International Journal of Social Robotics manuscript No.
(will be inserted by the editor)

How humans teach agents

A new experimental perspective

W. Bradley Knox · Brian D. Glass · Bradley C. Love · W. Todd Maddox · Peter Stone

Received: date / Accepted: date

Abstract Human beings are a largely untapped source of in-the-loop knowledge and guidance for computational learning agents, including robots. To effectively design agents that leverage available human expertise, we need to understand how people naturally teach. In this paper, we describe two experiments that ask how differing conditions affect a human teacher's feedback frequency and the computational agent's learned performance. The first experiment considers the impact of a self-perceived teaching role in contrast to believing one is merely critiquing a recording. The second considers whether a human trainer will give more frequent feedback if the agent acts less greedily (i.e., choosing actions believed to be worse) when the trainer's recent feedback frequency decreases. From the results of these experiments, we draw three main conclusions that inform the design of agents. More broadly, these two studies stand as early examples of a nascent technique of using agents as highly specifiable social entities in experiments on human behavior.

Keywords human-agent interaction · human teachers · shaping · single agent learning · reinforcement learning · misbehavior · attention

W. Bradley Knox and Peter Stone
University of Texas at Austin
Department of Computer Science
E-mail: {bradknox, pstone}@cs.utexas.edu

Brian D. Glass and W. Todd Maddox
University of Texas at Austin
Department of Psychology
E-mail: glass@mail.utexas.edu, maddox@psy.utexas.edu

Bradley C. Love
University College London
Department of Cognitive, Perceptual and Brain Sciences
E-mail: b.love@ucl.ac.uk

1 Introduction

Before agents begin learning a task, much engineering by technical experts goes into the agents' task-specific designs. However, human beings are a largely untapped source of knowledge and guidance *during* the learning process. We aim to create agents with relatively natural interfaces that allow anyone—including non-programmers—to guide the learning process of an agent. But to effectively design such agents, we need to understand how people naturally teach.

In this paper, we ask how human teachers can be affected by changes in their beliefs and in the pupils' behaviors. Specifically, we examine effects on the frequency with which teachers give feedback and the quality of the behavior learned from their feedback. We describe two experiments, each of which addresses this question by using a computational agent as the pupil of a human subject. The agents, described in Section 2, are built within the TAMER framework for agents that can be shaped by human trainers' reward and punishment signals [18]. For these experiments, we vary the conditions under which the humans teach and then look for differences in training statistics and agent performance.

In what we call the *critique experiment*, which tests the impact of taking on the role of teacher, there are two conditions: one in which subjects know the agent is learning from their feedback and another in which subjects believe they are merely critiquing a recording of a learning agent. We predicted that the participants' assumed roles would affect what the agents learn from the resulting feedback and the frequency at which trainers give feedback. Against our intuitions, the results of the critique experiment indicate that changing the trainer's role has little on these two dependent variables. These results suggest that either the quality of the trainers' feedback was not greatly altered by whether they considered themselves to be teaching or that the learning agents were robust to such changes in training.

Attempting to directly study the relationship between trainer engagement and the agent's learned task performance, we conducted a second experiment wherein the agent in one condition directly responds to changes in recent feedback frequency. This experiment, called the *feedback-frequency experiment*, considers whether a human trainer will give feedback more frequently if the agent acts less greedily (i.e., "exploring" or choosing actions believed to be worse) when the trainer's recent feedback frequency decreases. The results indicate that tying non-greedy action to a trainer's feedback frequency increases the overall frequency—and thus, the number of learning samples available. However, the effect on performance is unclear, which we discuss later. The feedback-frequency experiment yields two contributions that inform the design of agents than can learn from human teachers.

First, these results provide a strategy for increasing trainer engagement—lowering performance, especially when engagement drops—that could be incorporated in any agent that learns from a human teacher. Traditionally, learning agents receive feedback from encoded objective functions, called "reward functions" in reinforcement learning (RL) [34]; reward functions give regular feedback after each discrete time step. But human teachers are more complex than an encoded objective function—creating new challenges for learning—and yet can be more effective, especially given their ability to adapt to their pupils. The experiment described here adds to the currently small base of knowledge on how to create agents whose learning algorithms and behavior are designed with a respect for human trainers' strengths and limitations.

The second contribution of the feedback-frequency experiment is a proof-by-example that the common practice of categorizing all actions as either exploitation—greedily choosing actions currently thought to be best for the task—or exploration—trying other actions to learn whether they are actually superior—is insufficient when a human is in the learning loop. Since the human is reacting and adapting to the agent, the agent can take actions to intentionally affect the human's behavior. Rather than exploiting to get the highest appraisal or exploring to try new actions, the agent's actions might instead be used to communicate to, or even reinforce behavior of, the human trainer.

Additionally, this paper comprises a more general contribution. Social agents, including social robots, provide an emerging opportunity to study human social behavior [6, 9]. A computational agent's behavior can be parametrized and recorded much more thoroughly than can a human's behavior. Thus such studies allow more controlled conditions at the potential cost of less authentic interactions, yielding a different perspective from studies that use humans opposite the subjects, a perspective that has its own strengths and weaknesses. As our final contribution, these experiments illustrate this new experimental method, providing an instantiation of the previously unexplored version in which the agent learns during an interaction that is itself affected by the learning (i.e., socially-guided machine learning [36]). Also, we discuss the motivation for studying human behavior through human-agent interaction in the context of these experiments.

The remainder of the paper is organized as follows. In Section 2, we describe the learning paradigm and algorithm used in the experiment. Section 3 explains the experimental designs and results, which are then discussed in Section 4 along with our observations from the general practice of studying human behavior with interactive agents. Section 2.1 contains a discussion of related work.

2 Background and related work

In this section, we motivate our experiments, first by discussing related work in Section 2.1. Then in Section 2.2 we give the background on TAMER and the task that the TAMER agents are taught.

2.1 Related work

2.1.1 Agents learning from human teachers

The field of agents that learn from humans is young but already has a rich and varied literature. The most commonly studied mode of teaching is demonstration, for which Argall et al. [3] wrote an excellent survey. Successes of learning by demonstration include the domains of autonomous driving [27], multi-robot coordination [8], robot soccer [13], and helicopter flight [1]. Other modes, though given less attention, have also been studied: learning from advice [24,21], learning from numeric feedback or reward [15,37,18,40], and allowing the human to shape the agent's learning environment, facilitating the learning process [32,38]. For a more thorough review of the general topic of agents learning from human teachers, we refer the reader to Knox and Stone [18].

The concept of an agent using actions to affect a human teacher, though usually left out of the conversation about such human-oriented learning agents, has been explored previously. Nicolescu and Mataric [25] speak of "communication by acting," which is using behavior to communicate intentions and needs. They specifically consider how a robot can ask for help after failure and conduct experiments in which the robot repeatedly tries to execute a failed behavior to signal a need for help. A difference between their approach and ours is that, in their work, the human's requested assistance comes after learning, so the robot improves its current performance through assistance but it does not improve its autonomous performance.

2.1.2 How humans teach

The general question of how humans teach has been studied extensively. We review some of the more relevant work here.

Some work has specifically examined how humans teach social robots or other agents. Thomaz and Cakmak [38] examined how people teach a robot affordances (i.e., action-effect relationships) to manipulate objects. Among their findings, they observed that humans point out affordances that are rare in systematic exploration of the object configuration and action spaces. They also found that people often remove the object before the robotic action completes, possibly indicating that the remaining part of the action would not have caused a desirable effect. Kim et al. observed how people talk while teaching a robot and found that "people vary their vocal input depending on the learners performance history" [17]. In work by Koachar et al. [16], human subjects teach a complex task to a fake agent in a Wizard-of-Oz experiment. The teachers could give reward-based feedback, give demonstrations, teach concepts by example, and test the agent's skills. The authors found that "teaching by [feedback] was never employed by itself and in the 82% of cases where it was used, it followed another teaching type in all but 2 cases. 58% of the teachers who used feedback used it exclusively after testing." A consistent finding across all of these studies is that human teachers break implicit and explicit expectations built into the learning system (e.g., removing an object before an action is complete), suggesting that agents should be robust to at least some such violations.

Looking particularly at teaching by explicit reward and punishment, there has been much research on how humans and other animals learn [5] and, complementarily, how people *should* teach [29,28]. However, little has been said about how people actually *do* teach by explicit reward and punishment and, complementarily, how pupils should learn from it—as this paper does. One exception is by Thomaz and Breazeal [37], who had people teach a task to a software agent by reward and punishment (the agent also had another feedback source). They found that people gave more reward than punishment and that people appeared to be using the feedback mechanism to give guidance to the agent, again interestingly breaking protocol.

2.1.3 Studying human social behavior with human-agent interaction

Here we discuss the budding practice of studying *human-human interaction* using *human-agent interaction* experiments. We do not include in this category studies that draw conclusions that are only of interest to the human-robot interaction or human-agent interaction communities. We save our discussion of the motivation for using agents in lieu of humans for Section 4.4, where we can interweave our experimental results.

Replacing humans with agents in experiments on human social behavior has been proposed by numerous researchers [6,22,9]. Of the relevant social robotics studies which we are aware, all used both human-human and human-robot interaction [14,30]. In one [23], people converse with either a human, a Wizard-of-Oz robot (i.e., a robot controlled by a human but pretending to be autonomous), or an openly remote-controlled robot. Researchers examined which direction subjects moved their eyes when breaking eye contact after having been asked a question. The results on the effect of which conversational partner was used were inconclusive, which the authors attribute to high variance and a small sample size. In another study [11], each subject watched two videos, one of a collaborative human assistant and another of a collaborative robot assistant. Afterwards, subjects rated the collaboration on multiple criteria, such as comfort with and trust in the assistant. Subjects were divided along two additional variables. Along one of these variables, subject nationality, results on collaboration ratings were consistent across human and robot versions (e.g., Chinese subjects gave higher trust ratings for both human and robot assistants than did subjects from the U.S.). The ratings were not consistent along the other variable, how strongly the subject is prompted to consider the assistant to be part of her ingroup (i.e., a group that the subject strongly identifies with).

From these studies, we see two patterns. First, the robots and humans were not perfectly interchangeable as social partners. However, the difference in their effects was usually by whether results were significant, not by significant results in opposite directions. And the results did agree a fair amount. Overall, their specific robotic partners created interactions that resembled those with humans in some situations, but not fully. We note, though, that results from studies with human actors following scripted interactions—as the human partners in the above social robotics experiments do—differ in their own way from the ground truth of authentic human-human interaction. The second pattern is that none of these experiments use agents to fully replace humans where their use would be problematic or to perform analysis that would be impossible with humans. Among previous work that employed computational agents or robots to study human interaction, our experiments stand out for random assignment and controls, for the relatively large sample sizes, and for the complexity of our agents.

2.2 Background

2.2.1 The TAMER learning agent

The experiments carried out in this paper involved human subjects training a computational learning agent that implements the TAMER framework, employing the algorithm published by Knox and Stone [18]. TAMER, explained below, has two main motivations: (1) to empower people—regardless of programming ability—to designate correct behavior, which will often be specific to the person training and (2) to speed learning compared to traditional reinforcement learning by transferring human knowledge about the task to an agent.

The TAMER framework is an answer to the Interactive Shaping Problem [18]. The Interactive Shaping Problem asks how an agent can best learn to perform a task given only real-valued feedback on its actions from a human trainer. This problem is put formally as follows.

> **The Interactive Shaping Problem** Within a sequential decision-making task, an agent receives a sequence of state descriptions ($s_1, s_2, ...$ where $s_i \in S$) and action opportunities (choosing $a_i \in A$ at each s_i). From a human trainer who observes the agent and understands a predefined performance metric, the agent also receives occasional positive and negative real-valued reward signals ($h_1, h_2, ...$) that are positively correlated with the trainer's assessment of recent state-action pairs. How can an agent learn the best possible task policy ($\pi : S \rightarrow A$), as measured by the performance metric, given the information contained in the input?

Human reward is delivered through push buttons, spoken word, or any other easy-to-learn interface.

The TAMER framework is designed around two insights. First, when a human trainer evaluates some behavior, she considers the long-term impact of that behavior, so her feedback signal contains her full judgement of the desirability of the targeted behavior. Second, a human trainer's feedback is only delayed by how long it takes to make and then communicate an evaluation. Thus, credit from human reward can be assigned within a small window of recent actions. Though it is tempting to treat human reward[1] as reward within a reinforcement learning framework, these insights suggest a different approach. In reinforcement learning, agents use reward to estimate return, the long-term accumulation of reward. These estimates of return are considered the values of actions. However, human reward is more qualitatively analogous to a trivially delayed, noisy sample of expected return from the targeted behavior given the trainer's expectations of future behavior than it is to reward in an RL framework.[2]

Consequently, a TAMER agent does not try to predict and maximize long-term human reward. Instead, it tries to predict and maximize immediate reward, converting an apparent reinforcement learning problem into a supervised learning problem (with some credit assignment techniques which are described in past work on TAMER). Put simply, a TAMER agent assumes that the trainer has an internal feedback function, $H : S \times A \rightarrow \mathbb{R}$, and treats feedback as labels on state-action pairs, providing samples to learn $\hat{H}$, an approximation of H, via supervised learning. If acting greedily, the agent chooses the action that maximizes the output of $\hat{H}$ given the current state. In practice, all TAMER agents thus far have been greedy, since the trainer can punish the agent to make it try something different, making other forms of exploration less necessary.

Our experiments indicate that humans can train TAMER agents to perform tasks well (but imperfectly) within shorter time than a traditional RL agent would learn, reducing the costs of poor performance during learning.

2.2.2 The experimental task: Tetris

In this section, we describe how our human subjects trained TAMER agents and the task-specific agent implementations. Each TAMER agent was trained to play Tetris (Figure 1) as implemented in RL-Library [35] (with some visual adaptations),[3][4] a well known, computer-based puzzle game. In Tetris, pieces of various shapes fall from the top of the screen, and the player's task is roughly to fit each piece with previous pieces below to make solid horizontal lines. Each such line disappears upon filling its last hole(s), and the pieces above move down one position. Play ends when a piece cannot be placed because previous pieces are stacked too high, and the object of Tetris in our implementation is to clear as many horizontal lines as possible before play ends.

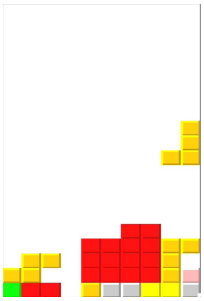

Fig. 1 A screenshot of RL-Library Tetris.

In this TAMER algorithm, the agent's action is a choice among potential piece placements, not each movement or rotation of a piece. $\hat{H}$ is represented by a linear model over 46 features that are extracted from a piece placement's effect on the Tetris board (i.e. state-action features). For the full time that the agent places the current piece, the trainer can give reward to the *previous* piece's placement; thus, credit is trivially assigned to the previous action, which is effective with slow action frequencies that give the trainer plenty of

[1] Following common practice in reinforcement learning, we use "reward" to mean both positively and negatively valued feedback.

[2] The trainer's assessment of return is, of course, dependent on her understanding of the task and expectation of future behavior, both of which may be flawed and will likely become more accurate over time.

[3] Tetris is one of five task domains for which TAMER has published results [18,33, 19,20].

[4] The specification of Tetris in RL-Library follows does not differ from that of traditional Tetris, except that there are no points or levels of increasing speed, omissions that are standard in Tetris learning literature [4]. We use RL-Library for convenience and its compatibility with RL-Glue, a software specification for reinforcement learning agents and environments.

time to respond. Each placement results in a set of features and a human reward value that together make a sample for supervised learning. The specific learning algorithm is incremental gradient descent. For more details on the TAMER agent's algorithm, consult Knox and Stone [18].

Subjects observed their agent's play on a computer screen and delivered real-time feedback, targeting the most recent block placement, through two keys. One key corresponded to positive feedback and the other to negative feedback, and each press of the button increased the feedback's intensity up to a predefined limit, yielding integer feedback values in the range $[-4,4]$. Subjects were given a short practice period to adjust to the training task before starting the actual training.

3 Experimental design and results

In this section, we discuss the designs and results of the two experiments. We first describe aspects of experimental design that were common to both experiments.[5]

We evaluated participants' teaching with descriptive analyses as well as simulations of their learned models' ($\hat{H}$s') performances. For descriptive analyses, we considered the human responses' frequency. All descriptive analyses were conducted over time in bins defined by intervals of 80 time steps. In other words, the first bin considered time steps 1 to 80, the second considered steps 81 to 160, and so on.

Simulations were performed offline for each subject at 80 time-step intervals, fixing $\hat{H}$ and using a greedy policy—and thus fixing the learned behavior—after 80, 160, ... time steps of training and then testing the fixed behavior's performance over 20 games (i.e., episodes). For our performance metrics, we use the mean number of lines cleared per episode by a TAMER agent over the 20 games in simulation at each time interval. This analysis evaluates the quality of a subject's training by simulating the performance of the fixed policies shaped from their feedback.

Subjects were drawn from the undergraduate community at the University of Texas at Austin.

3.1 The critique experiment: teaching vs. critiquing

In our first of two experiments, the *critique experiment*, we tested how donning the role of teacher affected subjects' feedback frequency and the effectiveness of their teaching.

3.1.1 Design

Subjects were randomly assigned to one of two conditions: Teaching or Critiquing.

1. Teaching (n = 27): Subjects were aware that the agent learns from his or her feedback.
2. Critiquing (n = 30): Subjects were told that they should critique a recording of an agent learning.

The authors' hypotheses about the conditions' effects on feedback frequency and agent performance varied. The dominant hypothesis was that when teaching, humans would satisfice aggressively, dramatically reducing their feedback once the agent appeared to be doing reasonably well. This reduction in feedback might harm the agent's performance compared to one that received a consistent level of feedback. If the non-teaching subjects trained better agents, it would suggest that human trainers need to be fooled into providing large amounts of feedback over time to maximize a TAMER agent's performance. Another intuition was that the Teaching subjects would be more engaged and attentive, leading to a contrasting hypothesis that the teaching group would give more feedback and achieve better performance. The plausibility of either result motivates this experiment.

3.1.2 Results

Our results focus on the question of whether frequency of feedback and agent task performance differed between the two conditions. We found that they did not differ. More detailed results are below. For our analyses, one subject in each condition was removed for not responding during the experiment, and two subjects were removed from the Critiquing group for not completing at least 720 time steps of training (as did the remaining 57 subjects).

Plots of feedback frequency and performance by condition are respectively shown in Figures 2 and 3. A 2 (condition) x 9 (interval) repeated measures ANOVA indicated no significant main effect of condition nor an interaction of interval and condition for the dependent measure of frequency of responding (all $F[2,55] < 0.83$, $p > 0.60$). Considering agent performance (i.e., lines cleared by the simulated TAMER agent), there was no significant main effect of condition nor an interaction of interval and condition (all $F[2,55] < 1.14$, $p > 0.33$).

Seeking to assess how similar the effects of the two conditions are, we calculated a Bayes factor for the data. A Bayes factor is the odds that the null hypothesis is true when the alternative is a distribution over alternative hypotheses. We examined performance at the end of the nine intervals, giving something akin to a final skill level, and feedback frequency over all intervals. Using an effect-size scaled parameter of 1 for specifying the distribution of alternate hypotheses, we calculate the JZS Bayes factor to be 4.28 for performance and 4.67 for feedback frequency [31]. Thus, under this parameter—which is recommended as a default parameter because it favors neither outcome—the null hypotheses for both metrics is more than four times more probable than

[5] Instructions given to subjects can be found at `http://www.cs.utexas.edu/~bradknox/papers/12ijsr`.

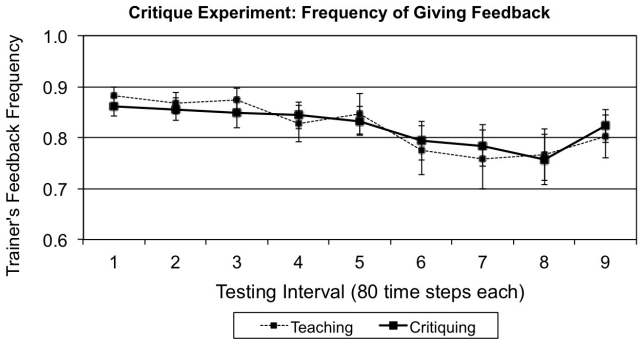

Fig. 2 Feedback frequency from the human trainer over 9 bins of 80 consecutive time steps each. On all plots with error bars, the bars show standard error.

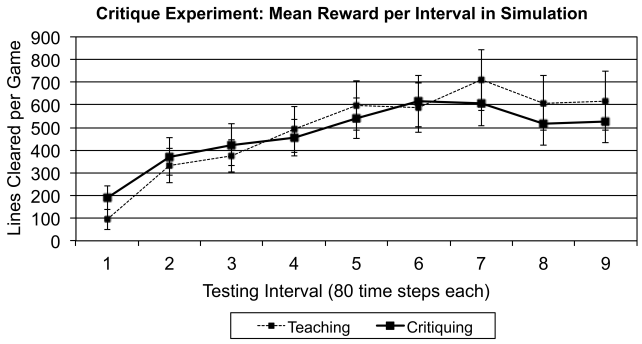

Fig. 3 Performance (lines cleared per game) of policies fixed at the end of 9 intervals, each 80 time steps in length. In other words, the tested agents have learned from the first 80, 160, ..., and 720 time steps.

the alternatives, given the data. A Bayes factor above 3 is commonly considered substantial evidence that the null hypotheses is approximately true, giving us confidence to conclude that the subjects' roles had similar effects.

Though we are not focusing on the difference in the amounts of positive and negative reward given, we report that the mean absolute value of positive reward per time step was greater than that of negative reward across all conditions of both experiments (all $p < 0.025$ in paired t-tests). This finding confirms observations by Thomaz and Breazeal [37].

In summary, the difference in participants' roles did not significantly affect any of the the dependent variables. Looking at performance, a Bayes factor analysis suggests that similarity between the two groups can explain the lack of significance, as opposed to merely too few subjects or too high of variance.

This critique experiment influenced the following experiment on feedback-frequency in several critical ways. First, because teaching and critiquing trainers behaved and performed similarly, all conditions in the feedback-frequency experiment involve a teaching role for the subject. Second, because subjects' frequency of responding was quite high in the critique experiment, we changed the subjects' instructions from "If it has made a [good/bad] move, press ..." to "If you feel it is necessary to [reward/punish] it, press ...". From this change in instructions, we hoped to both lower their baseline frequency and give subjects more leeway to determine their own frequency, two consequences that we expected to increase any differences in frequency created by the different conditions. Lastly, after the conditions of this critique experiment did not significantly affect the rate of feedback that some authors predicted would improve performance, we were motivated to more directly manipulate feedback frequency by making the agent react to it.

3.2 Feedback-frequency experiment: Varying action greediness with feedback frequency

In this section, we describe the feedback-frequency experiment, which investigates a human-agent interaction scenario in which the computer agent reacts to waning human feedback by behaving worse. By controlling the parameters of the computer agent's reaction to its human trainer's frequency of feedback, we were able to evaluate the human behavioral response under three conditions. The specification of conditions below relies on the term *greedy*, which in this context means choosing the action a that maximizes a prediction of immediate human reward, $argmax_a[\hat{H}(s,a)]$. To be concise and ease reading, we sometimes refer to non-greedy actions as "misbehavior", since agents are taking actions that they currently believe to be suboptimal (though they may actually be optimal).

3.2.1 Design

Subjects were randomly assigned to one of three conditions: Reactive Non-greedy, Greedy, or Yoked Non-greedy.

1. Greedy (n = 19): The TAMER agent always chose the action with the highest predicted feedback value.[6]
2. Reactive Non-greedy (n = 30): The TAMER agent's level of greediness was negatively correlated with the recency-weighted frequency of human feedback. (The frequency is "recency-weighted" because more recent opportunities for feedback are weighted more heavily in the frequency calculation.) For this group and the Yoked Non-greedy group, details about calculating feedback frequency and its effect on action selection are described below in this section.
3. Yoked Non-greedy (n = 30): To separate the effects of general misbehavior from misbehavior that occurs in response to the trainer, we added a third group in which agents explored without being tied to their respective trainers. In this Yoked Non-greedy group, the TAMER agent used the frequency from a matched trainer from the Reactive Non-greedy group instead of its own trainer's feedback frequency. In other words, we assigned each member of this group to a member of the Reactive Non-greedy group. The agent explored based on feedback frequency, identically to the Reactive Non-greedy group, except that the frequency at step i was determined from the feedback history of the matched subject from the

[6] The Greedy group can be considered similar to the Teaching group from the critique experiment. The two groups' instructions do contain differences, but both groups have identical TAMER agent algorithms and subjects are aware that they are teaching.

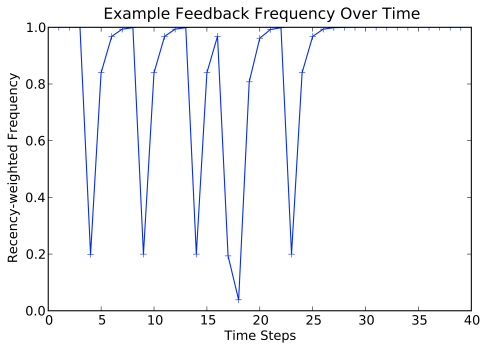

Fig. 4 An example trajectory of recency-weighted frequency over the first 40 time steps of training. The frequency varies dramatically, a consequence of the small decay factor, 0.2. In this example, the trainer refrains from giving feedback only six times.

Reactive Non-greedy group rather than the current subject's feedback history. Thus, whereas the agent acted with varying degrees of misbehavior, the level of misbehavior was not causally determined by the subject's behavior.

We hypothesized that the Reactive Non-greedy group would have the highest feedback frequency as well as the best performance. Our intuition was that, in line with the idiom "The squeaky wheel gets the grease" and popular wisdom that misbehavior can be a cry for attention, an agent that "misbehaves" when feedback frequency diminishes will be effectively training the trainer to give more feedback. And given more feedback, the agent would have more training samples to learn from, resulting in better task performance.

Calculating frequency To calculate a trainer's recency-weighted feedback frequency, each feedback instance is exponentially decayed over time. Thus, at each time step, we calculate $a := [decay * a] + (feedback \neq 0)$ and $b := [decay * b] + 1$, where a and b are initialized to zero and $feedback \neq 0$ resolves to 1 when feedback was given and 0 otherwise. Together, a and b define frequency: $freq := a/b$. In our experiments, the decay parameter was 0.2, which heavily weights the last few actions. An example frequency trajectory can be seen in Figure 4.

Choosing actions based on frequency Given a frequency, the agents in both non-greedy conditions choose actions. To choose, an agent ranks all available actions according to their predicted human reinforcement, $\hat{H}(s,a)$, and picks out five actions from that ranking: the best, the second-best, the action at the first quartile, the action at the median, and the worst. (Ambiguous quartile and median choices go to the better-ranked action.) Then, the agent chooses randomly from these five actions according to a probability distribution conditioned on frequency, where lower frequencies generally result in worse-ranked action choices. The distributions can be seen in Figure 5.

3.2.2 Results

We performed the same descriptive and model-based analyses as we did for the previous critique experiment.[7] An exception though, is that we find significant results here and thus do not perform the Bayes factor calculation, which we used to determine how similar the data was between conditions after finding a complete lack of significance. One Greedy subject, three Reactive Non-greedy subjects, and five Yoked Non-greedy subjects were removed for responding insignificantly during the experiment. Also, one Greedy subject, one Reactive Non-greedy subject, and four Yoked Non-greedy subjects were removed for training for less than the 800 time steps we used for analysis. For the two non-greedy conditions, subjects matched to removed subjects were also removed from analysis.

If non-greedy actions increase feedback frequency and tying non-greedy actions to trainer's recent feedback frequency further increases subsequent frequency, we expect the Reactive Non-greedy group to have the highest frequency, followed by the Yoked Non-greedy group, with the Greedy group having the lowest frequency. And since frequency increases the number of learning samples, we expect the same ordering of performance.

Trainer's feedback frequencies are shown in Figure 6, and performance after each training interval is shown in Figure 7. Note that the change in instructions described at the end of Section 3.1.2 was effective: the baseline feedback frequency, given by the Greedy group, is lower than the almost equivalent Teaching group in the critique experiment.

Surprisingly, 2 (condition) x 10 (interval) ANOVAs comparing the performance (i.e., lines cleared) of the Greedy group over all intervals to that of the Reactive Non-greedy and Yoked Non-greedy groups found significant effects by condition ($p = 0.015$ and $p = 0.024$, respectively), indicating superior learned performance within the greedy group.

[7] Performance is again tested offline, not during training, and the testing policy is greedy regardless of condition.

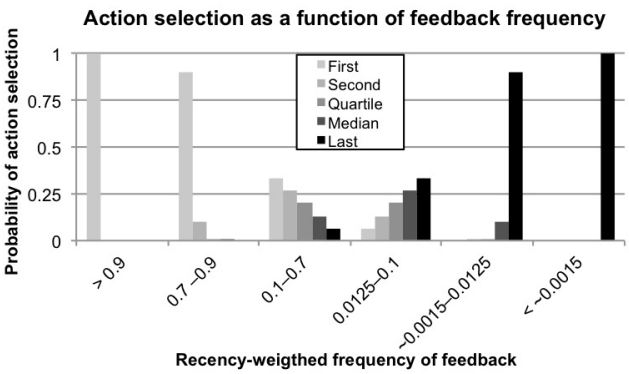

Fig. 5 Probability distributions over the five possible actions at different recency-weighted feedback frequencies.

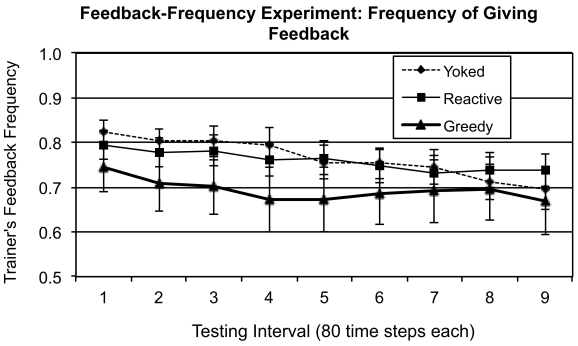

Fig. 6 Feedback frequency from the human trainer over 9 bins of 80 consecutive time steps each.

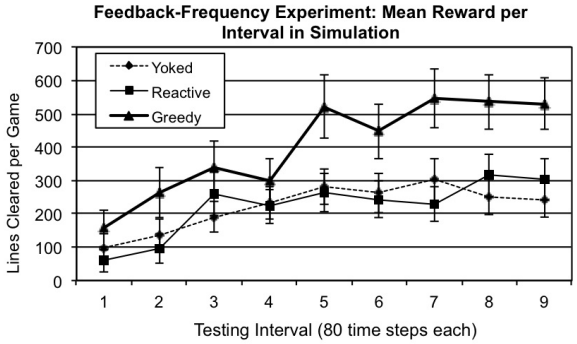

Fig. 7 Performance (lines cleared per game) of policies fixed at the end of 9 intervals, each 80 time steps in length.

The two non-greedy groups were not significantly different. Also, 2 x 10 ANOVAs comparing trainer's feedback frequencies found no significant differences.

Results for good trainers only Before acting intelligently, these learning agents go through a period of initial learning, during which their actions are generally of low quality. Additionally, many agents are never trained to a level at which greedy actions are generally good. Taking non-greedy actions when greedy actions themselves are not good lacks the qualitative characteristic on which we are focused: non-greedy action corresponding to decreased quality of action. Therefore, we repeat the analyses above, only examining the subset of subjects who were able to train their agents to consistently clear more than 10 lines on average across multiple time intervals. Additionally, we only use data starting at the third interval, where the percentage of agents that pass the 10-line standard first surpasses 90% (after pass rates of only 58.3% and 72.2% in the first two intervals), never dropping below after. The 10-line threshold was chosen for its position in the valley of the bimodal distribution of agent performance across subjects.[8] This more selective analysis gives a different perspective that is more focused on the effect of "misbehaving" to affect the trainer.

[8] Illustrating the bimodality of performance, there were 79 subjects across conditions. In the 9th testing interval, 23 agents clear between 0–1 lines; 47 clear more than 100. Only 2 agents clear 5–20 lines.

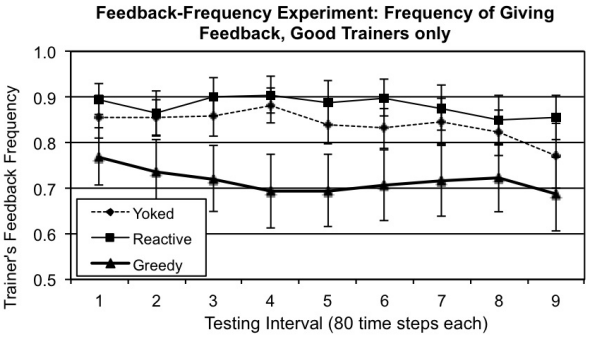

Fig. 8 Figure 6 with low-performing trainers removed.

After removing low-performing subjects and all subjects that were matched to those low-performing subjects, the condition sizes were Reactive Non-greedy, n = 10; Greedy, n = 16; and Yoked Non-greedy, n = 10. The feedback frequency by condition across time intervals for this smaller set of subjects is shown in Figure 8. Compared to the full set of subjects, all conditions have generally higher feedback frequencies. However, this frequency increase is more pronounced in the two non-greedy conditions. Further, the Reactive Non-greedy condition now results in more frequent feedback than the Yoked Non-greedy condition. Despite the small number of subjects being considered, the Reactive Non-greedy group's mean feedback frequency over intervals 3-10 has marginal significance ($p < 0.1$) in comparison to the lower feedback frequency of the Greedy group. Additionally, a non-parametric analysis of the same data is nearly significant, with an upper confidence bound of 0.83 for the Greedy group and a lower confidence bound of 0.82 for the Reactive Non-greedy group. Therefore, we expect that increasing the number of subjects would quickly strengthen the significance of the difference in feedback frequency.

The performance of these subjects across conditions is much more similar than in the full set of subjects and is no longer significantly different. However, removing subjects based on performance clearly biases group performance. This bias is motivated for analyzing feedback frequency but not for performance, and we therefore only base our performance results on the full set of subjects. We can say, though, that these results add further evidence (though biased) that increased feedback frequency and the consequently increased number of learning samples do not result in better performance in this experiment.

4 Discussion

In this section, we interpret and discuss the results of the experiments in Section 3. The first subsection draws a conclusion on agent design from the critique experiment; the second subsection does likewise from the feedback-frequency experiment. We then discuss the feedback-frequency exper-

iment's implications for the explore-exploit dichotomy that is pervasive within the field of learning agents. Lastly, we discuss the new technique of using social agents to study human behavior, using our experiments as examples and describing how these results may also be of interest outside of artificial intelligence communities.

4.1 Honesty is still the best policy

When agents learn to perform tasks, one clear objective is to maximize performance. The results from the critique experiment indicate that, contrary to the hypothesis that human trainers would need to be deceptively told that they are not teaching to do their best training, the human-agent system performs similarly when the human knows that he is engaged in a training session. Either the subject's role had little or no effect on his feedback, or the TAMER agent was robust to differences in feedback.

In addition to the performance objective that we explicitly study, it is also important to respect the desires and needs of humans. Deceiving human trainers to get the best performance is an ethically questionable trade-off. The results provide evidence that disclosing to the trainer that he is teaching maximizes both crucial objectives, performance-based and humanistic.

4.2 A tool for increasing feedback

When numeric feedback comes to an agent from an encoded reward function instead of a human, the problem is often framed as a reinforcement learning problem. These problems are usually formalized as Markov Decision Processes (MDPs). In an MDP, reward has a static distribution of frequency and quality. In contrast, human reward can be affected along both of these dimensions. From this observation, one may notice that one way to give highly effective feedback (though possibly imperfect feedback with certain function approximators) for a TAMER agent would be to give feedback at every time step and have as its value the expected return of MDP reward under the optimal policy from the most recent state-action pair, where the MDP reward follows the task's objective and credit is assigned only to the preceding time step. These two characteristics of feedback—frequency and quality, or, equivalently, the number of learning samples and the quality of their labeling—comprise two dimensions along which a particular human trainer's feedback can be improved.

The feedback-frequency experiment demonstrates one on-line technique for increasing the frequency of human feedback: lowering action quality. More specifically, when examining only the successful trainers (for which non-greedy actions would actually look worse), tying the action quality to the trainer's recency-weighted frequency further increased feedback frequency. Considering that there are likely other techniques that increase either frequency or quality of feedback, one product of our results is a proof-of-concept that this broader category of techniques exists, though the extent of its breadth is an open question. Also, the concept of an agent manipulating the trainer to improve feedback can be generalized to other modes of teaching, including demonstrations, which can also vary by frequency and quality.

Contrary to our expectations, though the agents' manipulations increased feedback frequency, they did not improve performance and even decreased it among the full set of subjects. Exploring created more learning samples, but we suspect these samples were less useful than those experienced by the Greedy group. We see two plausible explanations: the learning samples were in a less useful area of the state-action space, or the quality of trainer feedback worsened. The intuition behind the first potential explanation is that the learning samples created during greedy behavior help distinguish between the best few actions, whereas non-greedy behavior created samples that help distinguish between mediocre or worse actions, a type of differentiation that does not aid an agent trying to choose the best action. Further, the samples from non-greedy actions may have even been directly harmful; the representation of $\hat{H}$ is not highly expressive, and more accurately modeling reward for non-greedy actions likely lessens the accuracy of modeling high-quality actions. The other potential explanation is that the quality of the feedback within the non-greedy conditions suffered because of trainer frustration or some other effect of misbehavior on the trainer. Further analysis of the data might shed light on which of these explanations is correct. For instance, we could test each agent's performance with the same learning samples, except we label each sample with a static feedback function instead of with the variable set of humans that did label the samples. This relabeling would control for quality of feedback, directly testing how much the difference in the samples' locations in state-action space would affect performance. More generally, whether misbehavior can be used to increase interaction *and* learned performance is a promising question for future inquiry.

The agent's frequency-tied action selection can least speculatively be framed as a "manipulation". We might also consider it to be a form of communication with the human trainer, though we are careful not to imply that the trainer consciously understood some message from the frequency-tied actions, which she may or may not have. Another speculative but plausible interpretation is that when the agent lowers its action quality after the trainer's feedback drops in frequency, the human is being punished for inattentiveness. This interpretation is more compelling if the human trainer is emotionally vested in the agent's performance, which fits anec-

dotally with comments made by subjects and the authors' experience in informally training agents themselves.

One lesson of this feedback-frequency experiment is that agent designers should be careful not to make the mistake of considering pedagogy to be a single-directional manipulation, that though teacher and student do interact, it is the student who is significantly changed through the interaction. On the contrary, the student has expectations of the teacher and beliefs about how the teacher should meet his or her needs, and an effective student will teach the teacher how to meet those needs.

4.3 Non-greedy action is not necessarily exploration

When referring to agents that learn to estimate some notion of the relative values of various state-action pairs (i.e., not policy-search learners), researchers generally consider actions to be either exploratory or exploitative. This dichotomy between exploration and exploitation holds strictly in traditional reinforcement learning, where an action a is exploitative if it is chosen greedily, such that $(a = argmax_a Q(s,a))$, and contrapositively any action chosen non-greedily, typically resulting in $a \neq argmax_a Q(s,a)$, is exploratory [34].[9]

At the intersection of learning agents and human-agent interaction are agents that, like TAMER agents, learn interactively from human teachers. In past work, many of these agents only exploit [18,2,26] and some, especially those that use reinforcement learning, explore or exploit [37,15, 40]. However, we will argue that the non-greedy actions taken by agents in the Reactive Non-greedy group of the feedback-frequency experiment are neither exploration nor exploitation.

4.3.1 Is it exploitation?

Retaining the notion that exploitation involves greedy action selection, the Reactive Non-greedy group's non-greedy behavior was not exploitation by definition. This conclusion generalizes to any agents that learn the values of state-action pairs for the task and cannot model the human as part of their value function, though they may be able to model the impact of their actions on the trainer's feedback frequency and quality.

4.3.2 Is it exploration?

In the terminology of reinforcement learning, any action a such that $a \neq argmax_a \hat{H}(s,a)$ is commonly referred as "exploration". But exploration in reinforcement learning, and in general if we want to keep the term close to its colloquial meaning, is undertaken to learn more about state-action pairs which are not experienced sufficiently during greedy behavior to create the desired level of behavioral improvement. The Reactive Non-greedy group in the feedback-frequency experiment may have received a wider range of state-action pairs in their learning samples as a result of their non-greedy behavior, but they also affected their feedback source. Their trainers' feedback frequency, on average, was higher than that of other groups, sometimes significantly so, giving the agents motivation beyond exploration to act non-greedily.

Through its non-greedy actions, an agent in the Reactive Non-greedy group does receive information about state-action pairs that it would likely not encounter during greedy actions. So, in a sense, exploration does occur. But exploration is not the only effect, and in an agent that predicts the effects of its actions and acts with goals, the exploration may be merely incidental to the intended result of increasing feedback frequency. Thus, while the agents' non-greedy actions had exploratory consequences, calling such actions exploration is incomplete, obscuring their desirable, non-exploratory effects.

4.3.3 Non-greedy action to manipulate the trainer

There is more than one reason to act non-greedily. Exploring is one reason, as is increasing a trainer's feedback frequency. If the learning agents community intends to embrace the use of humans as teachers for agents, it might reconsider the common practice of using the word "exploration" synonymously with non-greedy actions. Though exploration remains a critical form of non-greedy action, our results show that when a human trainer is in the learning loop, there are reasons to act non-greedily besides exploration.

4.4 Illustration of employing human-agent interaction to study human behavior

In this subsection, we conduct a more general discussion on the merits of using social robots or social software agents to study human behavior outside of human-agent interaction. Our experiments serve as motivating examples in this discussion.

Computational agents, both robotic and simulated, comprise an emerging tool for the behavioral sciences. In current practice for experiments on human behavior that require social interaction and constrained behavior on one side of the interaction, a human fulfills the role opposite the subject. Compared to this human actor,[10] a computational agent can

[9] Though exploration is often considered equivalent to non-greedy action, this definition does not fit all instances of its use in RL. For instance, an agent that employs an exploratory policy might have a greedy policy that sometimes agrees on what action to select. However, this is a semantic point that does not affect our assertion that the comprehensive dichotomy of explore/exploit is insufficient.

[10] A human opposite the subject could have fully scripted behavior, act naturally except in certain situations (like misbehaving at certain times), or simply act naturally. Additionally, the subject may believe either that this person is a fellow subject or that she is working for the experimenters. We call this human that would potentially

act more consistently, since its behavior is fully parametrized. Further, the conditions under which humans act may confound their performance. In our feedback-frequency experiment, for example, a human pupil's learning would likely be confounded by varying levels of mental effort to align actions to the constraints of each condition. The computational agent chooses its actions without meaningfully pulling resources from the learning algorithm (i.e., though they share computation time, there was plenty of time for both). Additionally, the computational agent can record every aspect of its "mental" process and behavior, allowing in-depth analysis later. Both experiments provide an example of such analysis, freezing learning at different points in time and testing performance. On the other hand, human actors have some clear advantages. The focus of studies on social interaction is generally human-human interaction, and human subjects probably interact more naturally with human actors than computational ones, though the extent of this difference will depend on the character of the computational agent. Thus, the relative generalizability of results from experiments with human actors increases from the authenticity of human-human interaction. Given the different strengths of human and computational agents, we expect both to play an important role in future behavioral studies, a view shared by some in the human-robot interaction community [6, 22, 9].

This paper provides analysis aiming to be valuable to a researcher of learning agents or human-robot interaction. However, these results may also be of interest to the educational community. There the relationship between classroom misbehavior and teacher attention is of real-world importance [39]. In a relatively recent article, Dobbs et. al [10], summarizing past research on the relationship between misbehavior and attention from teachers, write that "children who misbehave frequently receive more teacher attention than do children who rarely misbehave." One study found that the amounts of criticism and commands received from a teacher were negatively correlated with the level of on-task behavior from children [12]. Other research on this relationship has been correlational and often considers a potential causal relationship in the direction of attention causing misbehavior. Using real children as misbehaving confederates in a randomized controlled trial is an untenable proposition. But with interactive agents, we were able to establish the first causal connection between misbehavior and teacher attention, showing that performance-oriented misbehavior can increase attention.

5 Conclusion

This paper describes two experiments that consider how human beliefs and agent behavior affect a human's teaching. The first, the critique experiment, showed similar feedback frequency and agent performance between subjects placed in a teaching role and subjects in a critiquing role, indicating that either the role had little effect on the subject or it did affect the subjects' feedback quality but the resultant differences did not affect the TAMER agent's performance. The second, the feedback-frequency experiment, demonstrated a technique that agents can use to increase the frequency of trainer feedback: acting non-greedily. Additionally, when we filter for agents that show sustained decent or better performance, the frequency increase is greatest when this non-greedy misbehavior occurs in response to decreases in the trainer's feedback rate. Through this type of behavior, the feedback-frequency experiment also gives a specific example of how actions in the presence of a human trainer can be used for purposes other than exploration or exploitation. This result shows that the explore/exploit dichotomy is inadequate for describing actions by an agent learning interactively from a human. Together, these experiments 1) lend support to the efficacy of the TAMER approach—actively taught and thus far greedy—to learning from human reward and punishment, and 2) identify forms of human-agent interactivity that do or do not impact agent performance.

This research may serve as a model for other research that studies humans by having them interact with robots. The generality of our findings would be buttressed by repeating these two experiments in different contexts: especially using a robotic agent, different tasks, and even a different teaching modality, such as Learning from Demonstration. Nonetheless, the results presented here provide interesting, sometimes surprising results that apply to designers of learning agents, including social robots. And the unexpectedness of some of our conclusions indicates that further studies of human teaching stand to provide much counterintuitive guidance in the design of agents that learn from human teachers.

An agent with the power to manipulate the trainer to its advantage should not necessarily use that power. We should consider when pulling a teacher in for more training is worth the cost in human effort. There are numerous potential approaches to this problem. For example, a more sophisticated agent might have some self-confidence measure and only engage the human when it lacks confidence in making decisions [7].

Lastly, this paper's two experiments serve as exemplars of using agents as parametrized social entities in experiments on human behavior. We hope that they will inspire and guide researchers to explore this nascent experimental technique, helping to expand the impact of human-agent and human-robot interaction into the behavioral sciences.

Acknowledgements This research was supported in part by NIH (R01 MH077708 to WTM), NSF (IIS-0917122), AFOSR (FA9550-10-1-0268), ONR (N00014-09-1-0658), and the FHWA (DTFH61-07-H-

be replaced by an agent a "human actor" for simplicity and to differentiate from the subject.

00030). We thank the research assistants of MaddoxLab for their crucial help gathering data.

References

1. Abbeel, P., Ng, A.: Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the twenty-first international conference on Machine learning, p. 1. ACM (2004)
2. Argall, B., Browning, B., Veloso, M.: Learning by demonstration with critique from a human teacher. In: Proceedings of the ACM/IEEE international conference on Human-robot interaction, pp. 57–64. ACM (2007)
3. Argall, B., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. Robotics and Autonomous Systems **57**(5), 469–483 (2009)
4. Bertsekas, D., Tsitsiklis, J.: Neuro-Dynamic Programming. Athena Scientific (1996)
5. Bouton, M.: Learning and Behavior: A Contemporary Synthesis. Sinauer Associates (2007)
6. Breazeal, C.: Designing sociable robots. The MIT Press (2004)
7. Chernova, S., Veloso, M.: Interactive policy learning through confidence-based autonomy. Journal of Artificial Intelligence Research **34**(1), 1–25 (2009)
8. Chernova, S., Veloso, M.: Teaching collaborative multi-robot tasks through demonstration. In: Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on, pp. 385–390. IEEE (2009)
9. Dautenhahn, K.: Methodology and themes of human-robot interaction: a growing research field. International Journal of Advanced Robotic Systems **4**(1), 103–108 (2007)
10. Dobbs, J., Arnold, D., Doctoroff, G.: Attention in the preschool classroom: The relationships among child gender, child misbehavior, and types of teacher attention. Early Child Development and Care **174**(3), 281–295 (2004)
11. Evers, V., Maldonado, H., Brodecki, T., Hinds, P.: Relational vs. group self-construal: untangling the role of national culture in hri. In: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, pp. 255–262. ACM (2008)
12. Fagot, B.: Influence of teacher behavior in the preschool. Developmental Psychology **9**(2), 198 (1973)
13. Grollman, D., Jenkins, O.: Dogged learning for robots. In: Robotics and Automation, 2007 IEEE International Conference on, pp. 2483–2488. IEEE (2007)
14. Hinds, P., Roberts, T., Jones, H.: Whose job is it anyway? a study of human-robot interaction in a collaborative task. Human-Computer Interaction **19**(1), 151–181 (2004)
15. Isbell, C., Kearns, M., Singh, S., Shelton, C., Stone, P., Kormann, D.: Cobot in LambdaMOO: An Adaptive Social Statistics Agent. AAMAS (2006)
16. Kaochar, T., Peralta, R., Morrison, C., Fasel, I., Walsh, T., Cohen, P.: Towards understanding how humans teach robots. User Modeling, Adaption and Personalization pp. 347–352 (2011)
17. Kim, E., Leyzberg, D., Tsui, K., Scassellati, B.: How people talk when teaching a robot. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, pp. 23–30. ACM (2009)
18. Knox, W., Stone, P.: Interactively shaping agents via human reinforcement: The TAMER framework. The 5th International Conference on Knowledge Capture (2009)
19. Knox, W.B., Breazeal, C., Stone, P.: Learning from feedback on actions past and intended. In: In Proceedings of 7th ACM/IEEE International Conference on Human-Robot Interaction, Late-Breaking Reports Session (HRI 2012) (2012)
20. Knox, W.B., Stone, P.: Reinforcement learning with human and MDP reward. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2012)
21. Kuhlmann, G., Stone, P., Mooney, R., Shavlik, J.: Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer. In: The AAAI-2004 Workshop on Supervisory Control of Learning and Adaptive Systems (2004)
22. MacDorman, K., Ishiguro, H.: The uncanny advantage of using androids in cognitive and social science research. Interaction Studies **7**(3), 297–337 (2006)
23. MacDorman, K., Minato, T., Shimada, M., Itakura, S., Cowley, S., Ishiguro, H.: Assessing human likeness by eye contact in an android testbed. In: Proceedings of the XXVII annual meeting of the cognitive science society, pp. 21–23 (2005)
24. Maclin, R., Shavlik, J.: Creating advice-taking reinforcement learners. Machine Learning **22**(1), 251–281 (1996)
25. Nicolescu, M., Mataric, M.: Learning and interacting in human-robot domains. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on **31**(5), 419–430 (2002)
26. Nicolescu, M., Mataric, M.: Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In: AAMAS, pp. 241–248. ACM (2003)
27. Pomerleau, D.: ALVINN: An autonomous land vehicle in a neural network. In: Advances in Neural Information Processing Systems 1. Morgan Kaufmann (1989)
28. Pryor, K.: Don't shoot the dog!: the new art of teaching and training. Interpet Publishing (2002)
29. Ramirez, K.: Animal training : successful animal management through positive reinforcement. Chicago, IL : Shedd Aquarium (1999)
30. Reed, K., Patton, J., Peshkin, M.: Replicating human-human physical interaction. In: IEEE International Conference on Robotics and Automation (2007)
31. Rouder, J., Speckman, P., Sun, D., Morey, R., Iverson, G.: Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review **16**(2), 225–237 (2009)
32. Saunders, J., Nehaniv, C., Dautenhahn, K.: Teaching robots by moulding behavior and scaffolding the environment. In: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, pp. 118–125. ACM (2006)
33. Sridharan, M.: Augmented reinforcement learning for interaction with non-expert humans in agent domains. In: Proceedings of IEEE International Conference on Machine Learning Applications (2011)
34. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press (1998)
35. Tanner, B., White, A.: RL-Glue: Language-independent software for reinforcement-learning experiments. JMLR **10** (2009)
36. Thomaz, A.: Socially guided machine learning. Ph.D. thesis, Citeseer (2006)
37. Thomaz, A., Breazeal, C.: Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance. AAAI (2006)
38. Thomaz, A., Cakmak, M.: Learning about objects with human teachers. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, pp. 15–22. ACM (2009)
39. Wolfgang, C.: Solving discipline and classroom management problems: methods and models for today's teachers. Wiley (2004)
40. Woodward, M., Wood, R.: Using bayesian inference to learn high-level tasks from a human teacher. In: International Conference on Artificial Intelligence and Pattern Recognition (AIPR-09) (2009)

W. Bradley Knox is a Ph.D. candidate in the Department of Computer Science at the University of Texas at Austin. He was an NSF Graduate Research Fellow from 2008–2011 and won the Pragnesh Jay Modi Best Student Paper Award at AAMAS in 2010. He is interested in machine learning, robotics, and psychology, especially machine learning algorithms that learn through human interaction.

Brian D. Glass received his Ph.D. in Cognitive Psychology from the University of Texas at Austin in 2012. He is currently a postdoctoral researcher at Queen Mary, University of London. His interests include modelling human decision making and learning in complex and interactive environments such as video gaming.

Bradley C. Love is a Professor in the Department of Cognitive, Perceptual and Brain Sciences at UCL. In 1999, he received a Ph.D. in Cognitive Psychology from Northwestern University. His research centers on basic issues in cognition, such as learning, memory, attention, and decision making, using methods that are informed by behavior, brain, and computation.

W. Todd Maddox received his Ph.D. in Computational Cognitive Psychology from the University of California, Santa Barbara in 1991 and is currently the Wayne Holtzman Chair and Professor of Psychology at the University of Texas at Austin. His research focuses on the computational cognitive neuroscience of normal cognition and cognition in various clinical populations.

Peter Stone received his Ph.D. in Computer Science in 1998 from Carnegie Mellon University. He is an Associate Professor in the Department of Computer Sciences at the University of Texas at Austin and the recipient of the IJCAI 2007 Computers and Thought award. Peter's research interests include machine learning, multiagent systems, robotics, and e-commerce.