Training a Tetris Agent via Interactive Shaping: A Demonstration of the TAMER Framework

W. Bradley Knox and Peter Stone
University of Texas at Austin
{bradknox, pstone}@cs.utexas.edu

1. INTRODUCTION

As computational learning agents continue to improve their ability to learn sequential decision-making tasks, a central but largely unfulfilled goal is to deploy these agents in real-world domains in which they interact with humans and make decisions that affect our lives. People will want such interactive agents to be able to perform tasks for which the agent's original developers could not prepare it. Thus it will be imperative to develop agents that can learn from natural methods of communication. The teaching technique of shaping is one such method. In this context, we define *shaping* as training an agent through signals of positive and negative reinforcement.[1] In a shaping scenario, a human trainer observes an agent and reinforces its behavior through push-buttons, spoken word ("yes" or "no"), facial expression, or any other signal that can be converted to a scalar signal of approval or disapproval. We treat shaping as a specific mode of knowledge transfer, distinct from (and probably complementary to) other natural methods of communication, including programming by demonstration and advice-giving. The key challenge before us is to create agents that can be shaped effectively. Our problem definition is as follows:

The Shaping problem Within a sequential decision-making task, an agent receives a sequence of state descriptions ($s_1, s_2, ...$ where $s_i \epsilon S$) and action opportunities (choosing $a_i \epsilon A$ 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 scalar reinforcement signals ($h_1, h_2, ...$) that are 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?

Expected benefits of learning from human reinforcement include the following:

[1]We use the term "shaping" as it is used in animal learning literature (in which it was initially developed by B.F. Skinner). There, shaping is defined as training by reinforcing successively improving approximations of the target behavior [3]. In reinforcement learning literature, it is sometimes used as in animal learning, but more often "shaping" is restricted to methods that combine the shaping reinforcement signal and the reward signal of the environment into a single signal [9].

Cite as: Training a Tetris Agent via Interactive Shaping: A Demonstration of the TAMER Framework, W. Bradley Knox and Peter Stone, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1767-1768.

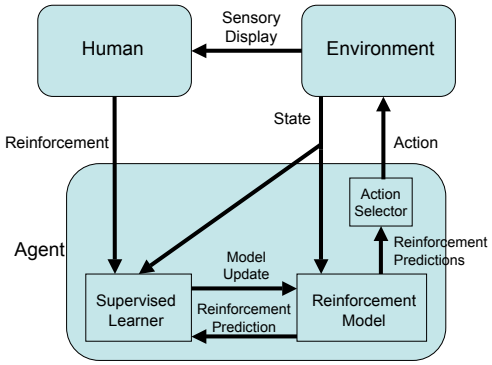

Figure 1: Framework for Training an Agent Manually via Evaluative Reinforcement (TAMER).

1. Compared to learning from the environmental reward of a Markov Decision Process, shaping can decrease sample complexity for learning a "good" policy, consuming less resources in real-world domains.
2. An agent can learn in the absence of a coded evaluation function (e.g., an environmental reward function).
3. The simple mode of communication allows lay users to teach agents the policies which they prefer, even changing the desired policy if they choose.
4. Shaped agents can learn in more complex domains than autonomous learning allows.

Previous results, described later, support the first three of these benefits. These results and other work on the TAMERframework have appeared in several publications [4, 7, 5] and will be presented as a full paper at AAMAS [8]. This technical report describes our framework for agents that can be interactively shaped, briefly discusses previously published experimental results in the domain of Tetris, and explains our demonstration of an interactively trainable Tetris TAMER agent at AAMAS 2010.[2]

2. THE TAMER FRAMEWORK

In our previous work on shaping, we introduced a framework called Training an Agent Manually via Evaluative Reinforcement (TAMER). The TAMER framework, shown in Figure 1, is an approach to the Shaping Problem that makes use of established supervised learning techniques to model a human's reinforcement function and bases its action selection on the learned model. If acting greedily, a TAMER agent chooses actions that are projected to receive the most reinforcement.

[2]Much of this report overlaps with previous, already cited work by the authors.

Table 1: Results of various Tetris agents.

Method	Mean Lines Cleared		Games
	at Game 3	at Peak	for Peak
TAMER	**65.89**	**65.89**	**3**
RRL-KBR [10]	5	50	120
Policy Iteration [1]	~ 0 (no learning until game 100)	3183	1500
Genetic Algorithm [2]	~ 0 (no learning until game 500)	586,103	3000
CE+RL [12]	~ 0 (no learning until game 100)	348,895	5000

Table 2: A comparison of the TAMER Tetris agent and various agents that learn from MDP reward.

The TAMER framework is designed for Markov Decision Processes that have the reward function R unspecified (MDP\R). A TAMER agent seeks to learn the human trainer's reinforcement function $H : S \times A \rightarrow \mathbb{R}$. Presented with a state *s*, the agent consults its learned model $\hat{H}$ and, if choosing greedily, takes the action *a* that maximizes $\hat{H}(s, a)$. Since the agent seeks only to maximize human reinforcement, the optimal policy is defined solely by the trainer, who could choose to train the agent to perform any behavior that its model can represent. Therefore, when the agent's performance is evaluated using an objective metric, its performance will be limited by the information provided by the teacher.

The principle challenge for autonomously learning agents (i.e., those that receive feedback in the form of MDP reward rather than human reinforcement) is to assign credit from environmental reward to the entire history of past state-action pairs. A key insight of the TAMER framework is that the difficult problem of credit assignment inherent in reinforcement learning is no longer present with an attentive human trainer. The trainer can evaluate an action or short sequence of actions, considering the long-term effects of each, and deliver positive or negative feedback within a small temporal window after the behavior. Assuming that credit is properly assigned within the temporal window (which we address in Knox and Stone [7]), we assert that a trainer can directly label behavior. Therefore, modeling the trainer's reinforcement function *H* is a supervised learning problem.

3. EXPERIMENTAL RESULTS

We developed TAMER algorithms for two contrasting task domains – Tetris and Mountain Car. Tetris has a complex state-action space and low time step frequency, and Mountain Car is simpler but occurs at a high frequency (seven actions per second). Our experimental data (see Table 1, for example) suggests that TAMER agents outperform autonomous learning agents in the short-term, arriving at a "good" policy after very few learning trials. It also suggests that well-tuned autonomous agents are better at maximizing final, peak performance after many more trials [4, 5].

4. THE TETRIS DEMONSTRATION

Our demonstration at AAMAS will begin with a short introduction. Then, we will allow audience members to train TAMERTetris agents. Trainers will use a handheld remote to deliver reinforcement. Additionally, the agent algorithm is improved since our experiments, yielding final performance that is anecdotally five to ten times as high as previously reported. Videos of an agent being trained can be found on YouTube at *tinyurl.com/tetrisbefore*, *tinyurl.com/tetristraining*, and *tinyurl.com/tetrisafter*.

5. CONCLUSION

The TAMER framework, which allows human trainers to shape agents via positive and negative reinforcement, provides an easy-to-implement technique that:

1. works in the absence of an environmental reward function,
2. reduces sample complexity, and
3. is accessible to people who lack knowledge of computer science.

The TAMER Tetris agent has been demonstrated previously with success, specifically at the 2009 IJCAI Robotics Exhibition [6]. Put simply, our experimental results suggest that TAMER outperforms autonomous learners with few learning samples and that autonomous learners perform better in the long term. From this observation, combining the complementary strengths of the two learning signals (MDP reward and human reinforcement) appears to be a promising approach. More recent work on the TAMER framework, which systematically tests and analyzes multiple techniques for combining TAMER with SARSA(λ), a commonly used reinforcement learning algorithm, will be presented as a full paper during AAMAS [8].

6. REFERENCES

[1] D. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[2] N. Bohm, G. Kokai, and S. Mandl. Evolving a heuristic function for the game of Tetris. *Proc. Lernen, Wissensentdeckung und Adaptivitat LWA*, 2004.

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

[4] W. Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *IEEE 7th International Conference on Development and Learning (ICDL)*, August 2008.

[5] W. Bradley Knox, Ian Fasel, and Peter Stone. Design principles for creating human-shapable agents. In *AAAI Spring 2009 Symposium on Agents that Learn from Human Teachers*, March 2009.

[6] W. Bradley Knox and Peter Stone. Interactive Shaping of a Tetris Agent Using the TAMER Framework. Technical report. In *Proceedings of the IJCAI Robot Workshop*, July 2009.

[7] W. Bradley Knox and Peter Stone. Interactively Shaping Agents via Human Reinforcement: The TAMER Framework. In *Proceedings of The Fifth International Conference on Knowledge Capture (K-CAP)*, September 2009.

[8] W. Bradley Knox and Peter Stone. Combining Manual Feedback with Subsequent MDP Reward Signals for Reinforcement Learning. To appear in *Proceedings of The Ninth Annual International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2010.

[9] A.Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. *ICML*, 1999.

[10] J. Ramon and K. Driessens. On the numeric stability of gaussian processes regression for relational reinforcement learning. *ICML-2004 Workshop on Relational Reinforcement Learning*, 2004.

[11] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[12] I. Szita and A. Lorincz. Learning Tetris Using the Noisy Cross-Entropy Method. *Neural Computation*, 18(12), 2006.