# A Large-Scale Study of Agents Learning from Human Reward

# (Extended Abstract)

Guangliang Li
University of Amsterdam
Amsterdam, the Netherlands
g.li@uva.nl

Hayley Hung
Delft University of Technology
Delft, the Netherlands
h.hung@tudelft.nl

Shimon Whiteson
University of Amsterdam
Amsterdam, the Netherlands
s.a.whiteson@uva.nl

## ABSTRACT

The *TAMER* framework, which provides a way for agents to learn to solve tasks using human-generated rewards, has been examined in several small-scale studies, each with a few dozen subjects. In this paper, we present the results of the first large-scale study of TAMER, which was performed at the NEMO science museum in Amsterdam and involved 561 subjects. Our results show for the first time that an agent using TAMER can successfully learn to play Infinite Mario, a challenging reinforcement-learning benchmark problem based on the popular video game, given feedback from both adult ($N = 209$) and child ($N = 352$) trainers. In addition, our study supports prior studies demonstrating the importance of bidirectional feedback and competitive elements in the training interface. Finally, our results also shed light on the potential for using trainers' facial expressions as a reward signal, as well as the role of age and gender in trainer behavior and agent performance.

## Categories and Subject Descriptors

I 2.6 [**Artificial Intelligence**]: Learning

## General Terms

Performance, Human Factors, Experimentation

## Keywords

Reinforcement learning, human-agent interaction

## 1. INTRODUCTION

As autonomous agents become more and more prevalent in our society, they have the potential to infiltrate all aspects of our daily lives, including healthcare, education, work, and leisure. The success of these autonomous agents will depend on their ability to efficiently learn from non-expert users in a natural way. Therefore, there is a great need for methods that facilitate the interaction between non-expert users and agents since it is through this interaction that such users transfer knowledge to such agents.

Inspired by human learning, researchers have developed many frameworks with which a human can assist an agent's learning. For example, giving the agent reward and punishment [4, 9], demonstration [1], guidance [10], advice [8], or even critiques of the agent's performance [2]. One of these approaches, called the *TAMER* framework [4], facilitates the agent learning from human-generated rewards that reflect the human trainer's judgement of the quality of the agent's actions. A TAMER agent learns from this feedback by creating a predictive model of the human trainer's feedback and myopically choosing the action at each time step that it predicts will receive the highest feedback value.

Moreover, previous work [5, 6] showed that the interaction between the agent and the trainer should ideally be bidirectional and that if an agent informs the trainer about the agent's past and current performance and its performance relative to others, the trainer will provide more feedback and the agent will ultimately perform better. However, due to the difficulty of recruiting subjects, these studies, like others [3] evaluating TAMER, were conducted using only 50-100 subjects. In this paper, we present the results of the first large-scale study of TAMER conducted at the NEMO science museum in Amsterdam with museum visitors as subjects. Doing so allowed us to recruit many more subjects ($N = 561$) and thus evaluate TAMER on a much larger scale than has previously been attempted. Our study provides large-scale support of our previous results demonstrating the importance of bidirectional feedback and competitive elements in the training interface but now in a new setting and also makes it feasible to investigate the potential of using facial expressions as reward signals and examine for the first time the role of both age and gender in the behavior and performance of trainers.

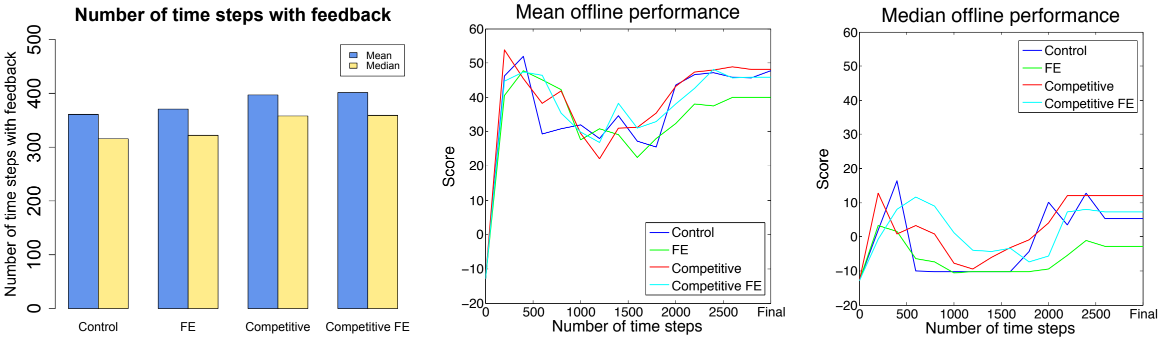
## 2. EXPERIMENTAL CONDITIONS

We have four conditions tested in our experiment: the *control condition* is the performance-informative interface replicated from [5] and implemented in the Infinite Mario domain, which shows the agent's past and current performance; the proposed *facial expression condition*, *competitive condition* and *competitive facial expression condition*.

Trainers in the control condition were told to use key presses to train the agent. They could give positive and negative feedback by pressing buttons on the keyboard to reward or punish the agent's previous action. The interface used in the facial expression condition is the same as in the control condition except that the trainers were told to use both key presses and facial expressions to train the agent.

In the interfaces used in the control and facial expression conditions, only the agent's own performance was shown to the human trainer. Previous work [6, 7] showed that putting people in a socio-competitive situation could further moti-

**Appears in:** *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*



a b c

**Figure 1: Number of time steps with feedback across the four conditions (a), mean (b) and median (c) final offline performance along the whole training process. (FE=Facial Expression.)**

vate them to give more feedback and improve the agent's performance. To verify this result as well as investigate how it plays out when the socio-competitive setting involves people who know each other and are training at the same time in the same room, we implemented the competitive condition that allows the agent to indicate the rank and score of the other members of the group, which typically consists of family members or close friends. The final condition is a combination of the facial expression and competitive conditions. Specifically, the interface is the same as in the competitive condition but, as in the facial expression condition, trainers were told to use both key presses and facial expressions to train the agent. In all four conditions, only the key presses were actually used for agent learning.

# 3. EXPERIMENTAL RESULTS

## 3.1 Feedback Given

As shown in Figure 1a, in terms of both mean and median, in the competitive condition, trainers gave more feedback than those in the control condition ($p = 0.05, r = 0.10$). Similarly, in the competitive facial expression condition, trainers gave more feedback than those in the facial expression condition ($p = 0.15, r = 0.06$). However, trainers in the facial expression condition and competitive facial expression condition gave similar amounts of feedback to those in the control and competitive conditions respectively. Thus, combined with the results of Mann Whitney U test between conditions, our results provide additional evidence that, even at a much broader age range (from 6 to 72) than was considered in [6], an agent's competitive feedback can increase the amount of feedback given by the trainers. Moreover, telling the subjects to use facial expressions as a separate channel for giving feedback for training agents does not decrease the amount of feedback given via key presses.

## 3.2 Performance

Our results show that an agent's competitive feedback can motivate human trainers to train agents better regardless of whether they are told to use facial expressions as a separate channel to train the agent or not. As shown in Figures 1b and 1c, agents in the competitive condition ultimately outperform those in the control condition, especially in terms of the median ($p = 0.2952, r = 0.04$). Similarly, agents in the competitive facial expression condition ultimately outperform those in the facial expression condition ($p = 0.2096, r = 0.05$). Moreover, agents in the facial expression condition and competitive facial expression condition ultimately perform worse than those in the control and competitive condition respectively ($p = 0.2072, r = 0.06$ and $p = 0.2827, r = 0.04$ respectively). Furthermore, these effects differ for females and males, children and adults.

# 4. ACKNOWLEDGMENTS

This research was part of Science Live, the innovative research programme of Science Center NEMO in Amsterdam that enables scientists to carry out real, publishable, peer-reviewed research using NEMO visitors as volunteers. Science Live is partially funded by KNAW and NWO. Guangliang Li is supported by the China Scholarship Council.

# REFERENCES

[1] P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. *ICML*, 2004.

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

[3] W. Knox, B. Glass, B. Love, W. Maddox, and P. Stone. How humans teach agents. *IJSR*, 2012.

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

[5] G. Li, H. Hung, S. Whiteson, and W. B. Knox. Using informative behavior to increase engagement in the tamer framework. *AAMAS*, 2013.

[6] G. Li, H. Hung, S. Whiteson, and W. B. Knox. Learning from human reward benefits from socio-competitive feedback. In *ICDL-EpiRob*, 2014.

[7] G. Li, H. Hung, S. Whiteson, and W. B. Knox. Leveraging social networks to motivate humans to train agents. In *AAMAS*, 2014.

[8] R. Maclin, J. Shavlik, L. Torrey, T. Walker, and E. Wild. Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression. In *National Conference on Artificial Intelligence*. AAAI Press, 2005.

[9] P. Pilarski, M. Dawson, T. Degris, F. Fahimi, J. Carey, and R. Sutton. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. *International Conference on Rehabilitation Robotics*, 2011.

[10] A. Thomaz and C. Breazeal. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. *Proc. of the National Conference on AI*, 2006.