

Shaping Agents via Human Reinforcement

Thesis Statement

W. Bradley Knox

University of Texas at Austin

bradknox@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, making decisions that affect our lives. However, with real-world deployment comes real-world costs. For such a deployment to be viable, agents will not be able to use hundreds or thousands of learning trials to reach a good policy when each suboptimal trial is costly. For example, an autonomous driving agent should not learn to drive by crashing into road barriers and endangering the lives of pedestrians.

Fortunately, for many of these tasks, humans have domain knowledge that could speed the learning process, reducing costly sample complexity. Currently, most knowledge transfer from humans to agents occurs via programming, which is time-consuming and inaccessible to the general public. It is important 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.

My thesis will focus on the question:

How can agents harness the information contained in signals of positive and negative evaluation from a human to learn sequential decision-making tasks?

Expected benefits of learning from human reinforcement²

¹We 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 [Bouton, 2007]. 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 [Ng *et al.*, 1999].

²In this thesis statement, I distinguish between human reinforcement

include the following:

1. Shaping decreases 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 the first three of these benefits. Some of these results have appeared in two published papers [Knox and Stone, 2008; Knox *et al.*, 2009].

2 Completed work: 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, makes use of established supervised learning techniques to model a human’s reinforcement function and then uses the learned model to choose actions that are projected to receive the most reinforcement.

ment and environmental reward within an Markov Decision Process. To avoid confusion, human feedback is always called “reinforcement.”

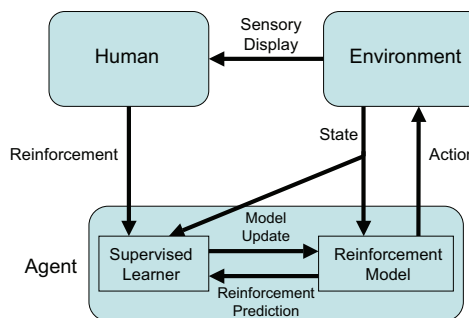


Figure 1: Framework for Training an Agent Manually via Evaluative Reinforcement (TAMER).

At the highest level of description, the TAMER framework allows a human to train an agent to perform a task via positive and negative reinforcement signals. From the agent’s point of view, its goal is to model the human’s reinforcement while exploiting that model to earn as much immediate reinforcement as possible. The human trainer’s goal is, therefore, to provide reinforcement that leads the agent to perform the task well.

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. When the agent’s performance is evaluated using an objective metric, we expect its performance to 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 environmental reward) 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. Assigning credit within that window presents a challenge in itself, which we address by defining a probability density function $f(x)$ over the time window and distributing “credit” among the actions as follows. If t and t' are times of consecutive time steps, credit for the time step at t is $h \times \int_{t'}^t f(x) dx$. Assuming that credit is properly assigned within the temporal window, we assert that a trainer can directly label behavior. Therefore, modeling the trainer’s reinforcement function H is a supervised learning problem.

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). In Mountain Car, to assign credit from the human reinforcement to each recent action, we used a Uniform(0.2, 0.6) distribution, where the parameters are in seconds.

Our experimental data 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.

3 Future Thesis Work

There are many open questions and interesting directions that could be taken. The completed thesis will contain of the following:

Table 1: Results of various Tetris agents.

Method	Mean Lines Cleared		Games for Peak
	at Game 3	at Peak	
TAMER	65.89	65.89	3
RRL-KBR [2004]	5	50	120
Policy Iteration [1996]	~ 0 (random until game 100)	3183	1500
Genetic Algorithm [2004]	~ 0 (random until game 500)	586,103	3000

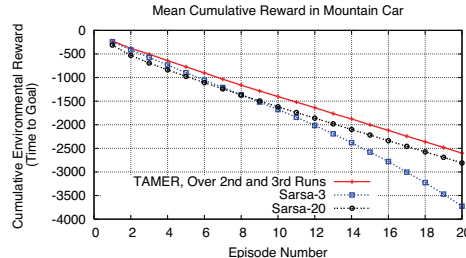


Figure 2: Mean cumulative environmental reward received for the mountain car task.

1. Our results indicate that TAMER agents learn quickly, and autonomous agents can reach higher final performance after a long training period. Given this difference in strengths, I aim to explore how best to use both human reinforcement, h , and, when available, environmental reward, r , relying on the former more heavily for early learning and on the latter for fine-tuning to achieve better results than either method can achieve in isolation.
2. I will investigate how to extend TAMER to a variety of human interfaces for communicating the reinforcement signal to the agent and to training scenarios that violate our initial assumptions.
3. Thus far, we have tested linear models with gradient descent updates within TAMER. I will test a wide variety of model representations and training methods and evaluate which, specifically and categorically, are effective in modeling human reinforcement and yielding good policies within sequential decision-making tasks.
4. In addition to asking how to create the best human-shapable agent, I will investigate how humans shape the agents. In the field of animal learning, not much work has been done on how people teach via reinforcement, making this a potentially fruitful line of investigation.

References

- [Bertsekas and Tsitsiklis, 1996] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- [Bohm et al., 2004] N. Bohm, G. Kokai, and S. Mandl. Evolving a heuristic function for the game of Tetris. *Proc. Lernen, Wissensentdeckung und Adaptivitat LWA*, 2004.
- [Bouton, 2007] M.E. Bouton. *Learning and Behavior: A Contemporary Synthesis*. Sinauer Associates, 2007.
- [Knox and Stone, 2008] W. Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *IEEE 7th International Conference on Development and Learning*, August 2008.
- [Knox et al., 2009] 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.
- [Ng et al., 1999] A.Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. *ICML*, 1999.
- [Ramon and Driessens, 2004] J. Ramon and K. Driessens. On the numeric stability of gaussian processes regression for relational reinforcement learning. *ICML-2004 Workshop on Relational Reinforcement Learning*, pages 10–14, 2004.