Teaching Agents with Human Feedback: A Demonstration of the TAMER Framework

W. Bradley Knox
Massachusetts Institute of Technology
bradknox@mit.edu

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

Cynthia Breazeal
Massachusetts Institute of Technology
cynthiab@media.mit.edu

ABSTRACT

Incorporating human interaction into agent learning yields two crucial benefits. First, human knowledge can greatly improve the speed and final result of learning compared to pure trial-and-error approaches like reinforcement learning. And second, human users are empowered to designate "correct" behavior. In this abstract, we present research on a system for learning from human interaction—the TAMER framework—then point to extensions to TAMER, and finally describe a demonstration of these systems.

Author Keywords

reinforcement learning; modeling and prediction of user behavior; end-user programming; human-agent interaction; interactive machine learning

ACM Classification Keywords

H.1.2 User/Machine Systems: Miscellaneous

OVERVIEW

Software-based control systems are often deployed in the service of end users that lack programming skills. Examples of these control systems (i.e., autonomous agents) include autonomous vehicles, personal robots, and game-playing agents. This abstract describes research on TAMER, a general framework (Figure 1) for control algorithms that learn from real-valued signals of user approval and disapproval (i.e., reward) through simple human-machine interfaces that do not require technical expertise on the part of the user. These algorithms provide two distinct benefits: (1) giving general users the ability to specify correct behavior for a control system and (2) incorporating available human task expertise to increase learning speed on tasks with predefined objective functions. This work makes progress towards answering the open question of how best to learn from human-generated reward, a potential source of guidance that will be abundant for many robots through social cues such as smiles and attention. The work to be demonstrated has resulted in a number of publications [5, 7, 2, 4], including the 2010 Best Student Paper at AAMAS [3], a 2012 finalist for the CoTeSys Cognitive Robotics Best Paper award at Ro-Man [6], and a paper at IUI this year [8].

IUI'13 Companion, March 19–22, 2013, Santa Monica, CA, USA.
ACM 978-1-4503-1966-9/13/03.

THE TAMER FRAMEWORK

The TAMER framework models a human's internal feedback function and chooses actions that maximize human reward as predicted by the current model. TAMER applies to both robotic and simulated agents. On multiple tasks, we have shown that TAMER agents can learn more quickly—sometimes dramatically so—than their more traditional counterparts, which specifically are tabula rasa agents that learn from a predefined evaluation function instead of human interaction. On Tetris for example (see Figure 2(a)), TAMER agents reached a mean performance of 50 lines cleared per game in less than 5 games of training, a visually impressive level of play that human-free algorithms required tens or hundreds of games to reach. 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 well suited for fitting robotic behaviors to an individual's unique demands, empowering those without programming skills to specify correct behavior. TAMER has been successfully implemented on the robot Nexi to teach interactive navigational behaviors such as following and avoidance of the trainer (Figure 2(b)).

As shown in Figure 1, TAMER breaks the process of learning behaviors from live human reward into three modules, making contributions to the design of each:

1. **credit assignment**, where delayed human reward is applied appropriately to recent events;
2. **regression** on experienced events and their consequential credited reward **to create a predictive model** for future reward; and
3. **action selection** using the model of human reward.

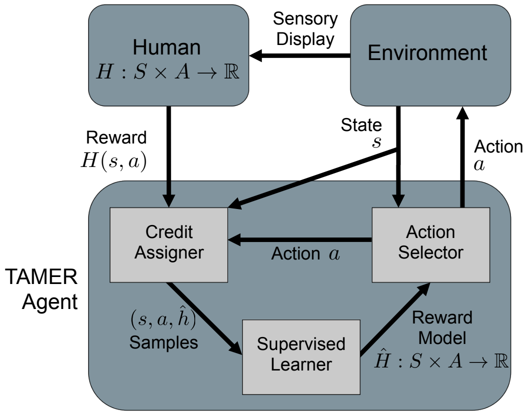

Figure 1. Conceptual diagram for the TAMER framework.

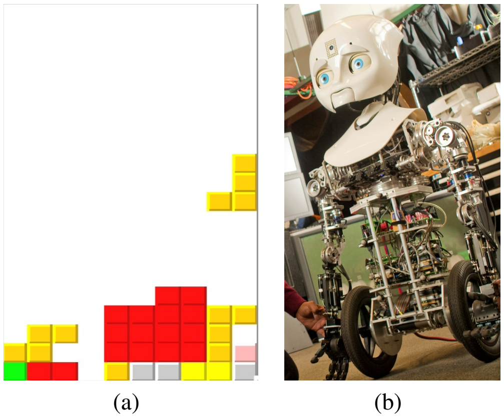
(a) (b)

Figure 2. Task domains to be demonstrated with TAMER agents: Tetris and interactive robotic navigation

TAMER differs from traditional reinforcement learning (RL) algorithms—generally powerful algorithms that cannot be naively applied to human-generated reward [6]—in multiple ways. For instance, human reward is delayed from the event that prompted it, and TAMER acknowledges this delay, absent in traditional reinforcement learning, and adjusts for it. And importantly, whereas RL algorithms attempt to maximize their accumulation of reward over the long-term, TAMER algorithms focus only on reward caused directly by immediate action. Thus when acting greedily, TAMER chooses the action expected to elicit the most reward within the current state. This myopic, shortsighted approach is akin to the relatively hedonistic way that pets and young children decide between options.

In recent research [6], we found that designing algorithms for learning from human reward is much easier with a myopic approach. We also investigated other projects with human reward and found that all five known projects were also myopic, though to a lesser extent than TAMER [1, 13, 12, 11, 9, 10].

A disadvantage of myopia is that it puts more of a burden on the trainer; the trainer must micromanage the agent's behavior. RL algorithms that seek long-term reward have opposite properties: they are easier on the trainer—when they work—but harder to design. The IUI paper by Knox and Stone [8] takes first steps toward non-myopic learning from human reward—to our knowledge, describing the first successful non-myopic approach to be published for *any task*. As it stands, TAMER is currently better at many complex tasks than the best non-myopic learning algorithms, and even as we move toward more farsighted learning, TAMER stands as the foundation on which we base our algorithms.

A DEMONSTRATION OF TAMER

We will demonstrate both TAMER and some non-myopic algorithms that build upon TAMER and are investigated in our IUI paper [8]. The demonstration will consist of simulated agents that can be trained interactively by conference attendees. We expect to demonstrate TAMER on Tetris and to teach interactive robot navigation tasks using a simulation of the robot Nexi. Attendees will use a presentation remote to deliver reward and punishment to the agents, which will be shown on a large display. Attendees are free to teach quite opposite behaviors. Nexi, for example, can be taught to follow a trainer's avatar or to avoid it. Video of training the real robot will also be shown. The non-myopic algorithms will be demonstrated on a simple grid-based navigation task.

SUMMARY AND VISION

This research constitutes a deep investigation into how agents should learn from human reward. TAMER is currently the only general framework prescribing how to learn exclusively from human reward. Whereas learning behavior from humans is currently dominated by demonstration-based approaches, TAMER helps establish a second major form of teaching, well grounded as a known teaching mechanism in humans and other animals.

We envision that the TAMER framework and related algorithms will empower human users to teach behaviors and improve their understanding of the agents through the interactivity of teaching. TAMER additionally is one of a number of approaches that can accelerate learning, which must be accomplished for assistive learning agents to be deployed widely.

REFERENCES

1. Isbell, C., Kearns, M., Singh, S., Shelton, C., Stone, P., and Kormann, D. Cobot in LambdaMOO: An Adaptive Social Statistics Agent. *Proceedings of The 5th Annual International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (2006).
2. Knox, W., Glass, B., Love, B., Maddox, W., and Stone, P. How humans teach agents: A new experimental perspective. *International Journal of Social Robotics, Special Issue on Robot Learning from Demonstration* (2012).
3. Knox, W., and Stone, P. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. *Proceedings of The 9th Annual International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (2010).
4. Knox, W. B. *Learning from Human-Generated Reward*. PhD thesis, Department of Computer Science, The University of Texas at Austin, August 2012.
5. Knox, W. B., and Stone, P. Interactively shaping agents via human reinforcement: The TAMER framework. In *The 5th International Conference on Knowledge Capture* (September 2009).
6. Knox, W. B., and Stone, P. Reinforcement learning from human reward: Discounting in episodic tasks. In *21st IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man)* (September 2012).
7. Knox, W. B., and Stone, P. Reinforcement learning with human and MDP reward. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (June 2012).
8. Knox, W. B., and Stone, P. Learning non-myopically from human-generated reward. In *International Conference on Intelligent User Interfaces (IUI)* (March 2013).
9. León, A., Morales, E., Altamirano, L., and Ruiz, J. Teaching a robot to perform task through imitation and on-line feedback. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (2011), 549–556.
10. Pilarski, P., Dawson, M., Degris, T., Fahimi, F., Carey, J., and Sutton, R. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *IEEE International Conference on Rehabilitation Robotics (ICORR)*, IEEE (2011), 1–7.
11. Suay, H., and Chernova, S. Effect of human guidance and state space size on interactive reinforcement learning. In *20th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man)* (2011), 1–6.
12. Tenorio-Gonzalez, A., Morales, E., and Villaseñor-Pineda, L. Dynamic reward shaping: training a robot by voice. *Advances in Artificial Intelligence–IBERAMIA* (2010), 483–492.
13. Thomaz, A., and Breazeal, C. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence 172*, 6-7 (2008), 716–737.