Understanding Human Teaching Modalities in Reinforcement Learning Environments: A Preliminary Report

W. Bradley Knox
University of Texas at Austin
bradknox@cs.utexas.edu

Matthew E. Taylor
Lafayette College
taylorm@cs.lafayette.edu

Peter Stone
University of Texas at Austin
pstone@cs.utexas.edu

Abstract

While traditional agent-based learning techniques have enjoyed considerable success, in recent years there has been a growing interest in improving such learning by leveraging humans as teachers. These human-in-the-loop methods have demonstrated substantial improvements by using human subjects in a variety of interaction modalities. Unfortunately, there are few, if any, guidelines about when one teaching modality is more appropriate than another. In addition to highlighting this important gap in the current literature, this paper presents a pilot study that compares two specific teaching modalities: learning by feedback and learning by demonstration, and proposes a set of hypotheses about their relative performance.

1 Introduction

There has been considerable success in allowing agent-based machine learning techniques to autonomously learn difficult tasks. In particular, *reinforcement learning* (RL) approaches have enjoyed multiple past successes. However, RL algorithms frequently need substantial amounts of data to learn a decent control policy. In many domains, collecting such data may be slow, costly, or infeasible. One promising approach towards solving RL problems in a more sample-efficient manner is to explicitly leverage human knowledge.

A number of recent papers have shown that human-in-the-loop techniques can be successfully leveraged to tackle difficult tasks [Isbell *et al.*, 2006; Thomaz and Breazeal, 2006; Knox and Stone, 2009; Judah *et al.*, 2010]. However, the growing number of methods appearing in the literature place different burdens on, and make different assumptions about, the human trainer. Due in part to these differences, there are currently few guidelines about what types of human training are appropriate under different situations.

This paper presents a pilot study that compares two teaching methods: learning from feedback and learning from demonstration, each of which are described in the following section. The primary contributions of this paper are as follows. First, we present initial results in which learning from demonstration outperforms learning from feedback. Second, we investigate secondary effects, including how the results are highly sensitive to the experimental design, and how the agent's online performance during teaching compares with the agent's offline performance. Third, we discuss important future directions to further investigate the relative merits of different learning methods.

2 Background

This section provides background on (1) reinforcement learning, the common framework for all tasks learned in this paper, (2) learning from feedback via TAMER, which allows a human to critique an autonomous learner, and (3) learning from demonstration, a method for learning to mimic a human.

2.1 Reinforcement Learning

Reinforcement learning (RL) is a flexible approach that allows agents to learn from experience. This section briefly introduces RL using the standard notation of Markov decision processes. At each *time step* the agent observes its state $s \in S$ as a vector of k state variables, where $s = \langle x_1, x_2, \ldots, x_k \rangle$. The agent then selects an action a from the set of available actions A. An MDP's reward function $R : S \times A \mapsto \mathbb{R}$ and its (stochastic) transition function $T : S \times A \mapsto S$ fully describe the system's dynamics. The agent attempts to maximize the long-term reward, which is determined by the (initially unknown) reward and transition functions. An agent chooses which action to take in a given state via a policy, $\pi : S \mapsto A$. π is modified by the learner over time to improve performance, i.e., maximize the expected total reward.

2.2 Learning from Feedback (LfF)

We define the *Learning from Feedback* (LfF) problem as a task where an agent attempts to maximize rewards from a human trainer.[1] In an LfF scenario, a human trainer observes an agent and reinforces its behavior through push-buttons, spoken word ("yes" or "no"), or any other signal that can be converted to a real-valued signal of approval or disapproval. The key challenge, then, is to create agents that can be effectively guided by such feedback.

An agent acting in an MDP 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). A human trainer provides occasional positive and negative scalar reinforcement signals ($h_1, h_2, ...$) that are correlated with the trainer's assessment of recent state-action pairs. The agent's goal is to learn a policy ($\pi : S \mapsto A$) that maximizes the human's expected long term reinforcement. In this case, the MDP's reward function is not used—following Abbeel and Ng's terminology [2004], we call this an MDP\R.

[1] In previous work [Knox and Stone, 2009], this problem was termed *Interactive Shaping Problem*, but we use LfF here to emphasize the main difference between learning from feedback and learning from demonstration.

TAMER Framework

The TAMER framework was introduced by Knox and Stone [2009] as one approach for an agent to learn from numeric reinforcement signals. These feedback signals are generated by a human trainer who observes the agent attempting to perform a task. TAMER is motivated by two insights about human reinforcement. First, reinforcement is only slightly delayed; the trainer can quickly assess the agent's behavior and deliver feedback. Second, the trainer observes the agent's behavior with a model of that behavior's long-term effects. Thus the reinforcement can be assumed to be fully informative with respect to the quality of the agent's recent behavior. Comparing LfF to RL, human reinforcement is more similar to an action value (sometimes called a Q-value), albeit a noisy and trivially delayed one, than it is to an MDP reward. Consequently, TAMER assumes human reinforcement fully encompasses the quality of a state-action pair and it uses regression to model a hypothetical human reinforcement function, $H : S \times A \mapsto \mathbb{R}$, as $\hat{H}$ in real time. In this pilot study, the regression algorithm for modeling H is k-nearest neighbors. In the simplest form of credit assignment, each reinforcement creates a label for the last state-action pair.[2] The output of the resultant $\hat{H}$ function—changing as the agent gains experience—determines the relative quality of potential actions, so that the exploitative action is $a = argmax_a[\hat{H}(s, a)]$.

2.3 Learning from Demonstration

Learning from demonstration (LfD) research explores techniques for learning a policy from examples, or demonstrations, provided by a human teacher. LfD can be seen as a subset of supervised learning, where the agent is presented with labeled training samples and must model the function which produced the data.

Similarly to RL and LfF, LfD can be defined in terms of the agent's observed state $s \in S$ and executable actions $a \in A$. Demonstrations are recorded as sequences of state-action pairs $\{(s_0, a_0), ..., (s_t, a_t)\}$, and these sequences typically only cover a small subset of all possible states in a domain. The agent's goal is to generalize from the demonstrations and learn a policy $\pi : S \mapsto A$, where π covers all states, that imitates the demonstrated behavior.

Many different algorithms for using demonstration data to learn π have been proposed. Approaches vary by how demonstrations are performed (e.g., teleoperation, teacher following, kinesthetic teaching, or external observation), the type of policy learning method used (e.g., regression, classification, or planning), and assumptions about the degree of demonstration noise and teacher interactivity (see [Argall *et al.*, 2009]). Across these differences, LfD techniques possess a number of key strengths. Most significantly, demonstration leverages the human teacher's task knowledge to significantly speed up learning by either eliminating exploration entirely [Grollman and Jenkins, 2007; Nicolescu *et al.*, 2008] or by focusing learning on the most relevant areas of the state space [Smart and Kaelbling, 2002].

Implemented LfD algorithm

For our pilot study, the subject fully controls the agent for a predefined number of teaching episodes, producing numerous state-action samples. We model π by k-nearest neighbor on those samples, similarly to our LfF implementation. As with TAMER, no MDP reward is received by the agent.

3 Expected Relative Strengths: LfF vs. LfD

This section briefly explores what the authors *expected* to be the relative strengths and weaknesses of LfF and LfD. Note that not all of these expectations are fully met in our pilot study; future work will aim to collect more data and conclusively show which of these expectations are (in)correct.

We expect LfD, where the teacher directly provides actions, to generally result in better learning than LfF, where the teacher can only give feedback on the one action chosen. Also, because the teacher controls the agent during demonstrations, whereas LfF lets the initially uneducated agent control, we expect that the first few teaching episodes of demonstration will be performed more effectively than those of LfF, reducing any cost associated with poor performance. Lastly, many LfD control interfaces will be familiar to subjects because of video games and other push-button controllers.

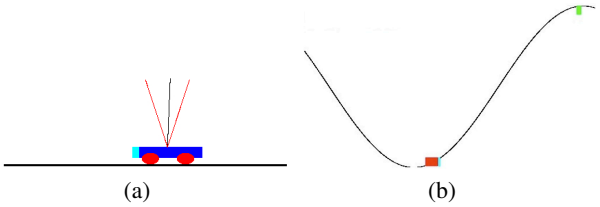
However, LfF provides some advantages over LfD. First, the interface for providing positive and negative feedback can remain constant across tasks, whereas a control interface for LfD is necessarily task-specific. Further, the actions available in a task may be too complex for teachers to control (e.g., a many-dimensional robot arm), making LfD infeasible. Second, the critical feedback of LfF may require less task expertise than demonstration; intuitively, one can evaluate the overall benefit of an action's consequences without knowing what action is optimal. Third, feedback likely places a lower cognitive load on teachers than demonstrating, using less of a human's valuable resources. Fourth, agents learning from LfF exhibit their learned policy during teaching, which informs the teacher of the learner's current strengths and weaknesses. In contrast, whereas a few LfD systems intersperse demonstration and exhibition of the learned policy [Argall *et al.*, 2007], the majority of systems do not show the learned policy during teaching, which may lead to poorly targeted demonstrations.

These relative strengths and weaknesses support the more general hypothesis that the relative performance of the two teaching modes strongly depends on the situations under which they are used. Factors such as the teacher's interface mastery, task expertise, and available cognitive resources will be critical. Additionally, the cost of poor performance in early episodes will also be an important factor. And the expressiveness of the employed model's representation (of $\hat{H}$ or π), may play an important role: poor expressiveness may increase the discrepancy between behavior during teaching and learned behavior in LfD.

4 Pilot Study

We designed a pilot study within the TAMER framework to compare the efficacy of LfF and LfD in two simple MDPs

[2] The trivial delay is dealt with using a credit assignment technique similar to that described in Knox and Stone [2009]. We plan to describe our current credit assignment technique exactly in a future journal article.



(a) (b)

Figure 1: Screenshots from (a) Cart Pole and (b) Mountain Car

using teachers who were not familiar with RL or MDPs. The study used 16 undergraduates who were taking a first semester computer science course, including both majors and non-majors.[3] Students trained agents in the *Cart Pole* and *Mountain Car* tasks (explained in the following subsection).

All students worked with Cart Pole first and Mountain Car second. Students were selected randomly to perform first one of the two training modalities—LfF or LfD—performing the other type of training second. Each student was given a chance to practice before his/her data was recorded, as described in the following handout to participants:

> You will have four teaching sessions for each task, first Cart Pole and then Mountain Car. Within each task, the order will be practice training by mode A, train by A, practice training by B, train by B. A and B are randomly chosen to be either demonstration or teaching by feedback.

4.1 Domains Used

Our study focused on two simple but well-studied RL domains: Cart Pole and Mountain Car.[4] Cart Pole is explained in our instructions to study participants:

> **Cart Pole:** The agent is a cart that moves left and right to keep a pole balanced upright between a V-shaped angular region. The goal is to keep the pole balanced between the V as long as possible. Cart Pole will automatically restart if you balance for 300 time steps. The green box shows which way the cart is moving (right or left).

The Mountain Car agent must drive an under-powered car up a mountain as our instructions explained:

> **Mountain Car:** The agent is a car that can choose to accelerate left, right, or not at all. The task is for the car to get to the goal (a marker on the top of the right hill) in the least time possible. The green box shows whether the cart is accelerating left, accelerating right, or not accelerating at all. It will restart after 500 time steps.

Note that in an optimal policy, the "no acceleration" action is never used as the agent should always increase its kinetic and potential energy.

4.2 Instructions for LfF and LfD

For completeness, this section provides the instructions given to participants for both training modalities.

> **Demonstration:** You will control the agent, which will later learn to imitate your demonstrations. In both tasks, "J" accelerates left and "L" accelerates right. (As these instructions note later, in the Mountain Car task, you can also not accelerate by pressing "K.")
> Note for teaching by demonstration:
> - The agent will keep performing the last action you give; you don't need to hold the button down or repeatedly press the button.

> **Teaching by feedback:** The agent is in control, but you reward behavior to encourage it and punish behavior to discourage it. The agent expects a small delay in reaction, so be quick but don't worry about being immediate. Faster button pressing is interpreted as stronger reinforcement. In both tasks, "/" (the button with "?" too) rewards and "Z" punishes.
> Notes for teaching by feedback:
> - The agent learns best when reinforcement is both consistent and given very shortly after the action/event being reinforced.
> - Up to a certain level, more frequent feedback generally makes for better learning.
> - Be careful not to give feedback for something that hasn't happened yet. In other words, don't give reinforcement for an action that you anticipate but has not occurred.
> - Not giving feedback communicates that recent behavior was neutral.

5 Results and Discussion

This section presents and discusses the results of our pilot study. Given the fairly small sample size in this pilot study, we do not consider statistical significance. There are two types of performance to consider: the online performance, which expresses the cost of training, and the offline performance of the learned policy, which shows how the agent would perform at any given stopping point.[5] In this second type of analysis, we test the policy after each episode, statically running it offline for 1000 episodes to measure its performance. When not specified, the reader can assume that "performance" refers to the performance of the learned policy, which is measured offline after each episode.

Section 3 discussed the comparative advantages the authors expected of the two teaching modes: three advantages for LfD and four for LfF. These formed our experimental hypotheses. We expected our experimental results, where applicable, to support these proposed advantages.

5.1 Main effect

The main effect with which we are concerned is the relative performances of LfF and LfD in each domain. On the simple measure of the teaching mode's relative performances, LfD constantly outperforms LfF in our experiments both online and by policy, as shown in Figures 2 and 3. However, we will

[3] Mountain Car data was removed for three students who did not finish teaching.

[4] Both tasks were adapted from RL-Library tasks and used RL-Glue to connect the agent, the environment, and the graphical visualizer [Tanner and White, 2009].

[5] For LfD, the online performance is the teacher's performance during demonstration.

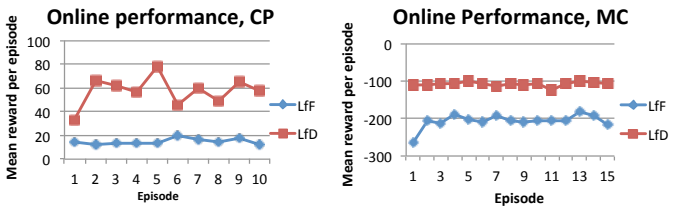


Figure 2: Mean performance per teaching episode during the teaching session for Cart Pole (CP) and Mountain Car (MC)

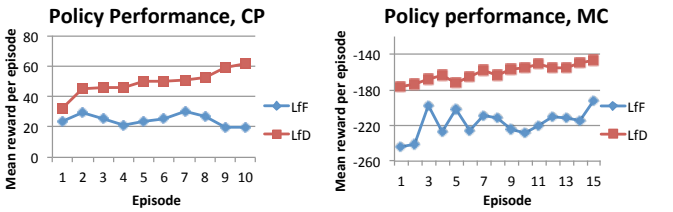


Figure 3: Mean performance of the learned policy, tested after each teaching episode

explore various interaction effects—how varying a second independent variable affects performance in ways that could not be predicted from the teaching mode alone—that strongly suggest that LfD's superior performance is not a general phenomena but is, as hypothesized, situationally dependent.

Also, we expected LfD to have an online advantage in early episodes (discussed in Section 3). However, it is difficult to separate this early advantage from LfD consistently being more effective overall, so we refrain from drawing conclusions about early performance from this study.

5.2 Interaction effects

This section points out three experimental manipulations that appear to affect mean performance, creating interaction effects: (1) the ordering of the teaching modes, (2) the wording of the instructions, and (3) the general experimental design as compared to a previous study. The first and third cases cause changes in the relative performance of the two teaching modes. For these evaluations and their respective figures (4, 5, and 6), we focus on the mean performance of the learned policy during teaching, approximated by averaging performance after each episode of the teaching session.

First, consider the order in which the subjects used the two teaching modalities. For each task, a subject either taught by demonstration and then feedback, or vice-versa. The order was determined by simulated coin toss. Unfortunately, this technique resulted in lopsided group sizes; in both domains, LfF was first much more often than LfD: 9 of 13 times in Mountain Car and 14 of 16 times in Cart Pole. This limits what conclusions can be made about ordering, particularly in Cart Pole. However, preliminary results suggest that LfF benefits much more from being second than LfD. As shown in Figure 4, there is almost no performance difference between LfF and LfD when each is the second teaching mode. In both tasks, LfF benefits from being second, whereas LfD receives a minor benefit in Mountain Car and a sizable decrease in performance in Cart Pole when it is second (though, again, our Cart Pole data is not well balanced in the ordering).

Second, the instructions appear to produce an interaction effect. After 11 subjects finished, the authors attempted to

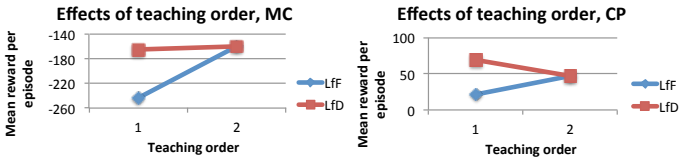


Figure 4: Performance of LfF and LfD when the teaching mode mode is first or second in order

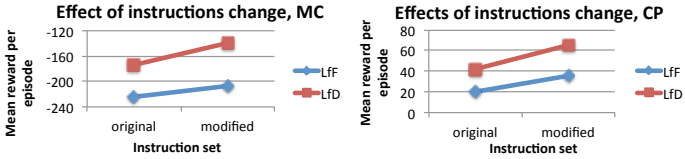


Figure 5: Performance of LfF and LfD after adjusting the subjects' instructions

improve instructions by verbally telling subjects to give feedback frequently for LfF. In our limited data, LfF performance improved on both tasks after this instructions modification (see Figure 5). Surprisingly, the LfD performance improved as well—fully exploring this surprising result is left to future work.

Third, we consider the effect of the general experimental design. In a past experiment [Knox and Stone, 2009], we also tested TAMER on Mountain Car. In this previous study, we prepared subjects more, announced high scores to spur competition and to give a high bar for good performance, and had other design differences. The online performance of the previous LfF study is much closer to that of LfD than the current study's LfF online performance, though LfD still results in the least costly teaching sessions (Figure 6). Policy performance data for the previous study was not immediately available, but the effects discussed in Section 5.4 suggest LfF would make further gains on LfD in policy performance.

5.3 Discussion of main and interaction effects

From the preliminary results presented above, we make *two tentative conclusions* that we will examine with a larger study, carefully designed with lessons learned from this one.

First, both teaching modes are sensitive to the experimental setup. It seems naïve to generalize by saying that one teaching mode is superior to the other. Therefore, our goal going forward will be to find experimental designs that straddle the line for which mode is more effective, allowing various manipulations—teaching order, the amount of subject preparation, or properties of the task—to affect which mode performs best. In other words, a manipulation that increases interacting variable A changes the best teaching mode from LfD to LfF, indicates that LfF becomes more desirable as A increases. Such experiments allow us to better understand the relative strengths and weaknesses of LfF and LfD.

Second, subjects need more preparation for LfF than for LfD, although it is possible that a small amount of preparation provides much more benefit to LfF than LfD. Successful teachers should have some understanding of both the interface and the task. Regarding the interface, LfD has the advantage of being more or less the same as playing a video game, something that people, especially computer science undergraduates, generally have much experience in. LfF has

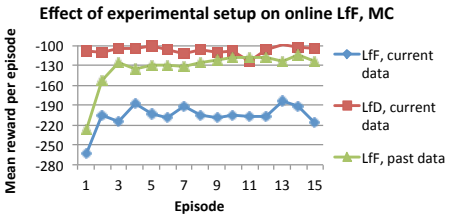


Figure 6: Comparison of Mountain Car results from Figures 2 and 3 with LfF results from a past study

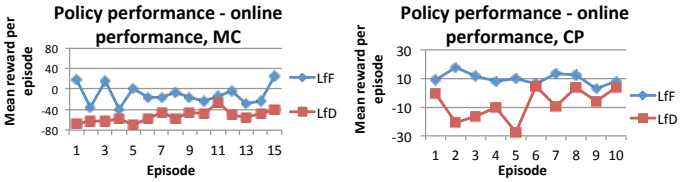


Figure 7: Policy performance minus online performance for LfF and LfD

a simple but alien interface, one that requires at least a small amount of practice and instruction to master. In regard to the task, we suspect that subjects who know little about a task can learn a good strategy more easily through direct control in LfD than by indirect control by feedback. Therefore, before training novel tasks by feedback, teachers should be able to perform the tasks themselves or see them competently performed. This tentative conclusion appears to contradict the second expected strength of LfF from Section 3, but it may operate in a different space of task expertise; LfF may require more knowledge about the quality of various state trajectories, but less knowledge about the transition model (mapping current state and action to next state to predict actions' effects), which may not be a factor for these tasks since the transition models are simple and easy for a trainer to learn.

5.4 Online vs. Offline Performance

If there is a large discrepancy between an agent's online behavior (during teaching) and the behavior of its learned policy (during offline testing), the teacher might not be aware of the weaknesses of the agent's learned policy.[6] For example, an LfD agent's policy representation might not be able to represent aspects of the demonstrations, causing the agent to enter regions of the state space that are far from any visited during the demonstrations and are thus unknown. We examine this discrepancy by looking at the difference in online performance and policy performance. Since an LfF agent is using its learned policy during teaching, we expect a lower performance discrepancy than for LfD, where the model of the demonstrations is hidden during teaching.

Figure 7, shows the performance differences—policy performance minus online performance—in the mean reward received per episode for both tasks. This difference can be seen as a type of error, where the policy performance is considered actual performance and the teaching performance is considered predicted performance. In Mountain Car, our expectations are met: LfF stays centered more or less just below zero error, whereas LfD policies consistently perform much worse than the online demonstration for that episode. In Cart Pole, however, both teaching modes have a similar amount of error over a training session. Although the policies learned from demonstration again underperform the demonstrations that guide them, LfF actually results in better end-of-episode policies than the agents had shown during teaching. A potential explanation of this surprising result is that the end-of-episode policy has more learning samples than the online agent during that episode. But this explanation does not suffice, since the agents do not improve online in the next episode by the amount of the error.

In summary, on one task LfF had less discrepancy between online performance and policy performance than did LfD, and on another task the two modes had similar magnitudes of discrepancies. Thus, our prediction was only partially borne out. We also note that LfF's learned policy did as well or better than during teaching, whereas LfD's policy did as well or worse.

6 Additional Related Work

To the best of our knowledge, no prior research focuses on comparing and evaluating different methods for integrating teaching into an agent acting in an MDP.

Related to all the techniques below, but fundamentally different, is that of *transfer learning*, which typically allows a *target agent* to directly access a *source agent*'s "brain." The challenge is typically to most effectively re-use the source agent knowledge when the target agent has different capabilities [Taylor and Stone, 2009]. A more recently investigated form of transfer relies on a human constructing a sequence of tasks for the agent to train on [Taylor, 2009; Zang *et al.*, 2010].

The remainder of this section surveys some of the most relevant related work that focuses on using a human to help teach an agent, categorizing systems along two dimensions. The first, and more important, is teaching modality, which determines the type of interaction the teacher and agent will have. The second is whether the agent has access to (and uses) reward from the underlying MDP.

6.1 Demonstration

This section discusses research related to LfD that were not discussed previously in Section 2.3.

A slightly different form of learning from human demonstrations is *Inverse Reinforcement Learning* [Abbeel and Ng, 2004]. For this form of LfD, the a human temporarily controls an agent within an MDP\R environment. Instead of learning a policy that directly mimics the human, an agent employing apprenticeship learning infers a reward function, R, from the human-provided examples, and then uses RL to optimize the policy on this inferred MDP, rather than attempting to directly mimic the human's policy.

LfD has also been used to improve reinforcement learning. Imitation learning may aim to directly mimic the agent/human's policy [Smart and Kaelbling, 2000; Price and Boutilier, 2003], or to use it as a starting point for optimizing the policy on the original MDP [Taylor *et al.*, 2011].

6.2 Reinforcement

Within psychology, *behavioral shaping* [Skinner, 1953] is a training procedure that uses reinforcement to condition the

[6] We use "offline" and "policy" as interchangeable modifiers of "performance" and "testing."

desired behavior in a human or animal. During training, the reward signal is initially used to reinforce any tendency towards the correct behavior, but is gradually changed to reward successively more difficult elements of the task. Shaping methods with human-controlled rewards have been successfully demonstrated in a variety of software agent applications [Blumberg *et al.*, 2002; Kaplan *et al.*, 2002].

Three other approaches integrated human training with autonomous learning. In Thomaz and Breazeal [2006], a table-based Q-learning agent in a virtual kitchen environment learns from a reward signal that is the sum of human reinforcement and MDP reward (a type of shaping rewards approach [Dorigo and Colombetti, 1994; Mataric, 1994]). In Isbell et al. [2006], an agent uses RL to learn social behavior from multiple sources of human reinforcement. Judah et al. [2010] alternate between "practice", where actual world experience is gathered, and an offline labeling of actions as good or bad by a human critic, an approach that blurs the line between feedback and demonstration. The human criticism is used to judge the expected value of candidate policies while also automatically determining the level of influence given to the criticism.

6.3 Advice

High-level advice and suggestions have also been used to bias agent learning. Such advice can provide a powerful learning tool that speeds up learning by biasing the behavior of an agent and reducing the policy search space. However, existing methods typically require either a significant user sophistication (e.g., the human must use a specific programming language to provide advice [Maclin and Shavlik, 1996]) or significant effort is needed to design a human interface (e.g., the learning agent must have natural language processing abilities [Kuhlmann *et al.*, 2004]).

7 Conclusions and Future Work

Several patterns emerge from the pilot study that directly address the teaching modes' expected comparative strengths, as described in Section 3. LfD generally results in higher performance than LfF, which we suspect to usually be the case. However, several interaction effects indicate that the relative performance of each method is highly situational, supporting our overall thesis that each method has unique strengths. More investigation is needed to determine which circumstances favor each method. Additionally, the performance of the policy learned from feedback was *at least* as good as the performance during teaching, whereas the performance of an LfD policy was *at most* as good as during the demonstrations (removing apparent episode-to-episode noise).

However, these patterns are not conclusive, because of a small sample size and changing conditions during the pilot study. In the future, we plan to redesign the experiment, better preparing the subjects to teach well, balancing the number of subjects in each condition, and carefully choosing what manipulations we use to further test when each teaching modality should be used. Additionally, we hope to use our understanding of the relative strengths of LfF and LfD to design an agent that appropriately uses both learning mechanisms.

Acknowledgments

This work has taken place in part in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by grants from the National Science Foundation (IIS-0917122), ONR (N00014-09-1-0658), and the Federal Highway Administration (DTFH61-07-H-00030). The first author is supported by an NSF Graduate Research Fellowship.

References

[Abbeel and Ng, 2004] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. *ACM International Conference Proceeding Series*, 2004.

[Argall *et al.*, 2007] B. Argall, B. Browning, and M. Veloso. Learning by demonstration with critique from a human teacher. HRI, 2007.

[Argall *et al.*, 2009] B. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469 – 483, 2009.

[Blumberg *et al.*, 2002] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. P. Johnson, and B. Tomlinson. Integrated learning for interactive synthetic characters. *SIGGRAPH*, 2002.

[Dorigo and Colombetti, 1994] M. Dorigo and M. Colombetti. Robot shaping: Developing situated agents through learning. *Artificial Intelligence*, 1994.

[Grollman and Jenkins, 2007] D. Grollman and O. C. Jenkins. Dogged learning for robots. *ICRA*, 2007.

[Isbell *et al.*, 2006] C. L. Isbell, M. Kearns, S. Singh, C. R. Shelton, P. Stone, and D. Kormann. Cobot in LambdaMOO: An Adaptive Social Statistics Agent. *AAMAS*, 2006.

[Judah *et al.*, 2010] K. Judah, S. Roy, A. Fern, and T. G. Dietterich. Reinforcement Learning Via Practice and Critique Advice. *AAAI*, 2010.

[Kaplan *et al.*, 2002] F. Kaplan, P-Y. Oudeyer, E. Kubinyi, and A. Miklosi. Robotic clicker training. *Robotics and Autonomous Systems*, 38(3-4):197 – 206, 2002.

[Knox and Stone, 2009] W. B. Knox and P. Stone. Interactively shaping agents via human reinforcement: The TAMER framework. In *KCAP*, 2009.

[Kuhlmann *et al.*, 2004] G. Kuhlmann, P. Stone, R. J. Mooney, and J. W. Shavlik. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *Proceedings of the AAAI Workshop on Supervisory Control of Learning and Adaptive Systems*, 2004.

[Maclin and Shavlik, 1996] R. Maclin and J. W. Shavlik. Creating advice-taking reinforcement learners. *Machine Learning*, 22(1-3):251–281, 1996.

[Mataric, 1994] M. J. Mataric. Reward functions for accelerated learning. *ICML*, 1994.

[Nicolescu *et al.*, 2008] M. N. Nicolescu, O. C. Jenkins, A. Olenderski, and E. Fritzinger. Learning behavior fusion from demonstration. *Interaction Studies*, 9(2):319–352, 2008.

[Price and Boutilier, 2003] B. Price and C. Boutilier. Accelerating reinforcement learning through implicit imitation. *JAIR*, 19:569–629, 2003.

[Skinner, 1953] B. F. Skinner. *Science and Human Behavior*. Colliler-Macmillian, 1953.

[Smart and Kaelbling, 2000] W. D. Smart and L. P. Kaelbling. Practical reinforcement learning in continuous spaces. *ICML*, 2000.

[Smart and Kaelbling, 2002] W. D. Smart and L. P. Kaelbling. Effective reinforcement learning for mobile robots. In *ICRA*, 2002.

[Tanner and White, 2009] B. Tanner and A. White. RL-Glue: Language-independent software for reinforcement-learning experiments. *JMLR*, 10, 2009.

[Taylor and Stone, 2009] M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *JMLR*, 10(1):1633–1685, 2009.

[Taylor *et al.*, 2011] M. E. Taylor, H.B. Suay, and S. Chernova. Integrating reinforcement learning with human demonstrations of varying ability. *AAMAS*, 2011.

[Taylor, 2009] M. E. Taylor. Assisting transfer-enabled machine learning algorithms: Leveraging human knowledge for curriculum design. In *The AAAI Symposium on Agents that Learn from Human Teachers*, 2009.

[Thomaz and Breazeal, 2006] A. L. Thomaz and C. Breazeal. Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance. *AAAI*, 2006.

[Zang *et al.*, 2010] P. Zang, A. J. Irani, P. Zhou, C. L. Isbel Jr., and A. L. Thomaz. Learn via human-provided sequence of tasks using training regimens to teach expanding. In *AAMAS*, 2010.