

# Experiments in Socially Guided Machine Learning: Understanding How Humans Teach

Andrea L. Thomaz  
MIT Media Lab  
20 Ames St., E15-468  
Cambridge, MA 02139, USA  
alockerd@media.mit.edu

Guy Hoffman  
MIT Media Lab  
20 Ames St., E15-468  
Cambridge, MA 02139, USA  
guy@media.mit.edu

Cynthia Breazeal  
MIT Media Lab  
20 Ames St., E15-468  
Cambridge, MA 02139, USA  
cynthiab@media.mit.edu

## ABSTRACT

In Socially Guided Machine Learning we explore the ways in which machine learning can more fully take advantage of natural human interaction. In this work we are studying the role real-time human interaction plays in training assistive robots to perform new tasks. We describe an experimental platform, Sophie's World, and present descriptive analysis of human teaching behavior found in a user study. We report three important observations of how people administer reward and punishment to teach a simulated robot a new task through Reinforcement Learning. People adjust their behavior as they develop a model of the learner, they use the reward channel for guidance as well as feedback, and they may also use it as a motivational channel.

## Categories and Subject Descriptors

I.2 [Computing Methodologies]: Artificial Intelligence

## General Terms

Algorithms, Human Factors

## Keywords

Machine Learning, Socially Guided Agents, Human-Robot Interaction

## 1. INTRODUCTION

Various Machine Learning (ML) works have addressed some of the hard problems that robots face when learning in the real-world [3, 4]. However, learning from a human teacher poses additional challenges for machine learning systems (e.g., limited human patience, intuitive affordance of human input). We are developing Socially Guided Machine Learning (SG-ML), which assumes people will teach machines through a social and collaborative process and shall expect machines to engage in social forms of learning.

Our claim is that current examples of human interaction with machine learning fall short of the SG-ML goal. What is needed is a principled understanding of a non-expert human's contribution to the learning process. SG-ML draws many open questions concerning "how do humans want to teach?". The *timing* of an untrained human's feedback has received little attention. This is true as well for the *meaning* of human feedback. For instance, when does feedback pertain to a task versus a specific action, or a state versus an aspect of a state?

Our goal is to design learning algorithms that allow robots to learn flexibly on their own from exploration, but take full advantage of human guidance to improve their exploration and make learning more efficient. Thus, we argue that robots need to move flexibly along the guidance/exploration dimension, reframing the machine learning problem as a collaboration between the human and the machine.

This paper presents a framework for studying SG-ML, and reports results from a user study with the system. We present observations of the teaching strategies that the human instructors employed: in addition to administering traditional feedback, users want to *guide* the agent and give anticipatory rewards. We also show that users read the behavior of the learner and adjust training as their mental model of the learning agent changes. Finally, we find users may want a separate channel for motivational feedback.

## 2. THE SOPHIE'S WORLD PLATFORM

To investigate SG-ML, we have implemented a Java-based simulation platform, "Sophie's World". Sophie's World is a generic object-based State-Action MDP space for a single agent, Sophie, using a fixed set of actions on a fixed set of objects. The implementation details are beyond the scope of this paper. In the experiment described below we have used a cooking scenario of this MDP space. The kitchen task has on the order of 10,000 states with between 2 and 5 actions available in each state.

Sophie's World presents a human trainer with interactive reward interface. Using the mouse, a human trainer can—at any point—award a scalar reward signal  $r = [-1, 1]$ . Additionally, the interface lets the user make a distinction between rewarding the whole state of the world or the state of a particular object (object specific rewards); however, both types of rewards are treated the same by the algorithm. These rewards feed into a standard Q-Learning algorithm with learning rate  $\alpha = .3$  and discount factor  $\gamma = .75$ . (Note that we use Q-learning here as an instrument to investigate

how humans provide reward and punishment, and do not advocate it as the ultimate reinforcement-based learning algorithm of choice for SG-ML given its known limitations in real-world robotics domains.)

### 3. EXPERIMENT

We obtained 18 volunteers from the campus community. Participants were asked to play a video game, in which their goal was to get the robot agent, Sophie, to learn how to bake a cake on her own. The agent achieves this goal through a series of actions (e.g. pick-up eggs, turn left, use-eggs on bowl, etc.). They got to decide when they were finished training Sophie. At this point the experimenter tested the agent and their game score was the degree to which Sophie finished baking the cake by herself. Participants received between \$5 and \$10 based on their game score.

#### 3.1 Results

Of the 18 participants, 13 successfully completed the task. Though the game is fairly simple, people had varied experiences. Total time spent with the game was varied (mean: 30.8 minutes; st. dev.: 16.7 minutes), as was the number of goals seen before declaring teaching done (mean: 3.9; st. dev.: 1.75). Despite participants' varied experiences, a few overarching similarities arise about their training behavior.

##### 3.1.1 Guidance versus Feedback

Even though the instructions clearly stated that communication and rewards were *feedback* messages, we saw that many people assumed the object specific rewards were future directed messages or guidance for the agent. This has been derived from both interviews and from the correlation of object/action pertinence and reward giving. Thus, people were giving rewards for actions the agent was *about to do* in addition to the traditional rewards for what the agent had just done. While delayed rewards have been discussed in the Reinforcement Learning (RL) literature [2], these *anticipatory* rewards observed from everyday human trainers will require new tools and attention in learning algorithms.

##### 3.1.2 Shifting Mental Models

We found two illustrations of the human trainer's propensity to adjust their behavior to the learner as they formed and revised their mental model of how the agent learns.

We expected that feedback would decrease over the training session, informed by related work in which Isbell et al. [1] observed habituation in an interactive teaching task. We found just the opposite: the ratio of rewards to actions over the entire training session had a mean of .77 and standard deviation of .18 and, we see an increasing trend in the rewards-to-actions ratio over the first three quarters of training. One explanation for an increasing trend is a shift in mental model; as people realize the impact of their feedback they adjusted to fit this model of the learner. This explanation finds anecdotal support in the interview responses. Many users reported that at some point they came to the conclusion that their feedback was helping the learning process and they subsequently gave more rewards.

A second expectation we had was that a human coach would naturally use goal-oriented and intentional communication. In most MDP scenarios a reward pertains to a

complete state, but in an SG-ML reward there is likely a particular aspect of the state being rewarded. We tried to measure this with the object specific rewards. In looking at the difference between the first and last quarters of training, we see that many people tried the object specific rewards at first but stopped using them over time. In the interview, many users reported that the object rewards "did not seem to be working." Thus, many participants tried the object specific rewards initially, but were able to detect over time that an object specific reward did not have a different effect on the learning process than a general reward (which is true), and therefore stopped using the object rewards.

##### 3.1.3 Positive Bias in Rewards

For many people, a large majority of rewards given were positive. A plausible hypothesis is that people are falling into a natural teaching interaction with the agent, treating it as a social entity that needs motivation and encouragement. Some people specifically mentioned in the interview that they felt positive feedback would be better for learning. This might indicate the need for a dedicated motivational channel in SG-ML systems. An alternative hypothesis is that negative rewards did not result in the appropriate reaction on the agent's part (such as an UNDO behavior).

### 4. CONCLUSION

Our SG-ML approach is founded on the premise that people will naturally want to teach machines through a social and collaborative process. This study presents empirical evidence in support of this assertion. For instance, even with a single communication channel, people used it to guide and motivate the robot. In addition, people's strategy for administering reward and punishment changed over time based on the robot's behavior.

These findings raise important issues for the design of reinforcement-based learning algorithms to better take advantage of the human teacher: 1) How to incorporate human guidance to improve the learner's exploration; 2) How might the teacher's encouragement be used to improve learning performance; and 3) How can the overt (expressive) behavior of the learner be used to help people dynamically adapt their teaching strategies to be more appropriate over time. As future work, we are modifying our reinforcement-based learning algorithm to incorporate these findings and testing their impact both on teaching behavior and the robot's learning performance in follow-up user studies.

### 5. REFERENCES

- [1] C. Isbell, C. Shelton, M. Kearns, S. Singh, and P. Stone. Cobot: A social reinforcement learning agent. *5th Intern. Conf. on Autonomous Agents*, 2001.
- [2] L. P. Kaelbling, M. L. Littman, and A. P. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.
- [3] M. Mataric. Reinforcement learning in the multi-robot domain. *Autonomous Robots*, 4(1):73–83, 1997.
- [4] S. Thrun. Robotics. In S. Russell and P. Norvig, editors, *Artificial Intelligence: A Modern Approach (2nd edition)*. Prentice Hall, 2002.