ARTICLE IN PRESS

Cognition xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Social is special: A normative framework for teaching with and learning from evaluative feedback

Mark K. Ho[a,*], James MacGlashan[b], Michael L. Littman[b], Fiery Cushman[c]

[a] Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Box 1821, Providence, RI 02912, United States
[b] Department of Computer Science, Brown University, 115 Waterman St, Providence, RI 02906, United States
[c] Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, United States

ARTICLE INFO

Article history:
Received 31 March 2016
Revised 13 December 2016
Accepted 9 March 2017
Available online xxxx

Keywords:
Reward
Punishment
Theory of mind
Social learning
Evaluative feedback
Teaching

ABSTRACT

Humans often attempt to influence one another's behavior using rewards and punishments. How does this work? Psychologists have often assumed that "evaluative feedback" influences behavior via standard learning mechanisms that learn from environmental contingencies. On this view, teaching with evaluative feedback involves leveraging learning systems designed to maximize an organism's positive outcomes. Yet, despite its parsimony, programs of research predicated on this assumption, such as ones in developmental psychology, animal behavior, and human-robot interaction, have had limited success. We offer an explanation by analyzing the logic of evaluative feedback and show that specialized learning mechanisms are uniquely favored in the case of evaluative feedback from a social partner. Specifically, evaluative feedback works best when it is treated as communicating information about the value of an action rather than as a form of reward to be maximized. This account suggests that human learning from evaluative feedback depends on inferences about communicative intent, goals and other mental states—much like learning from other sources, such as demonstration, observation and instruction. Because these abilities are especially developed in humans, the present account also explains why evaluative feedback is far more widespread in humans than non-human animals.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Parents scold; teachers grade; lovers pout; bosses bonus; colleagues grouse; nations sanction; citizens protest; eyes smile and mouths frown. In short, people rarely forgo an opportunity for *evaluative feedback*: reward or punishment of another person in a manner designed to change their future behavior. Although teaching by evaluative feedback is sometimes costly, the potential benefit is obvious: We can exploit the capacity of social partners to learn from reward and punishment to shape their future behavior to profit ourselves, our kin and our allies. In many instances, such as parenting, long-run benefits accrue not only to the teacher (e.g., a parent) but also to the learner (the child) as they learn more adaptive patterns of behavior. The ubiquity of evaluative feedback is unremarkable because it is so effective. Dozens of laboratory (Balliet, Mulder, & Van Lange, 2011; Fehr & Gächter, 2002) and field (Owen, Slep, & Heyman, 2012) studies show that humans can effectively shape the behavior of other humans through the use of selective reward and punishment. Our goal is to understand how.

More precisely, we ask whether there is anything special about learning from social rewards and punishments, as compared to ordinary environmental rewards and punishments. Evaluative feedback from social others take on many forms. For instance, a social other may redirect naturally occurring stimuli in order to inflict pleasure or pain on a learner; giving or withholding food, comfort, poison, and painful experiences all fall under this category. Evaluative feedback may also depend on uniquely human and intrinsically social signals such as verbal praise or reprimands, or a smile or scowl. Although these forms of evaluative feedback differ in many ways, they all involve (1) a social agent causing (2) a rewarding or aversive experience in (3) another social agent, and (4) in a manner ultimately designed to cause learning and behavioral change. What are the cognitive mechanisms that support this form of social teaching and learning in humans? Are they specially adapted to the social domain? Should they be?

At first blush, the answer seems obvious. The tendency of organisms to repeat what is positive and to avoid what is negative is fundamental to psychological theory, akin to gravity in physics

[*] Corresponding author.
E-mail addresses: mark_ho@brown.edu (M.K. Ho), jmacglashan@gmail.com (J. MacGlashan), mlittman@cs.brown.edu (M.L. Littman), cushman@fas.harvard.edu (F. Cushman).

http://dx.doi.org/10.1016/j.cognition.2017.03.006
0010-0277/© 2017 Elsevier B.V. All rights reserved.

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006

or natural selection in biology. The power of these rewards and punishments to shape human behavior is entirely unsurprising because rewards and punishments exert a gravitational force on the behaviors of non-human animals from the sea-slug (Cook & Carew, 1986) to the chimpanzee (Randolph & Brooks, 1967), and every lab rat (Guttman, 1953), cat (Populin & Yin, 1998; Thorndike, 1898) and pigeon (Skinner, 1948) in between. Moreover, brain imaging studies have confirmed that material rewards and inherently social rewards like facial expressions are processed in similar regions (Lin, Adolphs, & Rangel, 2012). Here, then, is a simple premise that has inspired much prior research: *Social rewards and punishment shape behavior by exploiting the same learning mechanisms that process environmental rewards and punishments*. This claim does not commit to any particular form of the learning (associative, causal, Bayesian, etc.). Rather, the key claim is that however we learn from rewards and punishments of non-social origin, we learn the same way from rewards and punishments originating from social partners. That is, we learn from the sting of criticism just as we would from the prick of a thorn.

Although parsimonious, this premise is closely associated with several unfulfilled programs of research. In the 1950s and 1960s, buoyed by decades of progress in animal learning, researchers began to apply principles of operant conditioning discovered in non-social learning tasks to the socialization of children (Aronfreed, 1968; Bryan & London, 1970; Sears, Maccoby, & Levin, 1957). There were some later successes in showing that behaviors like altruism could be reinforced (Gelfand, Hartmann, Cromer, Smith, & Page, 1975; Grusec & Redler, 1980). But as operant conditioning as a theory of social learning in humans lost adherents, the field eventually moved on to alternative models of social learning—for instance, by observation, instruction, or attribution—rather than learning by reinforcement as such (Grusec, 1997; Maccoby, 1992). There is something unsatisfying about this resolution: Humans obviously *do* reward and punish each other, so why can't our best models explain how this contributes to learning?

Similarly, buoyed by theoretical models that predicted the evolution of cooperation through punishment (Clutton-Brock & Parker, 1995) and reciprocal rewards (Trivers, 1971), biologists sought to document their prevalence among non-human animals. Again, these attempts yielded surprisingly few empirical successes (Hammerstein, 2003; Raihani, Thornton, & Bshary, 2012; Stevens, Cushman, & Hauser, 2005; Stevens & Hauser, 2004), and attention turned to alternative means of explaining non-human prosociality (West, Griffin, & Gardner, 2007). Again, something has been left unresolved: Given that animals are proficient at learning from environmental rewards and punishments, why don't they reward and punish *each other* more often?

In more recent decades, computer scientists have developed mathematical tools to build agents that embody the basic principles of non-human and human reward learning (e.g. Sutton & Barto, 1998). Yet, when they allow actual human participants to train these agents through reward and punishment, the results are spectacularly disappointing. Machines will often unlearn their initial training or even acquire unintended behaviors that human trainers fail to detect (Isbell, Shelton, Kearns, Singh, & Stone, 2001). Here, again, there is something left unfulfilled. Humans are happy to reward and punish agents employing artificial intelligence in order to improve their behavior. But if the agents are designed to *maximize* those rewards (and minimize punishment), they fail to learn what the humans are trying to teach. Where is the bug in the system?

Collectively, this evidence suggests that there is something special about the way that *human* learners respond to *social* rewards and punishments—and something correspondingly special about how human teachers structure those rewards and punishments. By understanding what that "special something" is, we will be in a better position to understand what human evaluative feedback is good for, why non-human animals are relatively less prone to use it, and how to build artificial intelligence that benefits from it.

Our approach to this problem leverages basic concepts borrowed from reinforcement learning, a framework that formalizes the problem of learning and decision-making based on reward and punishment (Dayan & Niv, 2008; Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). We provide a normative analysis of how teaching and learning from social evaluative feedback should be structured, contrast features of this approach to learning from non-social reinforcement, and compare each of these models against extant findings.

2. Adapting to non-social rewards and punishments

Like most animals, humans learn the value of actions as they experience positive and negative outcomes in the environment. For instance, a rat learns the value of pushing a lever when it experiences contingent food rewards (Guttman, 1953). A major goal of contemporary learning theory is to provide a formal account of the cognitive operations that enable this form of learning (Dayan & Niv, 2008). Many diverse answers to this problem have been proposed, but virtually all of them share a few key features. By summarizing these features, we can state with greater precision the potential similarities or dissimilarities between "traditional" reward learning (in non-social settings) and evaluative feedback (i.e. reward and punishment in a social setting).

2.1. The problem of learning value from reward

The central challenge of decision making for organisms is to choose the right behavior in any situation that arises. If the optimal, fitness-enhancing behavior were sufficiently consistent across time and individuals, then it could be specified entirely innately. For instance, koalas, an arboreal marsupial, mainly consume toxic eucalypts that are not difficult to find or competitively consumed by other species. In part due to the natural invariance of their main food source, koalas will only consume eucalyptus leaves that are attached to branches and not ones that have been plucked and placed on a flat surface (Tyndale-Biscoe, 2005). Reflexes, fixed action patterns, or unconditioned responses all fall into this category of innate stimulus-response mappings.

Of course, this approach is generally impractical: Many features of the world are not predictable from birth and stable across generations. Consider, for instance, the challenge of foraging for food. The timescale at which forests burn, herds migrate, ponds dry, and so forth, means that the most effective behaviors for obtaining food undergoes large changes within (and certainly between) generations. Thus, organisms must have an adaptive mechanism for altering their behavior in response to variable circumstances.

One solution to the problem of adapting behavior to partially predictable environments consists of two interacting representations: innate rewards and learned value (Littman & Ackley, 1991). First, an innate system designates the experience of certain actions, stimuli, or states of affairs as intrinsically rewarding or aversive because they are reliable indicators of fitness improvement or decline. Honey, for instance, could be experienced by an organism as intrinsically rewarding because of its high caloric content. Conversely, bee stings could be intrinsically aversive because they lead to swelling and potential infections.

Second, as an organism acts and undergoes different rewarding, aversive, and neutral experiences, a learning process flexibly updates a representation that predicts the contingencies of actions and experiences. For example, if an organism experiences eating

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006

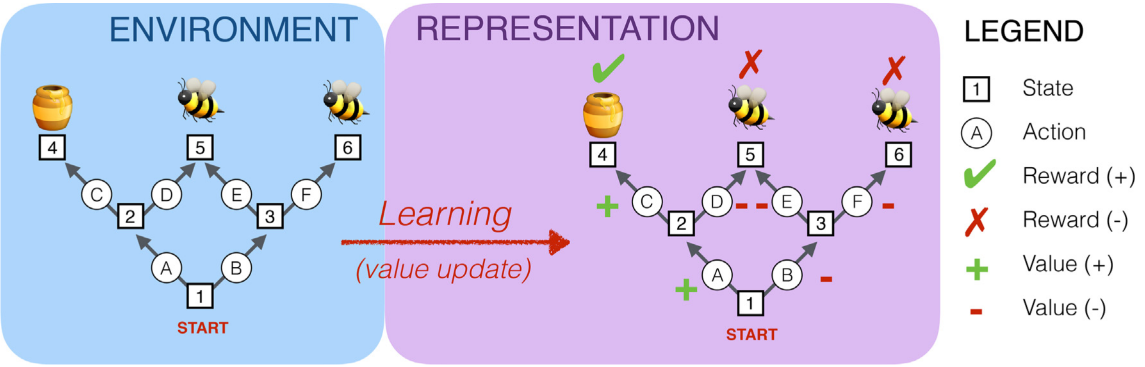

Fig. 1. Reward learning in a non-social setting.

honey as rewarding and bee stings as aversive, it will learn to take actions that make eating honey more likely and bee stings less likely. It would learn that sticking one's hand into a jar of honey leads to a greater balance of rewarding over aversive experiences than, say, sticking one's hand into a bee hive. We can think of the organism representing not just the experience of reward, but the *average future rewards* predicted by certain actions and states of affairs, otherwise known as their *value*.

This distinction between the fixed experience of reward and learned representation of value plays an essential role in the discussion that follows. In some sense, value is "reward and more": It incorporates environmental randomness, future possibilities, and causal relationships into a single, action-guiding representation that predicts future rewards. A representation of value assigns the motivational power of rewards to various actions that are learned to reliably cause those rewards. For instance, if a person finds eating honey rewarding, then she would learn to value actions such as sticking her fingers in jars of honey, walking to the appropriate aisle in the grocery, and perhaps even beekeeping. As we have emphasized, value can also be updated in response to changes in environmental contingencies. If the same person learned that the grocery on her block had stopped selling honey, she would eventually reduce the represented value of going to that store, reflecting the diminished rewards available.

Another crucial difference between reward (a feature of an experience) and value (a representation of future rewards) is that a single agent's value representations across actions and states are correlated. That is, the value of one action is directly determined by the value of the other actions and states it predicts. For example, the reward of eating honey determines the value of sticking one's hand in a jar of honey, which determines the value of opening a jar of honey. An incoherent set of value representations would place high value on opening the jar, but low value on sticking one's hand in the jar (that is, supposing that there is no other way to obtain the contents of the jar).[1] Put in folk-psychological terms, if an agent were to open a jar of honey, leave it on the counter, and walk away, one might ask "why did they open the jar if they were not planning on sticking their hand into the jar to get the honey?" In contrast, differently rewarding or aversive experiences that are sequentially or causally linked are not constrained to relate to one another in this manner. For instance, a lactose intolerant agent may simultaneously experience mint chip ice cream as delicious (i.e., "positive reward") but also experience nausea (i.e., "negative reward") after eating dairy. This is unfortunate, but it is not incoherent.

Fig. 1 presents a schematic representation of the relationship between reward and value. The learner proceeds through a series of states (1,2,3, etc.) and in each state faces a set of possible actions (A, B, etc.). Some transitions result in experiences of positive reward like eating honey (State 4) or experiences of negative reward (i.e. punishment) like being stung by bees (States 5 and 6). Others do not directly lead to bees or honey, but lead to states that subsequently lead to such outcomes. For example, taking Action A leads to State 2, which then leads to State 4, which has honey. Eventually, the learner should learn to assign Action A *positive value* in light of its relationship to future honey. By a similar logic, it should eventually learn that Action B has *negative value* since it only leads to the painful experience of bees.

2.2. Solutions to learning value from reward

In the non-social setting, a reinforcement learning "problem" is defined by the fixed experience of reward in an environment with certain contingencies. Its "solution" is an action-guiding representation of value that maximizes average future rewards. Thus, the reinforcement learning formulation can be understood as a "computational level" theory of reward-guided decision-making (Anderson, 1990; Marr, 1982). Much past and present psychological research on reward learning can then be viewed as characterizing the precise mechanisms that enable agents to calculate these solutions.

For instance, early psychological approaches to reward learning tended to avoid positing sophisticated internal mental processes, instead depending on simple associative models stated over sensorimotor primitives. Value could be approximated by associating rewards with previously visited states, akin to Thorndike's "Law of Effect" (Thorndike, 1898). In contrast, contemporary work on value-guided decision-making often invokes sophisticated mental representations and computations that go well beyond associative learning (Sutton & Barto, 1998). For example, organisms could construct an internal causal model or "cognitive map" that captures the relationships between actions and states, determine the value of each state, and then derive a plan that maximizes reward. Much current work seeks to characterize the diverse solutions humans use, spanning from simple association to explicit planning (Dolan & Dayan, 2013).

How humans and other organisms solve the problem of estimating value in a non-social setting has been more extensively reviewed elsewhere (Dayan & Niv, 2008; Lee, Seo, & Jung, 2012) and is beyond the scope of this article. Nonetheless, what all these approaches share is that they posit theories of how organisms

[1] Readers familiar with Markov Decision Processes will recognize this as a way of stating that mappings from states/actions to value must satisfy recursive Bellman equations, whereas mappings from states/actions to rewards do not have analogous constraints.

transform the *experience of reward* into effective, action-guiding *representations of value*. Our goal in the remainder of the paper is to assess whether this normative framework for conceptualizing non-social reward and value readily translates to social domains.

3. Adapting to or adopting from social rewards and punishments

Now that we have described non-social punishments and rewards, we can be more precise about what it would mean for social punishments and rewards (“evaluative feedback”) to operate in an identical manner. The learner would experience teachers’ evaluative feedback as rewarding or aversive and assign high value to states or actions that maximize reward. We can call the interpretation and use of rewards and punishments in this manner “reward feedback”, and the learning process “value update”, respectively.

This extension of reward learning to the social domain may sound perfectly natural, or perhaps even trivially obvious. But consider the following familiar scenarios:

(1) Maria is a toddler. She is trying to open a box but can’t quite do it. Her mother smiles and encourages her, but then has to leave. Motivated by her mother’s encouragement to continue trying, Maria eventually succeeds in opening the box even when she no longer receives positive feedback for trying.

(2) Allen’s mentor praises him when he writes in short, clear sentences, so Allen is sure to write this way even though he initially does not care for them. After Allen graduates his mentor stops reading his work, but Allen continues writing in short, clear sentences because he now finds them satisfying himself.

Rewards and punishments clearly play a role in these familiar forms of social learning, but the standard approach of “reward feedback” and “value update” cannot account for them. After social rewards are withdrawn, standard value update processes should re-assign neutral value to the behaviors. By analogy, if foraging a path stopped yielding honey, a reward maximizing agent would stop walking down it.

In cases like these, however, we do not expect that Maria or Allen will treat evaluative feedback simply as a form of reward to be maximized. Rather, the rewards and punishments of their teachers will be interpreted as *communicating information about value*. We expect Maria to interpret feedback as a suggestion by her mother “you are on your way towards opening the box!”. Similarly, we expect extrinsic feedback to eventually allow Allen to “internalize” the reward of clear writing such that it persists even after the feedback is withdrawn. In essence, we expect them to treat feedback not as a reward itself, but rather as a signal that behavior is on the right (or wrong) path to future success.

Our main insight is that human rewards and punishments do not simply influence an agent’s behavior through incentives, but can also in themselves “send a message” to a learner. Previous work in philosophy and psychology has examined the importance of signaling and recognizing communicative intent during teaching and learning. Much of this work has focused on verbal communication, but researchers have also studied non-verbal communicative behaviors such as demonstrating actions or selecting examples to illustrate a concept (Csibra & Gergely, 2009; Shafto, Goodman, & Griffiths, 2014; Sperber & Wilson, 1986). Importantly, although language is helpful, it is neither necessary nor sufficient for recognizing and learning from communicative intent. For instance, keying an adversary’s sports car is surely an instance of non-verbal communication that relies on presenting a social other with an aversive state of affairs.

Our goal in this section is to explain why “traditional” reward feedback and value update fail to support many familiar and useful forms of learning, while evaluative feedback understood as communication succeeds. To that end, we unpack the demands of learning in a social setting, describe the limitations of relying on standard non-social learning mechanisms, and explain why communicating the mental structure of reward and value is advantageous.

3.1. Learning in a social world

Why interpret evaluative feedback as communication, rather than ordinary reward? In order to analyze the unique features of learning and decision-making in the social setting, it is helpful to extend our graphical representation of reward learning to include both the teacher and the learner, as in Fig. 2. There are several key differences. First, part of the learner’s environment now includes the evaluative feedback provided by a teacher. Second, the teacher has its own mental structures that may differ from the learner’s and may already be well-adapted to its environment.

When we initially introduced this diagram, we assumed that while *honey* is reliably fitness-enhancing on an evolutionary timescale, the *path* to honey may change. This motivated an architecture in which honey is innately rewarding, but the value of paths is learned and updated. In the social setting, however, a teacher may already know the best path by which to obtain honey—i.e., she may have acquired a representation of value that would profit the learner. What’s more, the teacher may be benefitting from other foods, like avocadoes, that are even better for a learner’s health—but for which the learner presently experiences no reward. Thus, whereas in the non-social setting an agent can only *adapt* to the structure of its environment by seeking innate rewards, in the social setting agents can now *adopt* the mental structures of their social partners: both their useful representations of value, and even the very things they find rewarding.

A specific case helps to sharpen this point. Suppose a father wishes to teach his daughter to share her toys with her playmates. Thus, he punishes her when she hoards her toys but rewards her for sharing. What is the goal of his behavior? One possibility is to assume that she will treat his evaluative feedback identically to a non-social reward. If so, then his goal must be to shape her behavior by providing an external incentive for the behavior he desires her to perform (sharing). In other words, he hopes that she will fear his continued punishment and seek his continued praise, and so she will share. Intuitively, however, this explanation seems incomplete. At the very least, an obvious problem is that the daughter would no longer be motivated to share once the father is no longer around to shape her behavior.

In contrast, if the daughter is disposed to treat her father’s evaluative feedback as communication to be understood rather than reward to be maximized, she might adopt the father’s mental representations as her own. There are at least two possible ways this might occur. First, the father could hope to guide his daughter to acquire high value behaviors that lead towards an outcome that she would already consider rewarding. In other words, he wants to help her discover behaviors that she would choose, if only she knew their consequences. For instance, maybe sharing leads to lasting friendship. The daughter is not aware of this connection, but based on the father’s evaluative feedback provisionally accepts that sharing is somehow instrumentally valuable to achieving her other goals. This ameliorates the problem of withdrawal of feedback, since now her representation of value has been directly updated. If the benefits of sharing only accrue over many sessions of playing with toys with and without her father, it is all the more

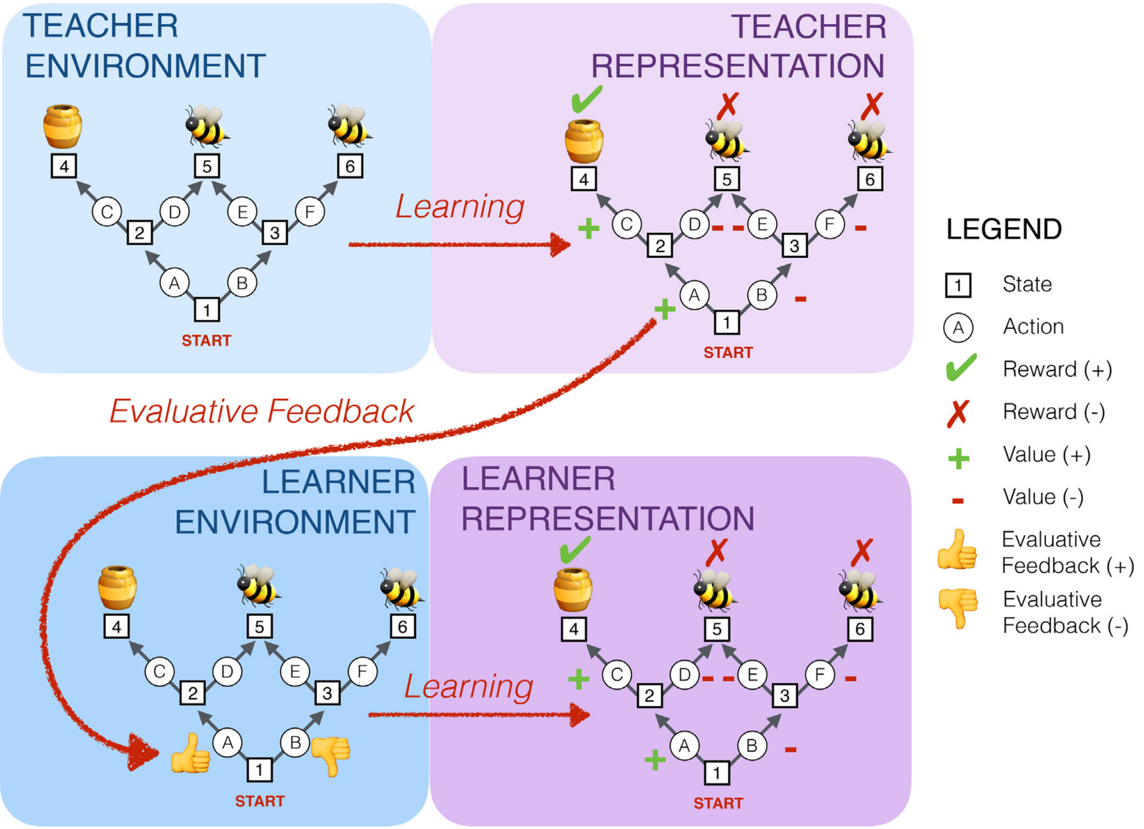

Fig. 2. Learning from evaluative feedback in the social setting.

important that she persist in the high value activity of sharing. Similarly, since sharing sequentially relates to other actions (e.g. making sure her partner enjoys the shared toy) she may also infer that those actions have high value.

Alternatively, rather than demonstrating a connection between sharing and reward, the father may hope to modify the very structure of his daughter's rewards. Presently, sharing does not count as a rewarding outcome or activity to her. He hopes to change this fact – to impart on her what economists call a "taste" or preference for fairness not merely as instrumentally valuable, but as an intrinsic good in itself. Socialization research typically refers to this process as "internalization" (Grusec & Goodnow, 1994), and it seems the most effective mechanism for affecting enduring changes in behavior. Not only does it guarantee that the daughter will continue sharing on her own, but it may also prompt her to search for novel ways in which to make sharing occur since she now values it for its own sake.

The father-daughter interactions sketched above all conform to the everyday experience of evaluative feedback, but reward feedback and value update can only provide an account of shaping. This is because in the social setting, learners do not just adapt to the contingencies of the environment but can also interpret evaluative feedback as communication about a desirable mental structure, and then adopt that mental structure as their own. Adoption may take two forms corresponding to the two basic mental structures that motivate behavior: Value (indicating an instrumental good) and reward (indicating an intrinsic good). In the next few sections, we spell out the methods, rationale and consequences of these approaches. Two key themes emerge: As compared with traditional "adaptive" approaches, the "adoption" of a teacher's mental structure is (1) persistent, allowing the learner to maintain valuable actions even in the absence of the teacher, and (2) inferentially rich, guiding the learner towards new sequences of valuable action.

3.2. From reward feedback to value feedback

First, we consider how learners might adopt their teacher's representations of value. In order to make it clear what this involves, and how it differs from interaction with non-social environments, we must begin with the distinction between the innately specified experience of reward and learned representation of value. In a non-social setting, an agent experiences reward and then uses this to compute value herself. If a teacher's evaluative feedback is similarly processed as reward we call this "reward feedback".

We contrast this standard approach with the alternative in which the learner interprets the teacher's evaluative feedback as a signal about value, which the learner then adopts directly. That is, rather than directly maximizing evaluative feedback, learners use it to update their representations of value. We call this "value feedback".

3.2.1. Reward feedback for positive or negative outcomes

Consider again the model environment in Fig. 1. In this environment, honey is a positive outcome and bees are a negative outcome. Suppose a learner reaches State 4, successfully obtaining honey. Under "reward feedback", a teacher would reward the learner because the teacher knows that entering State 4 is a smart move for obtaining honey. As should be clear however, this is not particularly useful or informative. After all, because the learner also directly experiences environmental rewards from entering State 4, the teacher's reward is redundant; a person does not need to be extrinsically rewarded for getting something that is already intrinsically rewarding to them.

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006

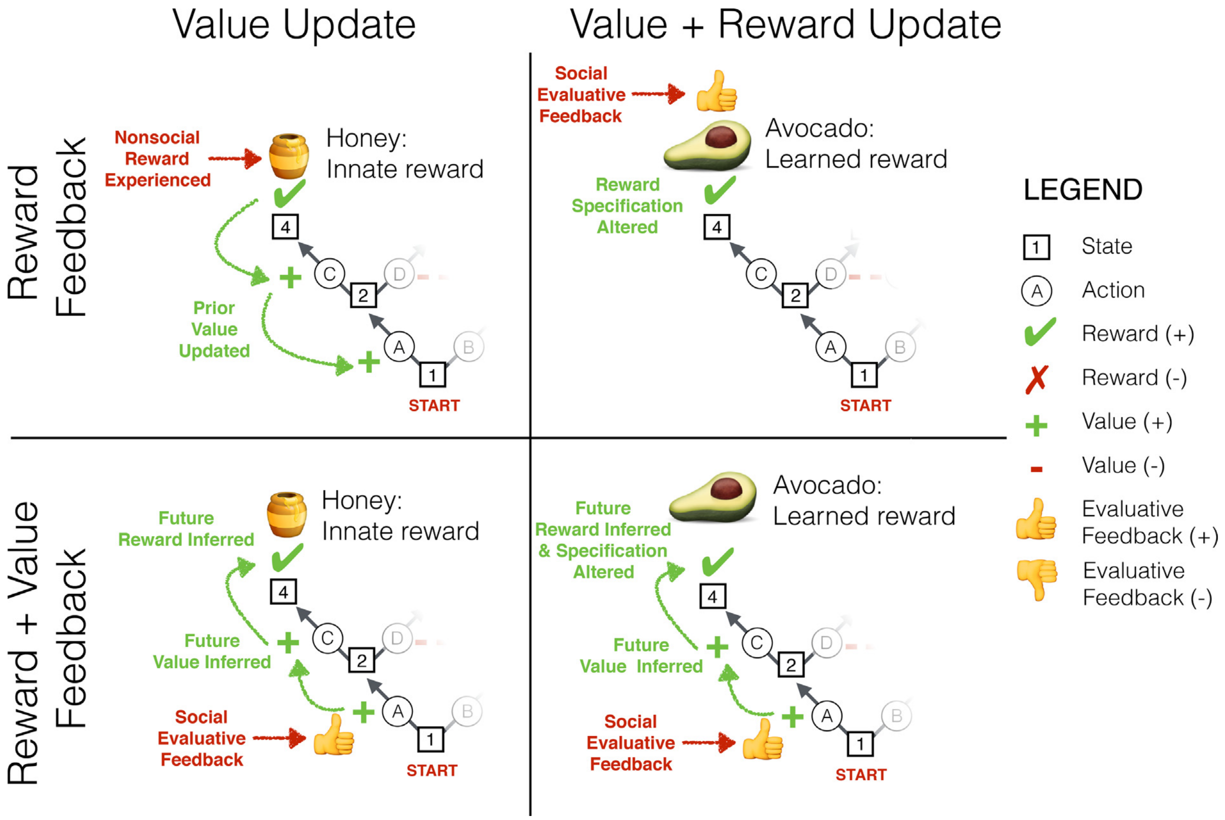

Fig. 3. Summary of adaptive and adoptive learning mechanisms discussed.

However, there are some circumstances in which reward feedback is not entirely superfluous. For instance, imagine that the bees in State 5 are usually absent, but that when they are present it is extremely unpleasant. Thus, the *average* future reward – i.e. value – of entering State 5 is negative but moderate. To teach this value, a reward feedback teacher could "smooth over" the environmental distribution of rewards by giving negative but moderate scoldings or material punishments even when the bees are absent. Although value is not being communicated directly (remember, this is still reward feedback), the correct value is more rapidly being estimated by the learner than it would be through direct experience. This indicates that reward feedback can be useful. Nevertheless, it is a somewhat roundabout method of having an agent learn value and, as we will see, indirectly inducing new value representations has serious limitations as the complexity of what is being taught increases.

Value feedback, in which evaluative feedback provides a direct commentary on the value of actions and states, has numerous advantages over reward feedback when teaching sequential behavior. Value, after all, is the recipe for obtaining reward. It is nice to be told, "You would love chocolate cake"; it is far more helpful to be informed how to bake one. As we shall see, value feedback provides more powerful and precise mechanisms to help a learner acquire a novel sequence of actions. This occurs principally because value feedback is *inferentially rich*: When a learner knows that a teacher considers an action valuable, she may infer the future rewards that underwrite that value.

3.2.2. The limits of teaching action sequences with reward feedback

Behaviorists coined the term "shaping" to refer to teaching a novel action sequence by reward feedback, and this terminology persists in contemporary reinforcement-learning approaches (Dorigo & Colombetti, 1994; Krueger & Dayan, 2009; Skinner, 1948). To train a pigeon to peck a button and eat food out of a hopper, for instance, one can "shape" its behavior by providing food for moving towards the button, then for touching the button, then for pecking it, and finally for moving close enough to the hopper to eat the food. Here, the teacher can teach the pigeon to value intermediate, but intrinsically unrewarding, actions (moving towards the button) by making them rewarding (providing food). Until the very end of this process, the value of intermediate actions derives entirely from the teacher's rewards and not from what the teacher knows is truly rewarding – eating the food in the hopper. This is directly analogous to the example of providing mild punishments even when bees are not in the bee hive in order to indirectly induce a desired value representation.

While shaping is possible both in principle and in practice, it involves some unexpected challenges. From a learner's perspective, difficulty arises in identifying what exact action or class of actions causes the teacher to deliver rewards or punishments. For example, how does a pigeon know if the reward it just received was due to walking towards the hopper, flapping its wings three seconds earlier, or any of the other vast number of actions it recently performed? Indeed, early learning theorists quickly discovered that their subjects would engage in causally inert "superstitious behavior" as a result of experiencing spurious correlations between actions and reinforcement (Skinner, 1948).

A second difficulty stems from the fact that a learner employing reward feedback will often learn to perform other, undesirable actions that maximize intermediate rewards but do not achieve the desired end state. For instance, suppose the pigeon takes two steps towards the button, and on each step receives food. If value were being directly modified by feedback then the pigeon would learn that steps towards the button have high value, so perhaps the button has high value and leads to rewards. In contrast, under reward feedback, the pigeon also assigns high value to those

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006

actions, but only because they are themselves now rewarding. Moreover, if moving towards the button is rewarding, then moving away from the button acquires high value *because it allows you to subsequently move towards the button*. (This effect is counterintuitive precisely because it violates the common sense logic of evaluative feedback.) In reinforcement learning, this is known as the problem of *positive reward cycles* – sequences of actions that begin and end in the same intermediate state yet yield a net positive reward (Ng, Harada, & Russell, 1999).

Of course, if moving away from the button is sufficiently punished by the teacher, then the positive reward cycle may be broken and the desired behavior can be induced. Nonetheless, two new obstacles arise. First, for any reward designed to make an action valuable, many punishments must be given to compensate for changes in the value of other actions. For example, if moving away from the button is inadvertently assigned high value, then the value of that action and similar actions needs to be "corrected" through punishment. This takes additional time and energy on the part of the teacher. Second, the teacher must be sufficiently cognitively sophisticated to devise a shaping policy that avoids such cycles in the first place. As a task's complexity grows, unreasonably large numbers of rewards and punishments will be needed to simultaneously teach the task and block the exploitation of positive cycles. Suppose a task like the button-pecking paradigm is divided into several stages from one side of the cage to the spot that needs to be pecked. To ensure forward progress towards the button and its eventual pressing, movements in the desired direction need to be rewarded, while movements in any undesired direction need to be punished. An agent that is trying to adapt to the teacher's reward function will attempt to explore all actions available at each stage, which requires the teacher to monitor and give the appropriate feedback for each possible action.

Recently, we illustrated the practical effects of positive reward cycles in an experimental setting (Ho, Littman, Cushman, & Austerweil, 2015a,b). Participants were asked to use rewards and punishments to train a virtual dog to reach a goal by following a specific path and avoiding certain areas. Virtually all participants rewarded the dog for walking along the path to the goal as well as reaching the goal. However, some dogs were programmed with standard reinforcement learning algorithms that embody the principle of reward feedback and so would backtrack to an earlier portion of the path to obtain more rewards from the participant. We found that people continued to reward these intermediate actions in a manner that allowed them to be exploited by the learning agent. In other words, people did not correct their evaluative feedback to match reward feedback; instead, their teaching matched value feedback. Consistent with the observation that shaping is difficult and computationally expensive, our results indicate that it does not come naturally to human teachers.

Thus, both theory and evidence indicate that shaping, the use of reward feedback to teach the value of intermediate actions, is difficult. In a way, this is because indirectly inducing a new value representation in a learner with rewards inherently requires more work and planning on the part of the teacher as opposed to simply communicating value directly. Intuitively, if a pigeon is rewarded for moving closer to a food button, it would ideally infer that this is an instrumentally valuable action—and perhaps that the food button is rewarding—and not that "moving closer to the food button" is itself source of reward.

Value feedback, of course, directly implements this preferred mechanism. As discussed in detail below, this allows evaluative feedback not just to motivate the behavior a learner has performed, but to license inferences about the value of future actions. Before presenting these advantages in detail, however, we turn to consider the second major form of "adoptive" learning: Adopting a teacher's reward specification.

3.3. From value update to reward update

In a non-social setting, an agent possesses an innate specification of reward that is the product of natural selection and updates a value representation over its lifetime. At the proximal level, there is no higher-order principle on which to reassign reward: The learning system is organized with reward as its highest-order principle. Put another way, adopting a new specification for reward to guide action would have to be based on what is already considered rewarding. But simply assigning proper value to guide action would be equally sufficient to changing one's rewards while also being responsive to changes in the environment. Additionally, if a learner is free to redefine its own rewards, there is the danger of severing the link between reward and fitness-improving behavior altogether. In summary, when interacting with a non-social environment, it does not make sense to update one's representation of reward.

In a social setting, however, reward need not be immutable. Rather, there can be good reasons to adopt the reward structure evinced by social partners. Recall that an organism's assignment of rewards determines its fitness. Some assignments will have better fitness consequences; others worse. Natural selection will favor mechanisms that reliably alter an organisms' reward function in a manner that enhances fitness. Internal to a system that learns and plans based on rewards, reward is the greatest good. But, the system itself is designed to maximize fitness, and so there may be conditions where evolution favors overwriting the specification of reward.

One obvious case is when social partners show greater signs of fitness (e.g., health, power, reproductive success, etc.). For instance, a child might evaluate which adults look healthier, and then adopt their reward preferences concerning food. If she notices that the most successful adults in her life seem to derive intrinsic pleasure from eating avocados, she would then adjust her own food preferences to derive intrinsic pleasure from eating avocados as well. A learner could also be innately predisposed to trust that certain social partners will teach or model effective reward specifications (for instance, her parents). She may also want to change what she finds rewarding when others' behavior and evaluation of her depends on her reward structure. For instance, social partners may be more likely to trust her if they conclude that she experiences the act of sharing as intrinsically rewarding, and not merely instrumentally valuable.

Internalization of others' reward specification is a form of cultural learning, and under this construal its adaptive rationale has been extensively investigated (Grusec & Goodnow, 1994; Heyes, 2016; Richerson & Boyd, 2008). Typical approaches to cultural evolution though treat stimulus-response structures as the unit of transmission, such as "Do unto others as you would have them do unto you", "Read aloud to your children every night" or "Don't wear white between Labor Day and Memorial Day". We aim to transport and elaborate this approach within the reinforcement learning framework, in which behavior is guided by value representations derived from reward. In this case, a role for cultural learning may be to acquire a new specification of reward.

Here, again, it is important to note how sharply this possibility differs from the standard reinforcement learning framework. Suppose that social rewards operated similarly to non-social rewards, and consider again what would happen when infant explores tasting an avocado and her father says, "Great job!". If the infant had no preference for avocado and was merely adapting to the contingencies of the environment, she would learn that tasting avocadoes has high value since it leads to praise. We termed this aspect of learning "value update". Importantly, there is nothing special about the fact that there are social others in this mode of learning. We can contrast this directly with if the infant assigned reward to eating avocadoes because she learned this from her father. This reward is intrinsic—the very act of eating avocado is now pleasurable in and of itself, not in a way that is contingent upon its connection to some further desirable state. We will call this mechanism "reward update".

3.3.1. Integrating value feedback and reward update

We have presented two ways in which a learner can "adopt" a teachers' mental representations (Fig. 3). First, she can use a teacher's evaluative feedback as a communicative signal of value, adopting this representation for herself. This "value feedback" method contrasts with the standard approach of treating evaluative feedback as a reward to be maximized. Second, if she can infer the teacher's specification of rewards, she may adopt the identical specification of reward for herself. This "reward update" method contrasts with the standard approach in which reward specification is innate and immutable.

These mechanisms can work in concert when a learner exploits the inferentially rich nature of value representation. Value representations are inferentially rich because they summarize the expected future reward that depends upon an action embedded in a longer organized sequence. Put more simply, they communicate that an action is instrumentally valuable with respect to some future reward. Thus, when a learner is rewarded for an action, she may not only represent that action as valuable, but also infer the valued sequence of actions that the teacher represents, along with the reward of the ultimate goal state. She may then adopt not just the local action, but also the extended sequence, and even the specification of reward that ultimately favors that sequence.

These virtues do not come for free, however. The inferences demanded on the part of the learner are mental state inferences. By interpreting the teacher's evaluative feedback as a communicative signal of value, she takes on the task of decoding its meaning: The hidden mental representations of action, sequence, goal and reward that she ultimately seeks to adopt. For the remainder of this section, we show the power afforded by successful mental state inferences under the value feedback/reward update regime. In Section 4, we detail the specific cognitive mechanisms that are necessary for such inference to succeed.

3.4. The advantages of adoptive learning mechanisms

In the non-social domain, adaptive learning mechanisms like reward feedback and value update rely on general cognitive mechanisms for inference and generalization. This can include generalization based on the similarity or category of stimuli, associating co-occurrent stimuli, or inference based on the causal structure of the world (Dickinson, 2012; Gershman & Niv, 2010). For example, if an agent learns that bees are reliable sources of aversive experiences, she may also infer that wasps will be as well in light of their similar features and causal properties. By relying on cognitive mechanisms that represent regularities and relations in the world, adaptive learning mechanisms generalize what is rewarding or aversive and what is valuable based on *environmental structure*.

Adoptive learning mechanisms, in contrast, can also leverage the *mental structures* of teachers to generalize the value and reward of new actions and states. Value feedback, in particular, is inferentially rich because of the properties of action-guiding representations of value. Reward update also allows for rich generalization because rewards themselves index invariant sources of fitness-enhancement. Here, we discuss three ways in which adoptive learning mechanisms are inferentially rich: inference from withdrawn or incomplete feedback, inference over plans, and inference across actions.

3.4.1. Inference from withdrawn or incomplete feedback

Reward feedback and value update rely on adaptive learning mechanisms that allow a teacher to shape a learner's behavior through rewards and punishments. However, once the teacher is no longer present, there is the very real possibility that the learner will unlearn everything that was taught. For instance, let us return to the example of the daughter being taught the importance of sharing by her father. Suppose that teaching is cut short and her father has only taught part of the connection between sharing and having close personal relationships. Under reward feedback and value update, in the absence of experienced rewards, the assignment of value to sharing begins to erode. Eventually, without the understanding that sharing is a means to an end other than her father's praise, the daughter stops sharing. This process, commonly known as *extinction* in psychology, severely limits the utility of learning by adapting to evaluative feedback.

In contrast, suppose the daughter could learn from evaluative feedback using adoptive mechanisms like value feedback and reward update. Reward update solves the problem of withdrawal in a straightforward manner: Once the daughter has internalized the experience of sharing as rewarding, she will want to share for its own sake. In the absence of her father's feedback, sharing remains intrinsically rewarding so she will share.

But what if the evaluative feedback is not internalized as a reward? Even in this case, value feedback still helps avoid the extinction of taught behavior. Because value feedback is interpreted as a signal about the *value* of actions, the absence of feedback is not undermining. When the daughter received positive feedback, she understood that sharing was valuable. If she had received negative feedback, this would have lowered her assignment of value. But the absence of feedback is not, itself, a signal: She has neither been told to raise nor lower her assignment of value of an action, and so it will remain the same, all else being equal. Moreover, as she continues to explore her environment and understand the relationship between sharing and other actions, states, and rewards, she may come to discover the ultimate reward that underwrites the value of sharing. Eventually, if she does not discover this, the value assigned to sharing will dissipate, but not because of the absence of her father's rewards *per se*.

In Section 4, we discuss some evidence from developmental research on learning by observation and demonstration that illustrates children's ability to persist in intermediate actions based on an assumption of long-term reward. This evidence suggests that similar mechanisms may also be brought to bear on learning from evaluative feedback.

3.4.2. Inference over plans with value feedback

Recall that value has two essential features: First, it is a summary of the rewards that could be obtained in the future by choosing optimally. Second, the value of different states and actions are mutually constrained by how they relate to one another as possible plans of action. This allows a learner to perform inference over plans from a teacher's feedback.

This can be illustrated with an example: Suppose that a mother punishes her toddler for climbing onto a chair that would allow him to climb onto the kitchen counter and play with a shiny object on the counter, which the mother knows to be a butcher's knife. Under value feedback, the child directly learns from punishment that climbing onto the chair has negative value. But climbing onto a chair has a clear intentional relationship to other actions: Once he climbs onto the chair, his salient next option is to climb onto the counter; once he climbs onto the counter, his salient next option is to play with the shiny object. In other words, the three actions—climb on chair, climb on counter, play with shiny object—constitute a single plausible plan. Since value representations across actions and states must be consistent (Section 2) this means that the values of these three actions are coupled to one another. As a result, even though the child has only directly received value feedback about the first action (climb on the chair), he can generalize to the value of the remaining actions in the sequence (climb on the counter and play with the shiny object). Thus, learning about value provides a basis for inferring future value in subsequent states or actions in a plan, even when these have not been directly experienced.

Of course, one can stipulate a representation of environmental structure that licenses the same inferences for an adaptive learner only using reward feedback. For example, in the environment of the child, chairs could co-occur with counters, and counters could co-occur with shiny objects. Upon receiving the mother's punishment after climbing the chair, the negative reward could be associated with the chair, counter, and shiny object. But any of those features of the environment could be associated with any number of other features in the environment. Why doesn't the child also assign negative reward to the refrigerator, which also always co-occurs but is unrelated to the action sequence? In contrast, value feedback offers a parsimonious way to account for inference over actions and states that are related by the mental structure of value.

Put colloquially, value feedback is especially useful when learners address the question "why am I receiving this reward/punishment?". A learner could integrate background knowledge about the environment and the teacher with new information in the form of evaluative feedback to infer the teacher's likely structure of value and reward. One way this could be accomplished is if the learner applies a theory of mind to the situation. The structure of this particular inferential problem is similar to the generic task of observing the behaviors of others and inferring their mental states, which can be formalized as the Bayesian inversion of a generative model in which mental states cause action (Baker, Saxe, & Tenenbaum, 2009; Koster-Hale & Saxe, 2013). In Section 4, we discuss this connection in more detail.

3.4.3. Inference across actions with value feedback

So far, we have considered the way that value feedback can signal information about the rewards of future actions (and, of course, the value of future actions as well). But value feedback may also signal information about the comparative value of *alternative actions*. For instance, suppose that a teacher provides evaluative feedback to a learner by rewarding desired actions and punishing all other actions. The learner could leverage this fact. If she is rewarded for doing action A out of the set of actions A-Z, for instance, she not only knows that A is valuable, she also knows that B-Z are *no more valuable* than A. In this case, she is not inferring the value (or reward) of future actions that follow from A; rather, she is inferring the value of alternative actions that might have been performed instead.

When is this form of inference valid? Clearly, it cannot work under reward feedback. A particular amount of reward obtained for one action in a given state has no necessary implications for the amount of reward obtained for another action in a given state.

In general, for the teacher's evaluative feedback to support useful inferences of this kind on the part of the learner, the teacher must choose feedback for each action as a function not only of its value representation of that action, but also of at least some unchosen actions. We have already considered one version of this idea: Reward the optimal action and punish all the suboptimal actions. There are a range of other possible methods that would license similar inferences, such as giving feedback in proportion to the teacher's probability of actually choosing the same action or in proportion to an action's similarity to the optimal action. Intuitively, all these approaches operate something like a game of "warmer/cooler" used to playfully guide a social partner to a hidden target. Moreover, this way of reasoning about actions *never taken* from feedback about actions taken parallels observational learners drawing conclusions about *undemonstrated* actions from pedagogically demonstrated actions (Bonawitz et al., 2011).

3.5. Summary

In the social setting, agents can adopt the mental structures of social partners. Teachers have their own value representations and reward specifications that are the product of natural selection and individual experience, and evaluative feedback is one way in which they can transmit them to learners. As we have discussed, however, the reward learning mechanisms that work in the non-social setting—reward feedback and value updating—have limited application in the social setting. In particular, these adaptive mechanisms only allow a teacher to use rewards and punishments to indirectly induce value in the learner through shaping. This makes it difficult or impossible for an agent to learn when being taught longer sequences of actions, when teaching is incomplete, or when the learner does not share the reward specifications of the teacher.

Value feedback and reward update, in contrast, naturally solve the problems associated with adaptive learning. Since feedback is no longer subject to maximization, teaching longer sequences of actions is not plagued by the emergence of positive reward cycles. Similarly, learners can still reap the benefits of teaching even when teaching is incomplete and value feedback is withdrawn. This is because under value feedback, the absence of feedback simply means that the learner's representation of value does not need to be changed. Finally, both value feedback and reward updating together allow a learner to adopt the teacher's reward specification. This is because value feedback permits inferences about the rewards likely motivating the teacher, while reward updating allows for their internalization.

Value feedback clarifies the intuitive notion of using evaluative feedback as a commentary on action rather than a socially-mediated form of reward to be maximized. In certain respects, the principles governing effective and efficient evaluative feedback are likely to mirror those that govern other forms of communication. Speakers and listeners can best communicate, for instance, by respecting certain shared assumptions about the nature of the discourse (Grice, 1957): That the speech act will efficiently convey true, precise and relevant information. Recent research formalizes the inductive benefits of this coordinated stance between teacher and learner (Frank & Goodman, 2012; Shafto et al., 2014). This provides a promising foundation to build formal accounts of the optimal structure of coordination between teacher and learner during evaluative feedback (Ho et al., 2015a).

Notably, although value feedback and reward update are novel ideas in the context of teaching by evaluative feedback, there is already a wealth of evidence for similar mechanisms in other domains of human learning and teaching. For example, for a learner to interpret evaluative feedback as a signal of value as opposed to a reward to be maximized, it is necessary for the learner to segregate its positive hedonic experiences into two streams depending on whether or not the relevant reward likely constitutes a pedagogical act on the part of a social partner. One stream would process non-social or non-pedagogical rewards by treating them in the ordinary way (as a hedonic experience to be maximized). When interacting with social partners with communicative intentions, however, the same experiences take on a different meaning and distinct computational role. In the following sections, we discuss this and other mechanisms in detail and how they relate to our proposal.

4. Evidence for mechanisms

For an agent to respond effectively to value feedback and to perform reward update, several cognitive abilities are required to augment traditional mechanisms of reward learning. First, humans must have the capacity to identify whether experiences of punishment and reward are being furnished by non-social sources (or non-communicative social sources) versus by communicative social agents. Second, they must have the capacity to infer the assignment of value and reward to a large set of partially unobserved states and actions from sparse samples of value from observed states and actions—i.e., to exploit the inferentially rich nature of value feedback. Third, they must have the capacity to update their own rewards (i.e., tastes, preferences, morals, etc.) based on social information—i.e., to engage in reward update. While the existence of these mechanisms has rarely been tested in the context of evaluative feedback, circumstantial evidence indicates that humans possess all three.

4.1. Identifying acts of evaluative feedback

To begin with, learners must successfully differentiate communicative evaluative feedback (a candidate for value feedback) from other events (candidates for reward feedback). This is not as simple as merely dissociating experiences dependent on their origin from animate versus inanimate sources, or even social versus non-social sources. Sometimes social rewards are not meant to communicate value. It may be, for instance, that a person's peers cannot suppress laughter when they hear him pass gas—beyond the fourth grade, such laughter is rarely an intentional act of positive evaluative feedback, however, and should not be interpreted as such. Among all the rewards and punishments that are causally related to a social other, it is necessary to identify the subset that are enacted communicatively.

Nearly a decade of research has demonstrated that adults and children are sensitive to whether an actor signals they want to communicate information. These ostensive cues include making eye-contact or saying the observer's name, and the subsequent recognition of communicative intent leads the observer to interpret the actor's behavior differently. For example, if a demonstrator signals her intent to teach an action and then performs an action, children are more likely to imitate her specific action and avoid exploring alternatives (Bonawitz et al., 2011; Sage & Baldwin, 2011). Similar work has shown that imitation is mediated by signaling communicative intent for longer sequences of actions (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011), unnecessary intermediate actions (Brugger, Lariviere, Mumme, & Bushnell, 2007), and unusual actions (Király, Csibra, & Gergely, 2013). Communicative cues also influence how children generalize functional properties (Butler & Markman, 2012; Butler & Markman, 2014), generalize preferences (Egyed, Király, & Gergely, 2013), unlearn actions (Hoehl, Zettersten, Schleihauf, Grätz, & Pauen, 2014), overimitate actions (Lyons, Young, & Keil, 2007), and themselves teach others (Vredenburgh, Kushnir, & Casasola, 2015). Similar findings have been established for adults teaching and learning about rules, categories, and causal relationships (Shafto et al., 2014).

Thus, when it comes to demonstrating actions or the properties of objects, human learners are sensitive to whether teachers signal their communicative intentions. Recognition of communicative intent leads to differential imitation and encoding, which suggests that similar processes may support learning from evaluative feedback.

4.2. Inferring and adopting unobserved values

In order for value feedback and reward update to achieve their greatest advantage, a learner must be able to infer how a teacher's value representations, reward specification, and beliefs about environment relate to each other even when these cannot be directly observed. In other words, learners must perform inference over a theory of rational action. The literature on children's cognitive development clearly demonstrates that even young infants have this capacity, and that older children leverage it during imitation learning. Specifically, humans understand how a rational agent's actions, goals, and environment mutually constrain one another such that one can infer information about one from any two of the others.

For example, after having been habituated to a geometric, puppet, or human agent moving towards another agent in a particular environmental configuration, 12-month olds and 9-month olds (but not 6-month olds) expect the agent to adapt its actions when the environment changes (Csibra, Gergely, Bɪró, Koós, & Brockbank, 1999; Gergely, Nádasdy, Csibra, & Bíró, 1995; Sodian, Schoeppner, & Metz, 2004). Woodward (1998), Woodward (1999) similarly demonstrated that infants around 6- to 9-months old expect grasping human hands, but not claws or limp hands, to be directed at objects rather than move along a certain path in space. This reflects infants' more general capacity to predict and understand the flow of other agents' behaviors in light of intentions and overarching goals (Baldwin, Baird, Saylor, & Clark, 2001; Woodward & Sommerville, 2000). These results clearly establish that even young infants have strong expectations about how invariant goals (i.e. rewards) and changing environments impact the value of actions and rational agents' propensity to perform them.

Related studies have shown that an infant's knowledge about what an agent wants and the actions it takes constrain expectations about the environment. Nine- and 12-month olds were shown an agent, a goal, and an object occluding part of the path between the starting location and goal location. When the agent moved to avoid the occluded area, measurements of looking times indicated that the infants were more surprised when it was subsequently revealed that there was no obstacle (Csibra, Bı́ró, Koós, & Gergely, 2003). These capacities continue to develop and sharpen in older children. By 5–6 years of age, children can make inferences about and design tests of other non-observable agentic properties like competence and subjective differences in rewards (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015). Similar results have been demonstrated in adults (Baker et al., 2009).

Beyond simply recognizing that others adapt their actions to pursue their goals, infants can use this interpretive schema to imitate – that is, adopt – the relevant aspects of a demonstrator's behavior. Bekkering, Wohlschlager, and Gattis (2000), for instance, showed that 3- to 5-year olds imitated actions (touching one's left/right ear with one's left/right hand) based on the space of possible goals that were shown in a set of demonstrations (one ear or both ears). When only one ear was ever touched in a session, children imitated the exact hand used, whereas when both ears were touched during a session, children did not imitate the action as closely. More intriguingly, in Meltzoff's (1995) study, 18-month olds shown a failed intentional action imitated the action and implied goal, even though the latter was never observed. Thus, even in the absence of observing a rewarding outcome, young children can treat an action as being valuable and "aimed at" future rewards or goals.

As when interpreting others' behavior, children take rational constraints on actions, goals, and the environment into account when imitating. In particular, imitating an action depends crucially on the availability of other, seemingly more valuable or efficient actions. One week after a demonstrator used an especially unusual action to accomplish a task (e.g. using one's head to turn on a light), children imitated the unusual action only when ordinary actions (e.g. using one's hands) were an option for the demonstrator (Gergely, Bekkering, & Király, 2002). That is, the action itself, rather than the goal, was imitated only when the more efficient action was available but clearly not utilized. Similarly, Lyons et al.

(2007) examined overimitation, which occurs when children imitate actions that are obviously unnecessary to accomplish a goal. Overimitation occurred even when the demonstrator left the room, it was clear that the goal needed to be reached urgently, and the participant was explicitly told not to perform unnecessary actions. These findings indicate that children encode new causal features into objects, which is equivalent to the learners inferring that the actions had value in virtue of unobserved causal properties.

In summary, ample evidence supports the existence of a mechanism in humans that learns about the value of actions as they relate to future actions and alternative actions. Much of this evidence comes from studies of children's imitative behavior. We propose that the same mechanism may help children respond rationally to communicative evaluative feedback.

4.3. Adopting others' rewards

Since early researchers started applying concepts from operant conditioning to evaluative feedback, internalization has posed a puzzle (Aronfreed, 1968; Bryan & London, 1970). If Betsy gets scolded for jumping on the bed by her father, why wouldn't she learn to start jumping once he leaves? Or, why would Ali continue saying "please" before each request long after his mother has stopped praising him? It is difficult to explain the persistence of incentivized behaviors in the absence of the incentives that motivate them. Yet, across a number of domains, it is clear that children persist in behaviors long after contingent reward and punishment disappear. This provides evidence that children possess the capacity for "reward update".

Young children readily reason about food preferences (Repacholi & Gopnik, 1997), and even 12 month-old infants often use social cues rather than intrinsic properties (e.g. color or texture) to evaluate foods (Shutts, Kinzler, McKee, & Spelke, 2009). This makes the acquisition of food preferences a valuable case study in human social learning (Shutts, Kinzler, & DeJesus, 2013). When teaching their children to eat a new food, for example, parents report using rewards for eating as frequently as any other method (Casey & Rozin, 1989). Moreover, these novel food preferences, once reinforced, can be permanently learned by children even once contingent rewards are removed (Hendy, Williams, & Camise, 2005; Lowe, Horne, Tapper, Bowdery, & Egerton, 2004). The clearest demonstration of this is given by Cooke et al. (2011), who investigated preferences for initially disliked vegetables after training with either tangible rewards (stickers) or social rewards (praise from a female experimenter). Both one month and three months after the tangible or social rewards were removed, rewarded children showed a constant greater intake of the target vegetable than children who were merely exposed to the vegetable or did not receive exposure. Relatedly, despite an innate aversion to the burning sensation associated with capsaicin (typically found in chili-peppers) many humans can eventually learn a stable preference for it through social-learning mechanisms, likely including evaluative feedback (Rozin & Schiller, 1980). This is the case even though it is extremely difficult to condition a preference for capsaicin in rats (Rozin, Gruss, & Berk, 1979). Collectively, these findings suggest that human preferences for capsaicin are internalized through a distinct, socially-mediated mechanism.

Similarly, early work on the development of altruistic behavior established that children would persist in performing initially unmotivated, prosocial actions even upon the cessation of contingent reinforcement. For example, 8- to 11- year old boys learned whether or not to donate money they won to an anonymous child via praise or scolding and continued this behavior even after two weeks (Rushton & Teachman, 1978). Subsequent research showed that the learning of prosocial behavior becomes progressively more stable and generalized between 5 and 10 years of age (Grusec & Redler, 1980). Relatedly, in a study designed to assess how mothers' childrearing behavior related to their childrens' emotional responses to sympathy-inducing films, Eisenberg et al. (1992) found greater sympathetic responses in daughters whose mothers reported using reinforcement to teach sympathy.[2]

Home studies of mother-child interactions and child behavior also show that children can internalize rules or behaviors that are taught by a mother through evaluative feedback. For instance, four- and five-year old children show evidence of internalizing norms that were taught by their mother, such as putting toys away or not playing with certain toys, as indicated by maintenance of the normative behavior in the absence of their mother. (This effect was markedly reduced among 2- to 3-year old children). Internalization does not occur, however, in learning settings where children need to receive constant rewards and punishments to perform the task. Rather, milder and sparser evaluative feedback tends to be associated with successful internalization (Kochanska & Aksan, 1995; Kochanska, Aksan, & Koenig, 1995). This is consistent with the hypothesis that children segregate specific forms of reward signals into a distinct processing stream designed around the communicative principles of evaluative feedback (Section 4.1).

4.4. Summary

Value feedback and reward update require (1) the capacity to recognize communicative intent, (2) the capacity to infer rewards from cues about value, and (3) the capacity to internalize new reward assignments. Although these abilities have not been extensively studied in the context of evaluative feedback, there is clear evidence for them from research on teaching by demonstration, imitation learning, and preference learning.

5. Evaluative feedback among humans and non-human animals

We have argued that human evaluative feedback should, and does, function by adopting the value representations and reward specifications of social partners. We have further argued that certain familiar aspects of learning from evaluative feedback, such as learning longer action sequences or internalizing new preferences, are difficult or impossible without dedicated social learning mechanisms. This includes a sophisticated ability to perform inference over a theory of rational action—i.e., theory of mind. This claim has a natural corollary: teaching with and learning from evaluative feedback should be much rarer or at least limited in non-human species that lack the ability to identify pedagogical intent, infer other's unobservable mental representations, or adopt others' preferences. We begin by situating this claim among current theories of animal teaching and learning, and then compare it against current evidence.

Current theories of social behavior in nonhuman animals agree that cognitive constraints limit the use of reciprocal reward and punishment (Hammerstein, 2003; Stevens & Hauser, 2004; Stevens et al., 2005). For instance, Raihani et al. (2012) argue that "the absence of language to explain the rationale for punishment" accounts for the dearth of evidence that animals punish for defection or free-riding. Of course, language can greatly facilitate reasoning about the rationale for social rewards and punishments, but we would suggest that it is often possible for humans to infer the likely communicative intent of reward and punishment even in the absence of explicit language. Perhaps the lack of linguistic communication in non-human animals is not the root cause, but a fur-

[2] We wish to note that these results are different from and consistent with work showing that *initially intrinsically* motivated behaviors can be undermined by extrinsic rewards (e.g. Warneken & Tomasello, 2014; Deci, Koestner & Ryan, 1999).

ther symptom of a more pervasive constraint on non-human animals to reason about communicative intent—a constraint that, as we have shown, severely limits the potential for useful social evaluative feedback.

In another point of convergence, there is an emerging consensus that associative learning processes can support at least some social learning of actions via observation (Heyes, 2012; Leadbeater, 2015). But, while many previous researchers have asked whether teaching in general occurs in non-human animals, they have generally focused on teaching that occurs by means other than evaluative feedback—for instance, teaching by demonstration, by intervention, or by constructing situations beneficial to learning. Skerry, Lambert, Powell, and McAuliffe (2013), for instance, primarily discuss teaching in humans and non-humans in the context of instruction and demonstration. Heyes (2016) examines social learners' use of meta-cognition to choose which conspecifics to imitate. And while Kline (2015) does address "teaching by evaluative feedback", the term is used to refer to a much broader class of teaching behaviors than rewarding or punishing a leaner's actions. We pursue a complementary approach by focusing specifically on the limitations of "traditional" non-social learning reinforcement learning mechanisms in the domain of evaluative feedback, contrasting this with the potential of value feedback and reward update.

Despite considerable interest in animal social learning—and in contrast to the robust literature on animal imitation—there is remarkably little evidence of widespread teaching and learning by evaluative feedback in non-human animals. This contradicts early predictions by ethologists and comparative researchers. Caro and Hauser (1992), for instance, predicted two forms of teaching in animals: opportunity provisioning, in which a learner is given a sequence of progressively more complex opportunities to practice a task, and "coaching", which is identical to what we term teaching by evaluative feedback. Although opportunity provisioning has been observed in several taxonomically unrelated species (Caro, 1980; Caro, 1994; Thornton & McAuliffe, 2006), non-anecdotal evidence for coaching has been harder to come by (we discuss some of the existing findings below). Similarly, although theorists studying animal cooperation predicted that rewards from cooperation and punishments for defections could reinforce cooperative behaviors in animals (Clutton-Brock & Parker, 1995; Trivers, 1971), empirical support for these claims is surprisingly sparse (Hammerstein, 2003; Raihani et al., 2012; Stevens & Hauser, 2004; Stevens et al., 2005).

Below, we pursue a simple explanation: The psychological mechanisms necessary to implement value feedback and reward update are much more developed among humans than non-human animals. Thus, evaluative feedback is familiar among humans, but surprisingly less conspicuous among non-human animals.

5.1. Evaluative feedback is ubiquitous and successful among humans

Rewarding cooperation and punishing defection is an effective way to enforce cooperative norms in experiments with people (Fehr & Gächter, 2002) and is arguably a universal feature of human societies (Henrich et al., 2004). In fact, humans are typically described as exceptional in the frequency, scope and magnitude of social reward and punishment (Hammerstein, 2003; Raihani et al., 2012; Stevens & Hauser, 2004; Stevens et al., 2005).

Likewise, as most parents can attest, the use of rewards and punishments to coach behaviors in children is a valuable and perhaps indispensable feature of the human parental repertoire. Much empirical data supports this common knowledge. For example, Owen et al. (2012) reviewed 41 studies over the last several decades focusing on the effect of verbal and non-verbal evaluative feedback on children's compliance with parental directives or instructions. Across naturalistic and experimental studies, the authors concluded that the use of evaluative feedback was not only commonplace but also effective at making children more compliant. The development of prosocial behaviors such as helping, sharing, and comforting others is another major area of study, mostly with observational methods. As with compliance to directives, parents shape the prosocial behavior of their children using different types of evaluative feedback. For instance, children who receive social approval or praise for spontaneous prosocial acts at home (e.g. cleaning up the dishes) are more likely to engage in prosocial behavior at school (Garner, 2006). Similarly, 4-year olds show more prosocial behavior when given explicit approval or praise versus no explicit response (Grusec, 1991).

5.2. Evaluative feedback is rarely observed among non-humans

In the animal literature, evaluative feedback has been explored in two domains. The first is cooperation between non-kin, while the second is "coaching" of an animal's offspring. Facilitating cooperation with rewards and punishments can be seen as a form of 'horizontal' behavior modification that results in fitness benefits for both animals, while coaching can be seen as a 'vertical' analogue that benefits offspring directly and the teacher indirectly. Yet, while both of these strategies have obvious adaptive value, they have been hard to reliably identify in non-human animals.

Game-theoretic analyses of cooperative situations predict that punishments and rewards by one partner can shape cooperation in the other. Clutton-Brock and Parker (1995), for example, predict that cooperation among non-kin in a species could be sustained though punishment or sanctioning of defectors. By imposing a cost on defection, there is less of an incentive to do so, leading to greater cooperation. Along similar lines, Trivers' (1971) theory of reciprocal altruism predicts that individuals could reward one another's behavior to sustain cooperation. The expectation of a future reciprocated reward theoretically serves to make an action that benefits another more valuable. Yet, counter to these predictions, there is little evidence that non-human animals use reward and punishment to motivate one another to cooperate, and not defect, in future interactions. Rather, the majority of cooperation in the literature can be explained by mechanisms that do not require incentivizing another organism's future behavior, such as mutualism or kin-selection (Hammerstein, 2003; Raihani et al., 2012; Stevens & Hauser, 2004; Stevens et al., 2005). And, aggression in non-human animals is more often directed at changing immediate behavior (e.g., reclaiming territory, establishing dominance or vying for a mate) than modifying future behavior (Stevens, 2004).

Similarly, there is surprisingly limited evidence that non-human animals use reward and punishment to teach their young. Here, again, it is crucial to distinguish between opportunity provisioning and coaching (Caro & Hauser 1992). There is unambiguous evidence of opportunity provisioning, which occurs when an adult provides young with opportunities to practice a task without the dangers normally associated with a task (Caro, 1980; Caro, 1994). A compelling example is that of adult meerkats, who will provide their pups with dead or disabled scorpions for practice (Thornton & McAuliffe, 2006).

In contrast, empirical support for coaching, which Caro and Hauser describe as when a teacher directly "alters the behavior of [a learner] by encouragement or punishment", is relatively poor. Much of the cited evidence, particularly in primates, comes from anecdotal reports by field researchers. For instance, Fletemeyer (1978) reported seeing a high-ranking male baboon eating a fruit laced with poison and subsequently threatening sub-adults and juveniles attempting to eat the food. Caro and Hauser (1992) and

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006

Boinski and Fragaszy (1989) report having observed similar coaching events in vervet monkeys and squirrel monkeys, respectively. One of the few attempts to quantitatively validate these reports studied whether rhesus and pigtailed macaques mothers used "puckering" to encourage their offspring to walk towards them (Maestripieri, 1995; Maestripieri, 1996). Out of 12 weeks of observation, there was only a single week that offspring who received puckering walked significantly more than those who did not. Moreover, latency in walking following puckering was neither related to infant age nor to number of previous puckers (Maestripieri, 1996). This provides, at best, only suggestive evidence that non-human primates teach using encouragement. The lack of evidence is also consistent with the general finding that social learning in non-human primates occurs primarily through curious learners and not active teachers (Thornton & Raihani, 2008).

Quantitative studies on coaching in non-primates similarly provide limited evidence that animals teach with reward and punishment. For example, Nicol and Pope (1996) reported that when separated maternal hens saw that their chicks were eating poisoned food, they engaged in more pecking and scratching at the ground, followed by demonstrations of "correct" behavior. However, the researchers did not rule out explanations other than coaching. For instance, the results could be explained by maternal hens drawing their chicks' attention away from the poisoned food and towards a demonstration of eating the unpoisoned food rather than the presentation of an aversive stimulus.

A report by Raihani and Ridley (2008) may provide evidence for teaching with reward and punishment among wild pied babblers. These birds produce a specific purr call when feeding nestlings and later use the same purr call to induce them to approach new food sources. This suggests that the chick associates food with the purr sound, which is later used by the parent for drawing the chick to novel food sources. This example qualifies as teaching, and it involves reward and punishment. However, it differs markedly from "coaching" in that it resembles an application of classical conditioning rather than operant conditioning. Indeed, many studies that purport to demonstrate coaching involve a parent 'luring' offspring towards themselves (e.g. Maestripieri, 1995; Maestripieri, 1996), rather than giving rewards or punishments *following* an action.

In the case that comes closest to true coaching, reported by West and King (1988), a female cowbird responds to a male cowbird's song with a 'wing stroke' (a suggestive copulatory gesture) the male sings the rewarded song more frequently. This example also qualifies as teaching in that the female cowbird controls the male cowbird's behavior through feedback, but it centers on the highly restricted and domain-specific context of a courtship ritual. In contrast, our analysis, as well as Caro and Hauser's, characterizes the logic of teaching with reward and punishment as an adaptation that facilitates knowledge or skill transmission.

Of course, it may just be difficult to isolate or detect true teaching by evaluative feedback in non-human animals. Thus, the paucity of data may only reflect the limits of the available scientific tools. But, if we take the current data at face value, it suggests that evaluative feedback is widespread among humans and much rarer in non-human animals.

A likely explanation is that humans possess a suite of especially powerful cognitive capacities for mental state inference, communication and internalization that together make evaluative feedback an adaptively favored behavior. Without the capacity to distinguish communicative signals, reason about and learn from others' value representations, and internalize others' preferences, evaluative feedback becomes inefficient and costly, and so is less likely to evolve as a teaching mechanism.

There is some evidence that other species lack the first two of these criteria (Csibra & Gergely, 2009; Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009). In particular, Call, Carpenter, and Tomasello (2005) used Meltzoff's (1995) paradigm to compare childrens' and chimpanzees' propensity to imitate incomplete or failed actions. As in the original study, human children copied the particular way in which an action was performed even if it failed to lead to a result. Chimpanzees, on the other hand, primarily learned to recreate the environmental results of an action (often called emulation) rather than the performed action itself. Surprisingly, chimpanzee learners in these studies appeared to understand what conspecifics were doing and could glean information from others' behaviors – e.g. learning to not perform an ineffective action. This suggests that chimpanzees spontaneously learn and reason about rewards and directly observable causal affordances in the environment, but less adeptly reason about another agent's value representations.

5.3. Summary

Although there is considerable evidence that other animals engage in social learning behaviors like observational learning or opportunity provisioning, a review of the literature suggests that teaching by evaluative feedback is much less common. Our account provides an explanation. Namely, evaluative feedback is evolutionarily beneficial primarily when learners can recognize communicative intent and reason about a teacher's mental states.

6. Evaluative feedback in human-robot interaction

A major goal of current robotics and reinforcement-learning research is to design algorithms that learn effectively from human interaction—for instance, from human rewards and punishments. By design, standard reinforcement-learning algorithms can learn behavior based on non-social rewards and punishments such as the "reward" of scoring points in Atari video games (Mnih et al., 2015). It was initially assumed that the same algorithms that respond effectively to non-social rewards and punishments could also be used to respond to human evaluative feedback. Early approaches applied commonly used algorithms such as Q-learning (a model-free method) or Rmax (a model-based method), treating evaluative feedback as a form of environmental reward to be maximized. This type of training corresponds to the standard formula of reward feedback plus value update.

Unfortunately, it doesn't work. Several studies show poor performance by standard reinforcement-learning algorithms trained by people, typically for the reasons anticipated above. First, during teaching with evaluative feedback, people nearly always produce positive reward cycles (Ho et al., 2015b) which entails that only reinforcement-learning agents that care about immediate rewards and ignore long-term rewards can be successfully trained (Knox & Stone, 2015). Second, in ongoing social interactions, people frequently withdraw feedback once learning is successful (Isbell et al., 2001). This directly contradicts the goal of these algorithms to maximize rewards, which eventually leads to an erosion of learned behavior – i.e. extinction.

Such algorithms need not fail in principle; rather, their failure reflects the structure of human evaluative feedback. For instance, if the 'teacher' is a computer program that can be automated to deliver rewards and punishments over the entire environment, positive reward cycles can be successfully avoided (Devlin & Kudenko, 2012; Ng et al., 1999). This involves the application of evaluative feedback that conserves the net amount of reward acquired by returning to a state. Yet, the computational demands of choosing evaluative feedback in real-time interaction appropriately turn out to be severe (see the discussion in Section 3.2.2).

Today's most effective, state-of-the-art algorithms for learning from human evaluative feedback instead tend to combine some form of value feedback and reward update. For instance, research into interactive teaching found that people did not give feedback contingent on *past behavior*, but often use it to signal information about *future behavior* (Thomaz & Breazeal, 2006; Thomaz & Breazeal, 2008). Loftin et al. (2014) developed learning algorithms that leverage this observation explicitly by using rewards and punishments to infer and adopt the reward specification the teacher is trying to teach.

In short, the recent history of building machines that learn from human-delivered rewards and punishments has shifted from shaping rewards (reward feedback/value update) to commentary on immediate behavior (value feedback/value update) to commentary on potential future behavior (value feedback/reward update). This is because of the difficulties of defining and providing feedback over the entire environment posed by shaping, the need to maintain behavior during the withdrawal of feedback, and the inductive richness of modifying value directly.

7. Conclusion

Psychologists once attempted to explain how human social learning was accomplished through reinforcement (Aronfreed, 1968; Bryan & London, 1970; Sears, Maccoby & Levin,1957), but this program of research died decades ago. We have attempted a *post-mortem* diagnosis: The research program was unfulfilled because it was widely assumed that mechanisms of learning from evaluative feedback are identical to mechanisms of learning from non-social rewards and punishments. Theoretical predictions predicated on this assumption failed to explain the empirical data, and so psychologists mostly gave up and moved on. In subsequent decades, far more research has been directed at alternative forms of social learning. These include learning via teleological inference (Csibra et al., 2003), imitation learning (Meltzoff, 1995), or learning from testimony (Koenig & Harris, 2005) among many others. Yet parents still punish, coaches still cajole, lovers still pout, and so on. These behaviors demand explanation.

Using concepts borrowed from reinforcement learning, we have argued that human evaluative feedback can be understood as a communication about value, rather than shaping using rewards. While the lonely world of adaptive learning is sensible in non-social settings, it becomes limited and puzzling when extended to *social* rewards and punishments. Adoptive learning is more suited for two reasons. First, value representations are *inferentially rich*, allowing for learning more from less. Second, teachers may have *more reliable mental structures*, including not only their knowledge of value, but also their specifications of reward. Learners can benefit from this by successfully inferring and internalizing a teacher's reward specifications. Because this relies on cognitive capacities particularly developed in humans (e.g. Theory of Mind), this can explain why evaluative feedback is widespread in humans but relatively rare in non-human animals. It also aligns with state-of-the-art methods in interactive reinforcement learning and robotics. Overall, this perspective not only gives us insight into evaluative feedback in humans, animals, and machines, but also helps us understand what links together the many varieties of human teaching and social learning.

Funding

Mark K Ho is supported by the National Science Foundation – United States Graduate Research Fellowship under Grant No. (DGE-1058262).This research was supported in part by grant N00014-14-1-0800 from the Office of Naval Research and grant IIS-1637614 from the National Science Foundation – United States.

Acknowledgements

We wish to thank Hyowon Gweon, Adam Morris, and William E. Moore III for their invaluable thoughts and feedback on earlier drafts of this manuscript.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

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

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

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72*(3), 708–717.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594–615.

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology Section A, 53*(1), 153–164.

Boinski, S., & Fragaszy, D. M. (1989). The ontogeny of foraging in squirrel monkeys, Saimiri oerstedi. *Animal Behaviour, 37*(Part 3), 415–428.

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*(3), 322–330.

Brugger, A., Lariviere, L. A., Mumme, D. L., & Bushnell, E. W. (2007). Doing the right thing: Infants' selection of actions to imitate from observed event sequences. *Child Development, 78*(3), 806–824.

Bryan, J. H., & London, P. (1970). Altruistic behavior by children. *Psychological Bulletin, 73*(3), 200–211.

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

Butler, L. P., & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. *Child Development, 83*(4), 1416–1428.

Butler, L. P., & Markman, E. M. (2014). Preschoolers use pedagogical cues to guide radical reorganization of category knowledge. *Cognition, 130*(1), 116–127.

Call, J., Carpenter, M., & Tomasello, M. (2005). Copying results and copying actions in the process of social learning: chimpanzees (Pan troglodytes) and human children (Homo sapiens). *Animal Cognition, 8*(3), 151–163.

Caro, T. M. (1980). Predatory behaviour in domestic cat mothers. *Behaviour, 74*(1/2), 128–148.

Caro, T. M. (1994). *Cheetahs of the serengeti plains: Group living in an asocial species*. University of Chicago Press.

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

Casey, R., & Rozin, P. (1989). Changing children's food preferences: Parent opinions. *Appetite, 12*(3), 171–182.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373*(6511), 209–216.

Cook, D. G., & Carew, T. J. (1986). Operant conditioning of head waving in Aplysia. *Proceedings of the National Academy of Sciences, 83*(4), 1120–1124.

Cooke, L. J., Chambers, L. C., Añez, E. V., Croker, H. A., Boniface, D., Yeomans, M. R., & Wardle, J. (2011). Eating for pleasure or profit the effect of incentives on children's enjoyment of vegetables. *Psychological Science, 22*(2), 190–196.

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science, 27*(1), 111–133.

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

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition, 72*(3), 237–267.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18*(2), 185–196.

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*(6), 627–668.

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

Dickinson, A. (2012). Associative learning and animal cognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 367*(1603), 2733–2742.

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron, 80*(2), 312–325.

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006

Dorigo, M., & Colombetti, M. (1994). Robot shaping: Developing autonomous agents through learning. *Artificial Intelligence, 71*(2), 321–370.

Egyed, K., Király, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science, 24*(7), 1348–1353.

Eisenberg, N., Fabes, R. A., Carlo, G., Troyer, D., Speer, A. L., Karbon, M., & Switzer, G. (1992). The relations of maternal practices and characteristics to children's vicarious emotional responsiveness. *Child Development, 63*(3), 583–602.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137–140.

Fletemeyer, J. R. (1978). Communication about potentially harmful foods in free-ranging chacma baboons, Papio ursinus. *Primates, 19*(1), 223–226.

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

Garner, P. W. (2006). Prediction of prosocial and emotional competence from maternal behavior in African American preschoolers. *Cultural Diversity and Ethnic Minority Psychology, 12*(2), 179–198.

Gelfand, D. M., Hartmann, D. P., Cromer, C. C., Smith, C. L., & Page, B. C. (1975). The effects of instructional prompts and praise on children's donation rates. *Child Development*, 980–983.

Gergely, G., Bekkering, H., & Király, I. (2002). Developmental psychology: Rational imitation in preverbal infants. *Nature, 415*(6873). 755-755.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165–193.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology, 20*(2), 251–256.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, 377–388.

Grusec, J. E. (1991). Socializing concern for others in the home. *Developmental Psychology, 27*(2), 338–342.

Grusec, J. (1997). *Parenting and children's internalization of values: A handbook of contemporary theory*. New York: J. Wiley.

Grusec, J. E., & Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology, 30*(1), 4.

Grusec, J. E., & Redler, E. (1980). Attribution, reinforcement, and altruism: A developmental analysis. *Developmental Psychology, 16*(5), 525.

Guttman, N. (1953). Operant conditioning, extinction, and periodic reinforcement in relation to concentration of sucrose used as reinforcing agent. *Journal of Experimental Psychology, 46*(4), 213–224.

Hammerstein, P. (2003). Why is reciprocity so rare in social animals? A protestant appeal. In *Genetic and cultural evolution of cooperation* (pp. 83–93). Cambridge, MA, US: MIT Press.

Hendy, H. M., Williams, K. E., & Camise, T. S. (2005). "Kids Choice" School lunch program increases children's fruit and vegetable acceptance. *Appetite, 45*(3), 250–263.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies (OUP catalogue)*. Oxford University Press.

Heyes, C. (2012). What's social about social learning? *Journal of Comparative Psychology, 126*(2), 193.

Heyes, C. (2016). Who knows? Metacognitive social learning strategies. *Trends in Cognitive Sciences, 20*(3), 204–213.

Ho, M. K., Littman, M. L., Cushman, F., & Austerweil, J. L. (2015a). Evaluative feedback: Reinforcement or communication? poster presented at the multi-disciplinary conference on reinforcement learning and decision making, Edmonton, Canada: Alberta.

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

Hoehl, S., Zettersten, M., Schleihauf, H., Grätz, S., & Pauen, S. (2014). The role of social interaction and pedagogical cues for eliciting and reducing overimitation in preschoolers. *Journal of Experimental Child Psychology, 122*, 122–133.

Isbell, C., Shelton, C., Kearns, M., Singh, S., Stone, P. (2001). Cobot: A social reinforcement learning agent. In *5th intern. conf. on autonomous agents*.

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.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research, 4*, 237–285.

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

Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *The Behavioral and Brain Sciences, 38*, e31.

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

Kochanska, G., & Aksan, N. (1995). Mother-child mutually positive affect, the quality of child compliance to requests and prohibitions, and maternal control as correlates of early internalization. *Child Development, 66*(1), 236–254.

Kochanska, G., Aksan, N., & Koenig, A. L. (1995). A longitudinal study of the roots of preschoolers' conscience: Committed compliance and emerging internalization. *Child Development, 66*(6), 1752–1769.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*(6), 1261–1277.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848.

Krueger, K. A., & Dayan, P. (2009). Flexible shaping: How learning in small steps helps. *Cognition, 110*(3), 380–394.

Leadbeater, E. (2015). What evolves in the evolution of social learning? *Journal of Zoology, 295*(1), 4–11.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience, 35*, 287–308.

Lin, A., Adolphs, R., & Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. *Social Cognitive and Affective Neuroscience, 7*(3), 274–281.

Littman, M. L., & Ackley, D. H. (1991). Adaptation in constant utility non-stationary environments. In *ICGA* (pp. 136–142).

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

Lowe, C. F., Horne, P. J., Tapper, K., Bowdery, M., & Egerton, C. (2004). Effects of a peer modelling and rewards-based intervention to increase fruit and vegetable consumption in children. *European Journal of Clinical Nutrition, 58*(3), 510–522.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, 104*(50), 19751–19756.

Maccoby, E. E. (1992). The role of parents in the socialization of children: An historical overview. *Developmental Psychology, 28*(6), 1006–1017.

Maestripieri, D. (1995). Maternal encouragement in nonhuman primates and the question of animal teaching. *Human Nature, 6*(4), 361–378.

Maestripieri, D. (1996). Maternal encouragement of infant locomotion in pigtail macaques, Macaca nemestrina. *Animal Behaviour, 51*(3), 603–610.

Marr, D. (1982). *Vision: A computational approach*. Freeman.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*(5), 838.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc. 16th y* (Vol. 99, pp. 278–287).

Nicol, C. J., & Pope, S. J. (1996). The maternal feeding display of domestic hens is sensitive to perceived chick error. *Animal Behaviour, 52*(4), 767–774.

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

Populin, L. C., & Yin, T. C. T. (1998). Behavioral studies of sound localization in the cat. *Journal of Neuroscience, 18*(6), 2147–2160.

Raihani, N. J., & Ridley, A. R. (2008). Experimental evidence for teaching in wild pied babblers. *Animal Behaviour, 75*(1), 3–11.

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution, 27*(5), 288–295.

Randolph, M. C., & Brooks, B. A. (1967). Conditioning of a vocal response in a chimpanzee through social reinforcement. *Folia Primatologica, 5*(1–2), 70–79.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology, 33*(1), 12–21.

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Rozin, P., Gruss, L., & Berk, G. (1979). Reversal of innate aversions: Attempts to induce a preference for chili peppers in rats. *Journal of Comparative and Physiological Psychology, 93*(6), 1001–1014.

Rozin, P., & Schiller, D. (1980). The nature and acquisition of a preference for chili pepper by humans. *Motivation and Emotion, 4*(1), 77–101.

Rushton, J. P., & Teachman, G. (1978). The effects of positive reinforcement, attributions, and punishment on model induced altruism in children. *Personality and Social Psychology Bulletin, 4*(2), 322–325.

Sage, K. D., & Baldwin, D. (2011). Disentangling the social and the pedagogical in infants' learning about tool-use. *Social Development, 20*(4), 825–844.

Sears, R. R., Maccoby, E. E., & Levin, H. (1957). *Patterns of child rearing* (Vol. vii) Oxford, England: Row, Peterson and Co.

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.

Shutts, K., Kinzler, K. D., & DeJesus, J. M. (2013). Understanding infants' and children's social learning about foods: Previous research and new prospects. *Developmental Psychology, 49*(3), 419–425.

Shutts, K., Kinzler, K. D., McKee, C. B., & Spelke, E. S. (2009). Social information guides infants' selection of foods. *Journal of Cognition and Development, 10*(1–2), 1–17.

Skerry, A. E., Lambert, E., Powell, L. J., & McAuliffe, K. (2013). The origins of pedagogy: Developmental and evolutionary perspectives. *Evolutionary Psychology, 11*(3), 500–572.

Skinner, B. F. (1948). "Superstition" in the pigeon. *Journal of Experimental Psychology, 38*(2), 168–172.

Sodian, B., Schoeppner, B., & Metz, U. (2004). Do infants apply the principle of rational action to human agents? *Infant Behavior and Development, 27*(1), 31–41.

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006

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

Stevens, J. R. (2004). The selfish nature of generosity: Harassment and food sharing in primates. *Proceedings of the Royal Society of London B: Biological Sciences, 271*(1538), 451–456.

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

Stevens, J. R., & Hauser, M. D. (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences, 8*(2), 60–65.

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

Thomaz, A. L., & Breazeal, C. (2006). Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *AAAI* (Vol. 6, pp. 1000–1005).

Thomaz, A. L., & Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence, 172*(6–7), 716–737.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. The Psychological Review: Monograph Supplements, 2(4), i–109.

Thornton, A., & McAuliffe, K. (2006). Teaching in wild meerkats. *Science, 313*(5784), 227–229.

Thornton, A., & Raihani, N. J. (2008). The evolution of teaching. *Animal Behaviour, 75*(6), 1823–1836.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.

Tyndale-Biscoe, H. (2005). *Life of marsupials*. CSIRO Publishing.

Vredenburgh, C., Kushnir, T., & Casasola, M. (2015). Pedagogical cues encourage toddlers' transmission of recently demonstrated functions to unfamiliar adults. *Developmental Science, 18*(4), 645–654.

Warneken, F., & Tomasello, M. (2014). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Motivation Science, 1*(S), 43–48.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Evolutionary explanations for cooperation. *Current Biology, 17*(16), R661–R672.

West, M. J., & King, A. P. (1988). Female visual displays affect the development of male song in the cowbird. *Nature, 334*(6179), 244–246.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2417–2428.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1–34.

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development, 22*(2), 145–160.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science, 11*(1), 73–77.

Please cite this article in press as: Ho, M. K., et al. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.006