Learning From Feedback on Actions Past and Intended

W. Bradley Knox
Department of Computer Science
University of Texas at Austin

Peter Stone
Department of Computer Science
University of Texas at Austin

Cynthia Breazeal
Media Lab
Massachusetts Institute of Technology

Categories and Subject Descriptors

I.2.6 [**Learning**]: Knowledge acquisition

Keywords

teachable agents, interactive shaping, human-robot interaction, reinforcement learning

1. INTRODUCTION

Robotic learning promises to eventually provide great societal benefits. In contrast to pure trial-and-error learning, human instruction has at least two benefits: (1) Human teaching can lead to much faster learning. For instance, humans can model the delayed outcome of a behavior and give feedback immediately, unambiguously informing the robot of the quality of its recent action. (2) Human instruction can serve to define a task objective, empowering end-users that lack programming skills to customize behavior.

The TAMER framework [3, 2] was developed to provide a learning mechanism for a specific, psychologically grounded [1] form of teaching—through signals of reward and punishment. TAMER breaks the process of interactively learning behaviors from live human reward into three modules: *credit assignment*, where delayed human reward is applied appropriately to recent events; *regression* on experienced events and their consequential credited reward *to create a predictive model* for future reward; and *action selection* using the model of human reward.

TAMER differs from traditional reinforcement learning (RL) algorithms—generally powerful algorithms that are intuitive but ultimately ill-suited for learning from human reward—in multiple ways. For instance, human reward is stochastically delayed from the event that prompted it, and TAMER acknowledges this delay, absent in traditional reinforcement learning, and adjusts for it. And importantly, human trainers consider the long-term effects of actions, making each reward a complete judgment on the quality of recent actions; therefore, predictions of near-term human reward are analogous to estimates of expected long-term reward in reinforcement learning, simplifying action selection to choosing the action with the highest expected human reward.

On multiple tasks, TAMER agents have been shown to learn more quickly—sometimes dramatically so—than counterparts that learn from a predefined evaluation function instead of human interaction. Further, the TAMER framework gives primacy to the desires of human trainers—learning only from these trainers—many of whom had no programming skills. Thus, TAMER is especially well suited for fitting robotic behaviors to an individual's unique demands, empowering those without programming skills to specify correct behavior.

Until now, TAMER has only been implemented on simulated tasks. The first contribution of this paper is an implementation of TAMER on the robot Nexi to teach interactive navigational behaviors, with proof-of-concept results.

Additionally, humans that have been instructed to give feedback sometimes do so for actions that have not yet occurred. Thomaz et al. [4] observed this tendency, as we have during informal testing of TAMER. These observations suggest that people may be inferring a robot's intention and giving feedback on the intended behavior. For further illustration, imagine your vocal reaction to a child walking towards a dangerous street or a dog staring at and inching towards a piece of steak within its reach. If your imagined reaction is similar to ours, your voice would carry strong negative affect. However, the hypothetical actors—the child and the dog—have done nothing wrong or harmful yet. Indeed, it is only their inferred intention that draws strong, negative feedback. The second contribution of this paper is to redesign TAMER to incorporate feedback both for past and intended actions. Allowing what we term "intentional feedback" both fits the learning system to natural human tendencies and introduces a crucial warning system: a robot need not experience a catastrophic event to learn to avoid it. Intending the event is sufficient.

2. TEACHING A ROBOT INTERACTIVE NAVIGATIONAL BEHAVIORS

We implemented TAMER to teach interactive navigational tasks on Nexi (Figure 1), a humanoid robot capable of social expression. Nexi moves with a two-wheel base and, in this task domain, senses its environment through a Vicon Motion Capture system that determines the 3-dimensional locations and orientations of the robot and the trainer. In our development and testing, we use both the physical robot and a simulation running on jMonkeyEngine.

In the task domain, the robot has 5 navigational actions: move forward, move backward, turn left, turn right, or stay. From the Vicon sensor data, the robot creates a number of features that describe it, the trainer, and their relation in the environmental space. A subset of these features are chosen to create the robot's state, the context that determines

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.
ACM 978-1-4503-1063-5/12/03.

what task behavior is correct. Trainer-based features make interactive behaviors possible.

We have successfully trained Nexi to exhibit several high-level behaviors. Two simple behaviors have been taught both in simulation and on the physical robot: facing the trainer and maintaining distance from the trainer in one dimension. For features, these behaviors respectively use the angle to the trainer and the distance to the trainer. A video of the final distance-maintenance behavior can be seen at `cs.utexas.edu/~bradknox/nexi`. In simulation only, two more complex behaviors were taught: follow the human (i.e., stay below a maximum distance) and avoid the human (i.e., stay above a minimum distance). For features, both behaviors use the distance and angle to the trainer from the robot. At the same URL, one can view a video of the follow behavior being trained from start to finish over less than five minutes. Note that when Nexi chooses a bad action, considerable time is lost in returning to the state before that action was taken; undesired actions have costs that go beyond their duration. This observation motivates the inclusion of intentional feedback, which can stop bad behavior before it occurs.

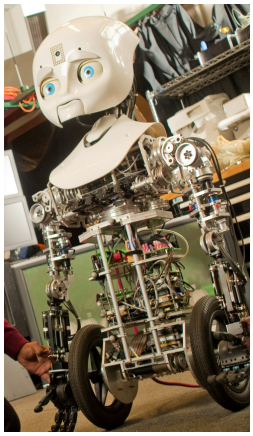

Figure 1: Nexi.

3. TAMER WITH INTENTIONAL BEHAVIOR

To make TAMER learn from intentional feedback, we make two major changes. First, in addition to regular task behavior, the robot exhibits intentional behavior that communicates its planned task behavior. Second, we adapt the learning algorithm to receive feedback on such intentional behavior.

In the current prototype version, the robot's intentional behaviors are announcements of its planned next task action. The name of the next planned action—"right", "left", etc.—is spoken while the current action is performed. For simplicity, we separate intentional behavior from task behavior, though in nature these behaviors often overlap. For instance, a stalking cat often freezes low to the ground before pouncing on its target; this behavior both communicates its intention to an observer and helps its hunting task by keeping the cat hidden.

The feedback interface we used for TAMER consisted of two push-buttons: one for positive and one for negative feedback on task behavior. We add two more buttons for positive and negative intentional feedback.

Learning from feedback on intentional behavior uses the same mechanisms as TAMER uses for task behavior. To learn a model of human reward, TAMER creates labeled state-action pairs on which a regression algorithm trains. These labels are determined through a credit assignment technique that maintains a window of recent task behavior and spreads credit from any reward received over the state-action pairs in this window [2]. For intentional behavior, we create a parallel window that likewise produces samples using the same credit assignment technique. A sample from intentional feedback consists of the predicted next state, the action that is planned for that state and exhibited through intentional behavior, and the credited intentional reward for that intended state-action pair. Thus, just as the robot is exhibiting one task action at any time, it also exhibits one intentional action; the human trainer delivers stochastically delayed reward to either type of action, and reward is credited through identical, parallel credit assignment mechanisms. The labeled state-action pairs from both task and intentional behavior are added to a single body of training data for the model of human reward.

Only displaying the next action as intended behavior may not describe the full range of intentional behaviors—intentions can range over different time scales—but it nonetheless provides a powerful improvement (as we expect our comparative experiments to show) over the TAMER algorithm that we use as a starting point. Further, we believe it would not be difficult to extend this algorithm to more complex intentions; if more than one state-action pair creates a single intention behavior, they can all be stored in the credit assignment window and the weight of each pair's resultant sample would be adjusted by the proportion of impact it had in determining the intention behavior.

4. CONCLUSION

In this paper, we present the first implementation of TAMER on a robot and a reformulation of TAMER to allow trainers to influence intended behavior, powerfully giving them the ability to give feedback on bad behavior before it occurs, yielding the benefits but removing the cost of potentially catastrophic learning experiences.

To complete this project, we will create more natural intention behaviors—movements of the eyes, head, and torso to indicate planned directions of movement—and conduct experimental evaluations of training with and without intentional feedback. In future work, we will unify the input interface for reward on past and intended actions, using prosody—vocal acoustic characteristics such as pitch and volume—to determine whether feedback utterances are focused on intention or past action.

5. ACKNOWLEDGMENTS

This work has taken place in the Personal Robotics Group of the Media Lab at MIT and the Learning Agents Research Group (LARG) at UT Austin. We thank the sponsors of the Media Lab and of LARG, which is supported in part by NSF (IIS-0917122), ONR (N00014-09-1-0658), and the FHWA (DTFH61-07-H-00030). We also thank Sigurdur Orn Adalgeirsson, Nick DePalma, and Adrian Mullings for their extensive technical support of Nexi.

6. REFERENCES

[1] M. Bouton. *Learning and Behavior: A Contemporary Synthesis*. Sinauer Associates, 2007.

[2] W. Knox and P. Stone. Interactively shaping agents via human reinforcement: The TAMER framework. *The Fifth International Conference on Knowledge Capture*, 2009.

[3] W. B. Knox and P. Stone. TAMER: Training an agent manually via evaluative reinforcement. In *IEEE 7th International Conference on Development and Learning*, August 2008.

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