely for action-signaling than reward-maximizing conditions overall ($\beta = 0.40$, $SE = 8.9 \times 10^{-2}$, Wald $Z = 4.56$, $p < .001$) and as a function of day ($\beta = 0.11$, $SE = 2.2 \times 10^{-2}$, Wald $Z = 4.86$, $p < .001$). Training success with model-based was less than with model-free agents ($\beta = -0.79$, $SE = 0.18$, Wald $Z = -4.33$, $p < .001$), although the slope by day was steeper in the model-based conditions ($\beta = 0.13$, $SE = 3.9 \times 10^{-2}$, Wald $Z = 3.35$, $p < .001$). Additional parameter estimates and test results are reported in Table 2.

Feedback for intermediate actions over time. As with Experiments 3a and 3b, we found that participants did not reduce the rewards they gave for intermediate target actions more in reward-maximizing conditions than in the action-signaling conditions. We selected mixed-effects linear models with day, intermediate action, and their interaction as random effects across participants; and day, intermediate action, their interaction, and condition as possible fixed effects. The fixed effects of day, $\chi^2(1) = 83.41$, $p < .0001$, and action, $\chi^2(1) = 234.44$, $p < .0001$, but not their interaction, $\chi^2(1) = 0.17$, $p = .68$, were significant. Condition and higher-order interactions involving condition were not significant, all $\chi^2(5) < 0.23$, $p > .28$, indicating that patterns of feedback did not detectably differ between conditions. See Appendix C for parameter fits.

Discussion

Several consistent patterns emerge from Experiments 3a, 3b, and 3c. First, participants reliably produce positive reward cycles that a reward-maximizing agent will learn to exploit. This results in reward-maximizing learners failing to learn the entire task, even though they are always taught the subtask of staying on the path. Participants often give feedback that results in dramatic deviations from the target behavior (see Figure 10). Importantly, reward-maximizing learners do not all fail equally in the presence of positive reward cycles. Neutrally initialized and model-free learners tended to "learn" the target task better than optimistically initialized and model-based learners. However, it is worth noting that although these algorithms are more likely to perform the target sequence, it is only because they are doing a worse job at maximizing rewards: They tend to get stuck in locally reward-maximizing policies. In contrast, optimistic initialization helps ensure that a learner will find the globally best solution to a problem (Sutton & Barto, 1998). Model-based learners (that have a correct model) have more powerful learning processes that can leverage knowledge of environment dynamics to accomplish their goals (Brafman & Tennenholtz, 2003). How well an algorithm learns to maximize reward thus seems to be inversely related to participants' success at training it. This suggests that the principle of reward maximization does not guide their choice of teaching strategy.

Second, we did not find evidence that people adapt their feedback strategy while interactively teaching reward-maximizing agents. If people realized a learner was treating their feedback as incentives and updating how those incentives were being interpreted, they could decrease rewards for learned actions. This does not occur, even in the absence of random learner exploration (3b). Indeed, in Experiments 3a and 3b we observed a different pattern: People gave more rewards to reward-maximizing learners. Note that the latter are better at learning to exploit positive reward cycles than model-free agents. This indicates that when faced with agents that maximize feedback, people persevere in using feedback communicatively even though this strategy is not effective.

Finally, we also found that within the action-signaling conditions, people trained the uniform prior learner more quickly than the state-reward prior learner. This is because although a state-reward prior is effective in some tasks (Loftin et al., 2014), in our particular experiment it initially biases the learner away from the target task and toward ones where path tiles are rewarding and not simply neutral. That is, the state-reward prior assumes that certain locations are either worth $+1$, -1, or 0, and as a result needs to be explicitly taught that only the goal state is rewarding. Nonetheless, because the learner is doing policy inference based on feedback as signals, this initial bias can be quickly overcome, allowing the agent to eventually learn the task just as well as one with a uniform prior over policies.

To summarize, across different agent types (virtual dogs and children) and across different parameterizations of reward-maximizing learning algorithms, teachers persist in producing positive reward cycles. This indicates people use rewards and punishments as signals, which is effective at teaching action-signaling learners but not reward-maximizing learners. Overall, this provides further evidence that people teach using evaluative feedback as communication and not as reinforcement.

General Discussion

In several experiments, we provide converging evidence that people do not teach other agents as if they are maximizing rewards. Instead, people use rewards and punishments in a communicative manner, which leads to positive reward cycles that a reward-maximizing agent would learn to exploit at the expense of learning the complete task. We tested whether people produce positive reward cycles for isolated actions (Experiment 1), during a learner's successful improvement on the task (Experiment 2), and while interacting dynamically with several learning algorithms that implement reward-maximization or action-signaling (Experiment 3). We found strong evidence that people not only produce positive reward cycles but persist in producing positive reward cycles even

when it is not effective. In addition, we found that the use of rewards and punishments as communication generalized across virtual canine and human (children) agents. This indicates that using evaluative feedback as communication may not be specific to the type of agent a person is teaching.

These findings shed new light on how people structure rewards and punishments to teach, which has implications for pedagogy and social learning more broadly. In the following sections, we describe how our results relate to previous research and future directions for investigation.

Reinforcement Learning Approaches to Social Rewards and Punishments

Our main result, that people produce rewards and punishments that are not effective for teaching reward-maximizing agents, relates to existing RL accounts of social rewards and punishments in a number of ways. Much contemporary psychological research models behavior, decision-making, and cognition as processes ultimately driven by learning how to maximize environmental rewards (Dayan & Niv, 2008; Sutton & Barto, 1998). In recent decades, researchers have identified dopamine as a neural substrate for reward prediction errors (Glimcher, 2011; Schultz, Dayan, & Montague, 1997) and developed cognitive and neural models of the relationship between goal-directed, model-based systems and habitual, model-free learning mechanisms (Gläscher et al., 2010; Lee, Seo, & Jung, 2012; Otto, Gershman, Markman, & Daw, 2013). Moreover, a number of studies have shown that areas of the brain involved in processing nonsocial rewards, such as food, water, or money, are similarly activated by social rewards such as signals about reputation, praise, or facial expressions (Izuma et al., 2008; Jones et al., 2011; Lin et al., 2012). This has led some scientists to posit the existence of a "common neural currency" for representing hedonic experience in the brain (Ruff & Fehr, 2014).

There is an apparent tension between our experimental results and the general finding that people show signatures of processing social rewards similar to nonsocial rewards. However, this can be resolved by distinguishing between representing reward in the brain and learning about reward from different sources, whether they are nonsocial, social, or pedagogical. In the nonsocial case, rewards are experienced from the environment directly and can be learned directly. An algorithm can thus maximize the reward signal itself, as in standard RL (Sutton & Barto, 1998). In contrast, when rewards are generated socially or pedagogically, they are mediated by the internal representations and intentions of the social partner who is producing them. An algorithm would take that mediation into account while learning, and then represent what has been learned as if it were learned directly from the environment. In other words, an agent needs to engage in inference based on the reward signal. Whether and how this inference process proceeds in people is then an empirical question.

A key contribution of our work is then the following: We explore how a teacher's communicative intentions and representations mediate the use of rewards and punishments. This investigation complements existing research on how behavior and cognition is shaped by environmental contingencies (Collins & Frank, 2013; Dayan & Niv, 2008; Gershman et al., 2010) as well as how rewards are represented in the brain (Ruff & Fehr, 2014). Future work must explore how these phenomena integrate to form a complete picture of teaching with and learning from rewards.

In addtion, this article provides a methodological contribution to research on reward learning and teaching. Here, we use a novel, multistate task with cyclical state topology. Although some previous work with reward learning has used tasks with multiple states (e.g., the two-step task in Daw, Gershman, Seymour, Dayan, & Dolan, 2011), most studies use single-state bandit tasks in which participants must choose between "slot machines" that have different (sometimes changing) distributions over rewards and punishments. However, as discussed in the modeling section, single-state (or two-state) tasks cannot differentiate between rewards and punishments as reinforcement or signals. Our studies, in contrast, allow for the possibility of cyclical action sequences while maintaining the ability to simulate and analyze behavior based on formal modeling.

Mental State Inference and Teaching

Our experiments investigate how pedagogical rewards and punishments are mediated by a teacher's own representations and intentions. In particular, the results we report here indicate that people do not intuitively teach using rewards as incentives, but rather as signals to indicate that learners are "heading in the right/wrong direction" by labeling their action with respect to a target behavior. Importantly, this presupposes that a learner is capable of reasoning about target behaviors, and therefore how actions relate to one another through plans and intentions. These results thus indicate that pedagogical rewards and punishments attempt to leverage the broader capacity of *theory of mind*, the ability to reason about one's own or another's mental states (Premack & Woodruff, 1978).

Outside of reward and punishment, recent developmental work has highlighted the importance of theory of mind for teaching and social learning. For example, when children learn by demonstration, they make strong inferences about how the observable actions of a demonstrator (e.g., reaching for a cup of water) reflect their mental states, such as unobservable beliefs, desires, or intentions (e.g., believing the cup has water, wanting to drink a cup of water). This capacity for mentalizing facilitates learning about, imitating, and adopting others' goals, traits, and world knowledge such as causal relationships (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Hamlin, Wynn, & Bloom, 2007; Jara-Ettinger et al., 2015; Lyons et al., 2007; Meltzoff, 1995; Powell & Spelke, 2018).

The action-signaling model draws on these principles since it assumes that evaluative feedback is a signal for whether actions are correct with respect to a larger plan of action or intention. Indeed, the action-signaling model can be considered part of a broader class of Bayesian social cognition models. Inverse planning models in this family have been used successfully to model theory of mind in adults and children (Baker et al., 2009; Jara-Ettinger et al., 2015). Although they both fit within the same modeling framework, there is a key difference: In inverse planning models, the participant observes an agent take a sequence of actions and uses that to infer the agent's goals. In the action-signaling models, the learning agent selects actions and then the teaching agent provides evidence for the correctness of the action via evaluative feedback.

The relationship between intention-recognition and learning from social rewards and punishments has arisen in the animal social learning literature as well. Despite theoretical claims that teaching with rewards and punishments would be widespread among nonhuman animals (Caro & Hauser, 1992; Clutton-Brock & Parker, 1995), surprisingly little empirical evidence has borne out this prediction (Raihani, Thornton, & Bshary, 2012; Stevens, Cushman, & Hauser, 2005). Raihani et al. (2012) have suggested that this is in part due to the indeterminacy of learning from rewards and punishments. That is, it is often ambiguous why one is receiving rewards and punishments from another animal. This issue often arises in laboratory shaping in the form of "superstitious behavior" in which an animal learns to perform a causally irrelevant behavior that happened to co-occur with a behavior that caused a reinforcing stimulus (Skinner, 1974). Nonetheless, it is clear that humans reward and punish each other in a variety of contexts (Fehr & Gächter, 2002; Owen et al., 2012). Our finding that people primarily operate in the mode of using rewards and punishments communicatively—that is, under the assumption that the learner can reason about target actions and plans—may reflect an adaptation that overcomes the indeterminacy problem (Ho, MacGlashan, Littman, & Cushman, 2017).

An additional consequence of social rewards and punishments being communicative signals and not reinforcement is that effective interpretation of evaluative feedback relies on theory of mind. The literature on autism distinguishes between accounts that emphasize deficits in theory of mind ability (Baron-Cohen, Leslie, & Frith, 1985) and those that focus on the role of decreased rewards associated with social stimuli (Chevallier, Kohls, Troiani, Brodkin, & Schultz, 2012). These are often treated as competing accounts; however, the motivational influence of social rewards and punishments may rely on the receiver interpreting them using theory of mind representations. Our results suggest that those producing evaluative feedback expect receivers to reason about how actions relate to intentions.

Previous Work on Mental State Inference and Evaluative Feedback

Our work also complements previous work on the role of mental state inference and evaluative feedback. For example, Meyer and colleagues (Barker & Graham, 1987; Meyer, 1982, 1992; Miller & Hom, 1996) reported the "praise/criticism paradox", which occurs when the effects of praise or criticism on a learner's affective state or perceived competence changes depending on the teacher's state of knowledge. They found that participants interpreted a teacher's praise positively and criticism negatively when they believed the teacher did not know their ability, which is consistent with evaluative feedback as reinforcement. But when learners believed the teacher did know their ability on the task, the pattern flipped: Praise was viewed negatively and criticism positively. These results indicate that the interpretation of evaluative feedback is a function of reasoning about a teacher's mental states and not simply their apparent reward. Our model does not capture these results directly because it treats rewards and punishments solely as a function of whether an action is correct or incorrect. However, it could be modified to capture Meyer et al.'s results by having feedback rely on both the correctness of actions and the teacher's belief in their competence. For example, negative feedback for correct actions could be likely only if the teacher believes they are competent.

Related work in moral psychology has also examined the extent to which satisfying one's communicative goals when rewarding and punishing involves recognition by the receiver. For example, when punishing someone who had previously transgressed them, participants were only satisfied if the transgressor signaled that they understood it "as a message" and not simply experienced it as a negative outcome (Funk, McGeer, & Gollwitzer, 2014; Gollwitzer & Denzler, 2009; Gollwitzer, Meder, & Schmitt, 2011). Consistent with our finding that people persist in producing positive cycles even when visibly ineffective, this suggests that people approach using evaluative feedback as a communicative act rather than an opportunity to incentivize or disincentivize certain behaviors.

Finally, an important finding in the educational literature is that extrinsic rewards delivered by a teacher can undermine intrinsic motivation to perform certain activities (see Deci, Koestner, & Ryan, 1999 for a review). In our studies, the learning agents were not provided with any intrinsic motivation (i.e., the only rewards in the task came from human teachers) and participants were not given any indication otherwise. Nonetheless, research on the effects of extrinsic rewards and intrinsic motivation has distinguished between controlling rewards, in which it is clear that the teacher essentially uses feedback to shape behavior, and informational rewards, in which the teacher uses them to indicate competence (Deci et al., 1999; Deci & Ryan, 1980). Positive feedback interpreted as controlling led to decreased intrinsic motivation, whereas positive feedback interpreted as information about the learner's competence enhances intrinsic motivation. Controlling and informative rewards map roughly onto evaluative feedback as reinforcement and communication, respectively. Our work thus suggests that people have a strong bias toward using rewards and punishments informationally.

Child Rearing, Dog-Training, and Robot-Shaping

People often teach with rewards and punishments. When interacting with children, adults use reward and punishment to teach a variety of behaviors such as compliance with rules or social norms (Owen et al., 2012; Rogoff, Paradise, Arauz, Correa-Chávez, & Angelillo, 2003). Similarly, dog owners and dog trainers rely heavily on rewards and punishments in the form of verbal praise or scolding as well as food and electric shocks as we used (virtually) in our experiments (Hiby et al., 2004).

With respect to these interactions, the main implication of our studies is that people are highly biased to use rewards and punishments as communication and do not intuitively use them as reinforcement. This is especially surprising considering that children and dogs are adept at maximizing rewards and suggests that people are either completely ineffective at teaching with evaluative feedback or children and dogs interpret teachers' feedback as communicative (Azrin & Lindsley, 1956; Salzinger & Waller, 1962). Given that people robustly succeeded in teaching action-signaling learners in our studies, it is likely that children and dogs can interpret teachers' feedback as communicative. Moreover, findings on other forms of teaching and social learning support this possibility. For example, in teaching by demonstration and imitation, children will treat a stimulus

produced by a teacher differently based on it being marked as intentionally pedagogical (Bonawitz et al., 2011; Brugger, Lariviere, Mumme, & Bushnell, 2007; Buchsbaum et al., 2011; Sage & Baldwin, 2011). In addition, dogs but not wolves are also sensitive to cues to communicative intent such as eye contact (Téglás, Gergely, Kupán, Miklósi, & Topál, 2012; Topál, Gergely, Erdõhegyi, Csibra, & Miklósi, 2009). Recognition of a teacher's communicative intent could similarly mediate whether dogs or children treat social rewards and punishments as reinforcement or signals.

Finally, even though we do not directly test how people would teach agents that are explicitly designated as artificial intelligence or robots, our results are relevant to machine learning. In recent years, researchers have explored building machines that learn from human rewards and punishment. Some of this work treats it as reinforcement, but positive reward cycles are a common mistake even for those who understand reinforcement learning (Isbell et al., 2001; Ng et al., 1999). They can be overcome by having learning agents myopically privilege immediate rewards over future ones (thus preventing multistate cycles), but this prevents useful generalization to future actions (Knox & Stone, 2015). Some recent work has explored interpreting human evaluative feedback as signals and found that these machines were easier for people to teach (Griffith, Subramanian, Scholz, Isbell, & Thomaz, 2013; Loftin et al., 2014). Our results indicate that this will continue to be a promising direction for future developments as this is more in line with people's natural intuition for teaching nonrobots with evaluative feedback.

Limitations and Future Directions

The work here examines teaching with reward and punishment by contrasting social inference (e.g., Baker et al., 2009; Meltzoff, 1995) and reinforcement learning (Dayan & Niv, 2008; Sutton & Barto, 1998). Our experimental results indicate that people do not readily use rewards and punishments as reinforcements but rather as communicative signals that indicate the correctness or incorrectness of actions. These findings provide a starting point for investigating how people teach with rewards, and here we outline several important limitations of the current work as well as possible future directions.

Specified models and alternative theories. Our modeling strategy in this article has been to characterize the different constraints associated with reward-maximizing and action-signaling teaching. Focusing on broad classes of learning models has allowed us to rule out that people use feedback purely as reinforcements (e.g., via the positive reward cycle analysis in Experiments 1a and 1b, simulation results in Experiment 3a, and results of the interactive paradigm in Experiments 3a–3c) as well as identify additional features consistent with action-signaling (e.g., the state training strategy in Experiment 1a or temporal trajectory of feedback in Experiments 2 and 3). Nonetheless, this approach does not allow us to make strong predictions about fine-grained behaviors (e.g., the probability of slider/button responses) or specific parameters involved in peoples' action-signaling strategies. An important next step then is to develop and contrast more specified models that can predict peoples' strategies across teaching settings.

In addition, although throughout this article we have emphasized two teaching strategies based on reinforcement and signaling, this does not exhaust the space of possible strategies. By design, our experiments create situations in which by pursuing a reasonable action-signaling strategy, people produce a positive reward cycle that a reward-maximizing agent would learn to exploit. We then find that people pursue (and generally stick to) such strategies. However, there may be domains that could tease apart an action-signaling strategy from alternative strategies or that qualify action-signaling in important ways. For example, we have discussed how people sometimes use a state training strategy (Experiment 1a), which is consistent with action-signaling in our paradigm but may not be in an alternative setting.

Relatedly, the two strategies we emphasize are not always mutually exclusive. As we note early in this article, there are situations in which the two strategies are identical (e.g., teaching single actions), and one might expect that in many settings people use a mixture of the two strategies. Indeed, that a reward can signal correctness and punishment signal incorrectness depends on them being inherently reinforcing and aversive, respectively. If rewarding or punishing signals were truly arbitrary signals, then the mapping between correctness and valence could just as easily be reversed, a possibility that is intuitively unlikely. Future modeling work thus needs to explore when and how the communicative use of rewards and punishments derive from their properties as reinforcing and aversive stimuli. Similarly, future work could examine whether certain contexts can prompt people to adapt and take a purely reward-maximizing stance.

Additional settings for teaching and learning. The task used throughout this article was designed with the goal of distinguishing reward-maximization and action-signaling. In addition, by using the same environment map, we have been able to systematically modify various aspects of the teacher–learner interaction and build up toward full-fledged ongoing interaction. There are a number of ways that the task could be generalized to examine how people use rewards. For instance, rather than having only a single agent shape another's behavior, both agents could be engaged in teaching and learning. Most studies of repeated games involve single actions and assume participants treat payoffs purely as incentives (Rand & Nowak, 2013). But as with bandit tasks, this can mask people's use of rewards and punishments as signals. Paradigms in which actions unfold over time (Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016) or participants can reward and punish one another may help elucidate whether people use rewards communicatively in other contexts.

It will also be important to investigate how teaching strategies interact with environments that involve complex representations. The presence of both action-signaling and state-based feedback in Experiment 1 suggests different patterns of feedback can be driven by different expectations of how learners carve up the state space. A teacher and learner may not always share prior knowledge about how states and/or actions should be represented or relate, and evaluative feedback can be used to teach features of states (e.g., the difference between path tiles and garden tiles). In addition, evaluative feedback in everyday settings is often interwoven with other types of teaching such as verbal statements or demonstrations. An important question is

how teachers choose between these different types of teaching and how they are integrated during learning.

Finally, our findings raise the question of whether human or animal learners explicitly recognize teaching rewards and punishments as communicative. Future research will need to test whether this is the case, and if so in what contexts they treat evaluative feedback as reinforcement or communication. We will also need to examine whether the types of mismatches we have identified experimentally (e.g., in Experiment 3) occur in everyday settings. For instance, do parents construct positive reward cycles that children learn to exploit?

Conclusion

We have presented a formal analysis of teaching by evaluative feedback as reinforcement and as communication and have shown that people have a strong tendency toward the latter. This challenges the widely held assumption that people use rewards and punishments to shape behavior. Instead, these findings reveal that evaluative feedback is more similar to other forms of pedagogy like teaching by example or by demonstration (Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016; Shafto et al., 2014). Specifically, people use rewards and punishments communicatively, suggesting a shared representational basis for teaching across seemingly disparate teaching behaviors.

References

Aronfreed, J. (1968). *Conduct and conscience: The socialization of internalized control over behavior*. New York, NY: Academic Press.

Azrin, N. H., & Lindsley, O. R. (1956). The reinforcement of cooperation between children. *The Journal of Abnormal and Social Psychology, 52,* 100–102.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1,* Article 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113,* 329–349. http://dx.doi.org/10.1016/j.cognition.2009.07.005

Barker, G. P., & Graham, S. (1987). Developmental study of praise and blame as attributional cues. *Journal of Educational Psychology, 79,* 62–66.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition, 21,* 37–46. http://dx.doi.org/10.1016/0010-0277(85)90022-8

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120,* 322–330.

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.

Brugger, A., Lariviere, L. A., Mumme, D. L., & Bushnell, E. W. (2007). Doing the right thing: Infants' selection of actions to imitate from observed event sequences. *Child Development, 78,* 806–824. http://dx.doi.org/10.1111/j.1467-8624.2007.01034.x

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

Butler, L. P., & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. *Child Development, 83,* 1416–1428.

Butler, L. P., & Markman, E. M. (2014). Preschoolers use pedagogical cues to guide radical reorganization of category knowledge. *Cognition, 130,* 116–127. http://dx.doi.org/10.1016/j.cognition.2013.10.002

Caro, T. M., & Hauser, M. D. (1992). Is there teaching in nonhuman animals? *Quarterly Review of Biology,* 151–174.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E., & Schultz, R. (2012). The Social Motivation Theory of Autism. *Trends in Cognitive Sciences, 16,* 231–239. http://dx.doi.org/10.1016/j.tics.2012.02.007

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373,* 209–216. http://dx.doi.org/10.1038/373209a0

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review, 120,* 190–229. http://dx.doi.org/10.1037/a0030852

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13,* 148–153. http://dx.doi.org/10.1016/j.tics.2009.01.005

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69,* 1204–1215. http://dx.doi.org/10.1016/j.neuron.2011.02.027

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18,* 185–196. http://dx.doi.org/10.1016/j.conb.2008.08.003

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125,* 627–668. http://dx.doi.org/10.1037/0033-2909.125.6.627

Deci, E. L., & Ryan, R. M. (1980). The empirical exploration of intrinsic motivational processes. *Advances in Experimental Social Psychology, 13,* 39–80.

Devlin, S., & Kudenko, D. (2012). Dynamic potential-based reward shaping. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 433–440). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415,* 137–140. http://dx.doi.org/10.1038/415137a

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336,* 998. http://dx.doi.org/10.1126/science.1218633

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin, 40,* 986–997. http://dx.doi.org/10.1177/0146167214533130

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117,* 197–209. http://dx.doi.org/10.1037/a0017808

Gerstenberg, T., & Goodman, N. (2012). Ping pong in church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1590–1595). Austin, TX: Cognitive Science Society.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66,* 585–595. http://dx.doi.org/10.1016/j.neuron.2010.04.016

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America, 108*(Suppl. 3), 15647–15654. http://dx.doi.org/10.1073/pnas.1014269108

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology, 45,* 840–844.

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology, 41,* 364–374. http://dx.doi.org/10.1002/ejsp.782

Griffith, S., Subramanian, K., Scholz, J., Isbell, C., & Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 2625–2633). New York, NY: Curran Associates, Inc.

Grusec, J. E., & Kuczynski, L. (Eds.). (1997). *Parenting and children's internalization of values: A handbook of contemporary theory* (Vol. 24). Hoboken, NJ: Wiley.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods, 48,* 829–842.

Hamlin, K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. L. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science, 16,* 209–226. http://dx.doi.org/10.1111/desc.12017

Hamlin, K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450,* 557–559. http://dx.doi.org/10.1038/nature06288

Hiby, E. F., Rooney, N. J., & Bradshaw, J. W. S. (2004). Dog training methods: Their use, effectiveness and interaction with behaviour and welfare. *Animal Welfare, 13,* 63–69.

Ho, M. K., Littman, M. L., Cushman, F., & Austerweil, J. L. (2015). Teaching with rewards and punishments: Reinforcement or communication? In D. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 920–925). Austin, TX: Cognitive Science Society.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 3027–3035). New York, NY: Curran Associates, Inc.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition, 167,* 91–106. http://dx.doi.org/10.1016/j.cognition.2017.03.006

Isbell, C. L., Shelton, C. R., Kearns, M., Singh, S., & Stone, P. (2001). Cobot: A social reinforcement learning agent. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (Vol. 14, pp. 1393–1400). Cambridge, MA: MIT Press.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58,* 284–294. http://dx.doi.org/10.1016/j.neuron.2008.03.020

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition, 140,* 14–23. http://dx.doi.org/10.1016/j.cognition.2015.03.006

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., . . . Casey, B. J. (2011). Behavioral and neural properties of social reinforcement learning. *Journal of Neuroscience, 31,* 13039–13045. http://dx.doi.org/10.1523/JNEUROSCI.2972-11.2011

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1679–1684). Austin, TX: Cognitive Science Society.

Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain Sciences*. Advance online publication. http://dx.doi.org/10.1017/S0140525X14000090

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

LaValle, S. M. (2006). *Planning algorithms*. Cambridge, UK: Cambridge University Press.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual review of neuroscience, 35,* 287–308. http://dx.doi.org/10.1146/annurev-neuro-062111-150512

Lin, A., Adolphs, R., & Rangel, A. (2012). Impaired learning of social compared to monetary rewards in autism. *Frontiers in Neuroscience, 6,* 163. http://dx.doi.org/10.3389/fnins.2012.00143

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In W. W. Cohen & H. Hirsh (Eds.), *Proceedings of the 11th international conference on machine learning* (pp. 157–163). San Francisco, CA: Morgan Kaufmann.

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 G. Ashok, R. Morris, P. Pu, S. Sen, & K. B. Venable (Eds.), *Proceedings of the 28th AAAI Conference on Artificial Intelligence* (pp. 937–943). Palo Alto, CA: AAAI Press.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 19751–19756. http://dx.doi.org/10.1073/pnas.0704452104

Maccoby, E. E. (1992). The role of parents in the socialization of children: An historical overview. *Developmental Psychology, 28,* 1006–1017. http://dx.doi.org/10.1037/0012-1649.28.6.1006

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31,* 838–850.

Meyer, W.-U. (1982). Indirect communications about perceived ability estimates. *Journal of Educational Psychology, 74,* 888–897.

Meyer, W.-U. (1992). Paradoxical effects of praise and criticism on perceived ability. *European Review of Social Psychology, 3,* 259–283.

Miller, A. T., & Hom, H. L. (1996). Conceptions of ability and the interpretation of praise, blame, and material rewards. *The Journal of Experimental Education, 65,* 163–177.

Neumann, R., Rafferty, A. N., & Griffiths, T. L. (2014). A bounded rationality account of wishful thinking. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1210–1215). Austin, TX: Cognitive Science Society.

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

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science, 24,* 751–761. http://dx.doi.org/10.1177/0956797612463080

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,* 364–385.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences, 21,* 425–433.

Powell, L. J., & Spelke, E. (2018). Third party preferences for imitators in preverbal infants. *Open Mind: Discoveries in Cognitive Science, 1,* 183–193. http://dx.doi.org/10.31234/osf.io/2svkh

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1,* 515–526. http://dx.doi.org/10.1017/S0140525X00076512

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution, 27,* 288–295. http://dx.doi.org/10.1016/j.tree.2011.12.004

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences, 17,* 413–425.

Rogoff, B., Paradise, R., Arauz, R. M., Correa-Chávez, M., & Angelillo, C. (2003). Firsthand learning through intent participation. *Annual Review of Psychology, 54,* 175–203. http://dx.doi.org/10.1146/annurev.psych.54.101601.145118

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80,* 1–28. http://dx.doi.org/10.1037/h0092976

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15,* 549–562. http://dx.doi.org/10.1038/nrn3776

Sage, K. D., & Baldwin, D. (2011). Disentangling the social and the pedagogical in infants' learning about tool-use. *Social Development, 20,* 825–844.

Salzinger, K., & Waller, M. B. (1962). The operant control of vocalization in the dog. *Journal of the Experimental Analysis of Behavior, 5,* 383–389. http://dx.doi.org/10.1901/jeab.1962.5-383

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599. http://dx.doi.org/10.1126/science.275.5306.1593

Sears, R. R., Whiting, J. W., Nowlis, V., & Sears, P. (1953). Some child-rearing antecedents of aggression and dependency in young children. *Genetic Psychology Monographs, 47,* 135–236.

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. http://dx.doi.org/10.1016/j.cogpsych.2013.12.004

Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning, 22,* 123–158. http://dx.doi.org/10.1007/BF00114726

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

Skinner, B. F. (1974). *About behaviorism*. New York, NY: Knopf.

Stevens, J. R., Cushman, F. A., & Hauser, M. D. (2005). Evolving the psychological mechanisms for cooperation. *Annual Review of Ecology, Evolution, and Systematics, 36,* 499–518.

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

Téglás, E., Gergely, A., Kupán, K., Miklósi, A., & Topál, J. (2012). Dogs' gaze following is tuned to human communicative signals. *Current Biology, 22,* 209–212. http://dx.doi.org/10.1016/j.cub.2011.12.018

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*(4), i–109. http://dx.doi.org/10.1037/h0092987

Topál, J., Gergely, G., Erdõhegyi, A., Csibra, G., & Miklósi, A. (2009). Differential sensitivity to human communication in dogs, wolves, and human infants. *Science, 325,* 1269–1272.

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

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17,* 89–100.

(*Appendices follow*)

Appendix A

Theorem and Proof Regarding Positive Reward Cycles

We are interested in characterizing when a combination of feedback and learning strategies leads to a positive reward cycle. To do so, we introduce the following definitions.

Given a state space $\mathcal{S}$, learner actions $\mathcal{A}^{\mathcal{L}}$, and transition function $T : \mathcal{S} \times \mathcal{A}^{\mathcal{L}} \rightarrow \mathcal{S}$, a reward maximizing learner can have a discount rate $\gamma \in [0,1]$, a horizon $H \in \mathbb{N} \cup \{\infty\}$, and intrinsic rewards $R : \mathcal{S} \times \mathcal{A}^{\mathcal{L}} \times \mathcal{S} \rightarrow \mathbb{R}$. A static teaching strategy $\mathcal{T}$ is one in which the same feedback function $F : \mathcal{S} \times \mathcal{A}$ is used at every timestep. Let $\pi_0=(\pi_0, \pi_1, \pi_2, \ldots, \pi_t, \ldots)$ be a sequence of learned policies, π: $\mathcal{S} \rightarrow \mathcal{A}^{\mathcal{L}}$, that results from the learning strategy $\mathcal{L}$ interacting with the teaching strategy $\mathcal{T}$. Standard RL algorithms seek to maximize expected reward (Sutton & Barto, 1998) in that they converge to a policy that optimizes value from all $s \in \mathcal{S}$. We define a reward maximizing learning algorithm $\mathcal{L}^{\mathrm{RM}}$ as this type of algorithm. That is, $\lim_{t\rightarrow\infty} \pi_t = \pi^{\mathrm{RM}} = \mathrm{argmax}_\pi V^\pi(s), \forall s$, where $V^\pi(s_0) = \sum_{t=0}^{H} \gamma^t[R(s_t, a_t, s_{t+1}) + F(s_t, a_t)]$, where $a_t = \pi(s_t)$ and $s_{t+1} = \mathcal{T}(s_t,a_t)$. A teaching strategy is *effective* for teaching a target policy π^* from initial state s_0 to a learning strategy $\mathcal{L}$ if $\lim_{t\rightarrow\infty} \pi_t = \pi^*$.

Next, we define the exact conditions under which feedback "counts" as producing a positive reward cycle given the learner discount, horizon, and intrinsic rewards.

Definition

A *positive reward cycle* is defined as a cyclical trajectory that deviates from the target policy and yields a greater cumulative discounted reward for a reward maximizing learning algorithm $\mathcal{L}^{\mathrm{RM}}$ than following the target policy. That is, $\zeta = (\tilde{s}_0, \tilde{a}_0, \ldots, \tilde{s}_k, \ldots, \tilde{s}_H)$, $k < H$ is a positive reward cycle with respect to a reward-maximizing learning algorithm $\mathcal{L}^{\mathrm{RM}}$ with intrinsic rewards R, discount γ, and horizon H; target policy π^*; and feedback F if and only if $\tilde{s}_0 = \tilde{s}_k$, there is at least one a_t in ζ such that $\pi^*(s_t) \neq a_t$, and $V(\zeta) = \sum_{t=0}^{H} \gamma^t[R(s_t, a_t, s_{t+1}) + F(s_t, a_t)] > V^{\pi^*}(\tilde{s}_0)$.

Based on this, we can prove the following theorem:

Theorem

Given a reward maximizing learning algorithm $\mathcal{L}^{\mathrm{RM}}$ with discount rate γ, horizon H, and intrinsic rewards R, a static teaching strategy $\mathcal{T}$ with feedback function F for teaching π^* from s_0 is not effective for teaching π^* if there is a positive reward cycle.

Proof

A reward maximizing learning algorithm $\mathcal{L}^{\mathrm{RM}}$ will learn the policy $\lim_{t\rightarrow\infty} \pi_t = \pi^{\mathrm{RM}} = \mathrm{argmax}_\pi V^\pi(s), \forall s$. If there is a positive reward cycle, then there is a state for which it is not the case that the target policy maximizes value from that state. Therefore, $\pi^{\mathrm{RM}} \neq \pi^*$ and $\mathcal{T}$ is not an effective strategy for teaching π^*.

Remarks

First, most algorithms in reinforcement learning fall under the definition of reward-maximizing (Sutton & Barto, 1998). Additionally, although the proof is for a *static* teaching strategy, extending the idea to the dynamic case is conceptually similar and interested readers should look at Devlin & Kudenko (2012) for generalizations of the shaping theorem presented in Ng et al. (1999). Finally, the experiments and analyses we report in the main portion of the article assume that learners have no intrinsic motivation (i.e. $R(s, a, s') = 0$ $(\forall s, a, s')$. We include the R term in this formulation as it makes explicit how intrinsic motivation influences whether a positive reward cycle will be learned.

(*Appendices continue*)

Appendix B

Algorithmic Details

Model-Free Algorithm

The Model-Free algorithm we implement is Q-learning with eligibility traces (Sutton & Barto, 1998; Watkins & Dayan, 1992). Like other Model-Free algorithms, Q-learning does not have an explicit representation of transitions in the world or rewards, but updates an estimate of action-values as it explores the environment. It is also guaranteed to converge to the reward-maximizing action-values in the limit. An eligibility trace allows the algorithm to log recently visited states and associate later rewards with those states without needing to revisit them. This enables the algorithm to more quickly propagate information about rewards to states that were recently visited without affecting its final convergence. Formally, Q-learning updates its current action-values according to the following rule (where $0 < \alpha < 1$ controls the learning rate and λ is an eligibility trace decay rate):

$$q_t(s, a) \leftarrow q_t(s, a) + \alpha\delta_t e_t(s, a), \ \forall\, s, a \tag{5}$$

where

$$\delta_t = f_{t+1} + \gamma \max_{a'} q_t(s_{t+1}, a') - q_t(s_t, a_t) \tag{6}$$

and

$$e_t(s, a) = \begin{cases} 0, & \text{if } q_t(s_t, a_t) < \max_{\bar{a}} q_t(s_t, \bar{a}) \text{ and } (s, a) \neq (s_t, a_t) \\ 1, & \text{if } q_t(s_t, a_t) < \max_{\bar{a}} q_t(s_t, \bar{a}) \text{ and } (s, a) = (s_t, a_t) \\ \gamma\lambda e_{t-1}(s, a) + 1, & \text{if } q_t(s_t, a_t) = \max_{\bar{a}} q_t(s_t, \bar{a}) \text{ and } (s, a) = (s_t, a_t) \\ \gamma\lambda e_{t-1}(s, a), & \text{otherwise} \end{cases}, \ \forall\, s, a. \tag{7}$$

δ_t denotes the reward prediction error and e_t denotes the eligibility trace function at a timestep t.

Model-Based Algorithm

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. For example, algorithms like Rmax have the learner simultaneously learn the transition model of the world and the world reward function (Brafman & Tennenholtz, 2003). Here, however, we assume that the learner has a complete and accurate model of state transitions and is mainly concerned with learning and maximizing the teacher's feedback function.

In our implementation, at each timestep, the reward function is updated:

$$R_{t+1}(s_t, a_t) \leftarrow (1 - \alpha)R_t(s_t, a_t) + \alpha f_t, \tag{8}$$

where $0 < \alpha < 1$ is the learning rate. This reward function is combined with the transition function to calculate an optimal value function $V_t^*(s) = \max_a R_t(s, a) + \gamma\sum_{s'} T(s, a, s')V_t^*(s')$ and an optimal (reward-maximizing) policy π_{RM}.

Action-Signaling Implementation

The action-signaling learner treats feedback as a direct signal of whether the learner's action matches the desired action in the teacher's target policy. A positive response from the teacher indicates that the action is "correct" from the teacher's perspective, while a negative response indicates that it is "incorrect". Here, feedback is not treated as a quantity to maximize, rather, it is diagnostic of an unknown variable: the teacher's target policy, π^*. Thus, learning from evaluative feedback can be modeled as a social or communicative inference problem in a manner similar to inferring an actor's goals based on their behavior or a speaker's meaning based on their utterances. Following recent research in computational models of social cognition and language, we model these inferences using Bayesian models (Baker et al., 2009; Frank & Goodman, 2012; Loftin et al., 2014).

For Experiment 3, our learning algorithm updates a posterior probability over target policies π^* at each timestep according to the equation:

$$P(\pi^* \mid \cdot\, f_{1:T}, s_{1:T}, a_{1:T}) \propto P(\pi^*)\prod_{t=1}^{T} p(f_t \mid \cdot\, \pi^*(s_t), a_t). \tag{9}$$

Eventually, the Learner Then Selects Actions with the Highest Marginalized Probability, for All $s \in \mathcal{S}$:

$$\pi_{\text{Action-Signaling}}(s) = \operatorname{argmax}_{a \in A(s)} \sum_{\pi^*} \mathbb{1}\{\pi^*(s) = a\} P(\pi^* \mid f_{1:T}, s_{1:T}, a_{1:T}), \tag{10}$$

where {condition} Is 1 if *condition* Is True, and 0 if It Is False.

The likelihood function, $p(f_t \mid \pi^*(s_t), a_t)$, is the probability of the teacher giving feedback f_t for the learner taking action a_t in state s_t given the target policy is π^*. It is represented as a sigmoid function between -1 and 1. The probability of reward when the action is correct or incorrect is, respectively:

$$p(f_t \mid \pi^*(s_t) = a_t) = \frac{1}{1 + \exp\{-\kappa f_t\}}, \tag{11}$$

and

$$p(f_t \mid \pi^*(s_t) \neq a_t) = \frac{1}{1 + \exp\{\kappa f_t\}}, \tag{12}$$

(*Appendices continue*)

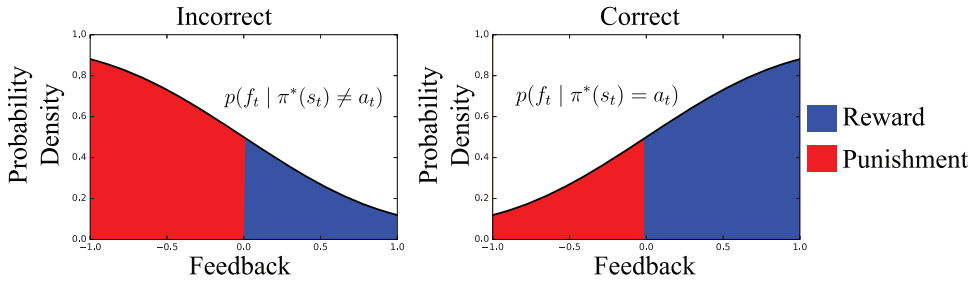

Figure B1. Graphs of likelihood functions for action-signaling models. $\kappa = 2$. See the online article for the color version of this figure.

where $\kappa > 0$ is the slope of the sigmoid function, controlling how smooth feedback is expected to be. As κ increases, the likelihood becomes more deterministic. Additionally, for all values of κ, this function integrates to 1 on the interval [–1, 1], making it a valid probability distribution. As plotted in Figure B1, positive feedback has higher probability when the action was correct, while negative feedback has higher probability when the action is incorrect. One consequence of this formulation is that if a teacher gives a neutral reward of 0, the likelihood given the action is correct is equal to if it is incorrect. As a result, noisy feedback and neutral feedback or the absence of feedback does not affect the learned policy.

Appendix C

Experiment 3: Additional Model Fits

Table C1
Experiment 3a: Fixed Effects in Linear Regression Model of Target Action Rewards

Predictor	Estimate	*SE*	*df*	*t*	*p*
Intercept	.63	.02	173.26	33.97	<.001
Day	-9.0×10^{-3}	3.1×10^{-3}	177.41	−2.92	<.01
Action	7.5×10^{-2}	9.3×10^{-3}	164.00	8.13	<.001
Action-signaling vs. reward-maximizing	-1.8×10^{-3}	1.0×10^{-2}	177.58	−.17	.86
Model-based vs. model-free	5.2×10^{-2}	1.8×10^{-2}	179.92	2.89	<.01
Neutral vs. optimistic initialization	-3.1×10^{-2}	1.8×10^{-2}	180.18	−1.75	.08
Uniform vs. state prior	1.9×10^{-2}	2.5×10^{-2}	176.66	.77	.45
Condition	-1.4×10^{-2}	3.6×10^{-2}	179.92	−.38	.70
Day × Action	-2×10^{-3}	1.4×10^{-3}	141.84	−1.40	.16

Note. $N = 180$ participants, $N = 6{,}830$ observations in total.

(*Appendices continue*)

Table C2
Experiment 3b: Fixed Effects in Linear Regression Model of Target Action Rewards

Predictor	Estimate	*SE*	*df*	*t*	*p*
Intercept	6.1×10^{-1}	1.9×10^{-2}	176.33	31.94	<.001
Day	-2.7×10^{-2}	3.1×10^{-3}	177.52	−8.62	<.001
Action	7.8×10^{-2}	8.6×10^{-3}	148.22	9.07	<.001
AS vs. RM	1.2×10^{-2}	1.3×10^{-2}	185.47	.91	.36
MB vs. MF	1.2×10^{-2}	2.2×10^{-2}	173.32	.56	.58
N. vs. O. initialization	-6.9×10^{-3}	2.2×10^{-2}	173.31	−.32	.75
U. vs. S. prior	2.5×10^{-2}	3.1×10^{-2}	190.88	.80	.42
Condition	2.7×10^{-2}	4.3×10^{-3}	173.29	.61	.54
Day × Action	-3.7×10^{-3}	1.7×10^{-3}	182.44	−2.23	<.05
Action × AS vs. RM	-2.0×10^{-2}	4.7×10^{-3}	178.72	−4.36	<.001
Action × MB vs. MF	1.0×10^{-2}	8.2×10^{-3}	177.61	1.22	.22
Action × N. vs. O. Initialization	-4.4×10^{-3}	8.2×10^{-3}	177.81	−.53	.60
Action × U. vs. S. Prior	5.9×10^{-3}	1.1×10^{-2}	178.37	.52	.61
Action × Condition	-2.9×10^{-2}	1.6×10^{-2}	177.43	−1.77	.08

Note. AS = action-signaling; RM = reward-maximizing; MB = model-based reward-maximizing; MF = model-free reward-maximizing; N. = neutrally initialized reward-maximizing learners; 0. = optimistically initialized reward-maximizing learners; U. = uniform prior; S. = state-based prior. $N = 180$ participants, $N = 7{,}418$ observations in total.

Table C3
Experiment 3c: Fixed Effects in Linear Regression Model of Target Action Rewards

Predictor	Estimate	*SE*	*df*	*t*	*p*
Intercept	.71	1.8×10^{-2}	177.40	39.31	<.001
Action	5.7×10^{-2}	7.9×10^{-3}	168.20	7.19	<.001
Day	-1.0×10^{-2}	2.9×10^{-3}	173.20	−3.76	<.001
Day × Action	-3.8×10^{-4}	1.4×10^{-3}	164.60	−.27	.79

Note. $N = 180$ participants, $N = 6{,}789$ observations in total.

Received July 19, 2017
Revision received October 25, 2018
Accepted December 17, 2018 ■