Socially Guided Machine Learning: Designing an Algorithm to Learn from Real-Time Human Interaction

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

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

Editor: Leslie Pack Kaelbling

Abstract

Socially Guided Machine Learning explores the ways in which machine learning can be designed to more fully take advantage of natural human interaction and tutelage. In this article we present a framework for studying the role real-time human interaction plays in training robots to perform new tasks. We have results from an initial user study using our experimental platform, Sophie's World, to understand how people administer reward and punishment to teach a simulated robot a new task through Reinforcement Learning (RL). Based on this study, we identified three modifications to a standard RL algorithm to make it more amenable to learning from real-time human interaction: an embellished communication channel with both guidance and feedback, transparency behaviors, and responsiveness to errors. We are evaluating these modifications in a follow-up study.

1. Introduction

Machine learning will play a significant role in the development of robotic assistants for human environments such as homes, schools, hospitals, and offices. It will be impossible to give machines all of the knowledge a priori that they will need to serve useful long term roles in our dynamic world; thus, the ability for non-experts to guide them easily will be key to their success. Recognizing that current machine learning techniques have met with great success in many applications, and various works have addressed some hard problems that robots face when learning in the real-world (Mataric (1997); Thrun and Mitchell (1993)), social learning in a human environment poses additional challenges for learning systems.

Socially Guided Machine Learning (SG-ML) takes the following position: people will teach machines through a social and collaborative process and shall expect machines to engage in social forms of learning. Robots should fully participate in the teaching/learning partnership, and the ability to utilize and leverage social skills is more than a good human interface, it can positively impact the underlying mechanisms to improve learning performance in real-time interactive training sessions.

In this article we describe a video game platform we used to look at how real people administer reward and punishment to a reinforcement learning agent. Based on this study, we identified three modifications to a standard reinforcement learning algorithm to make it more amenable to learning within the framework of a real-time human interaction: an embellished communication channel with both guidance and feedback, transparency behaviors, and responsiveness to errors. We are in the process of evaluating these modifications in a follow-up study.

2. Approach: A Human Interaction Perspective for Machine Learning

Our perspective reframes the machine learning problem as an interaction between the human and the machine. While significant attention has been paid to designing machine learning techniques that perform well on their own, less attention has been paid to designing for the performance of the complete human-machine teaching/learning system. It is useful to think of related works along the spectrum of "human interaction with a learning process". One extreme is a black box learning algorithm with inputs and outputs and little concern that a human may be involved in presenting the training data. The Machine Learning community has done much work in this scenario.

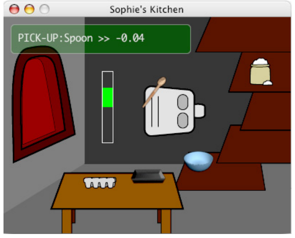
2.1 Designing for Human Input

Further along the spectrum, several examples exist in which a machine learning algorithm is designed to accept input explicitly from a human, often an expert. Natural language has been utilized to frame learning episodes in robot task learning (Nicolescu and Matarić (2003)), classification tasks (Steels and Kaplan (2001)), and navigation tasks (Lauria et al. (2002)). There are several examples of a human providing a reward signal for both robotic and software agent reinforcement learners (Kaplan et al. (2002); Isbell et al. (2001); Kuhlmann et al. (2004)). Several related works exist in which robotic and software agents learn by observing a human (Schaal (1999); Voyles and Khosla (1998); Lieberman (2001)). Many of these approaches constrain the teacher to a special interaction or language. An important question for SG-ML is "how do humans want to teach?".

2.2 Designing for Human Partnership

A social learning interaction is inherently a partnership, and both learner and teacher influence the performance of the tutorial dyad. While this observation is straightforward, the communication between human teacher and artificial agent has received little attention in traditional approaches. A third point along the spectrum is a scenario where the human provides input to the machine learning process and the process also provides feedback to the human; the learning experience is a tightly coupled partnership. Active learning is an approach that is in the spirit of this partnership scenario (Cohn et al. (1995); Schohn and Cohn (2000)), but the human's role is often constrained in terms of both input and output.

An important component of our approach is *transparency*, or ways in which the learner can communicate the state of its internal process. In other work, we have demonstrated aspects of collaboration and social learning on a humanoid robot, using social cues and gestures to achieve transparency and guide instruction (Breazeal et al. (2004a,b)). We look



PICK-UP:Spoon >> -0.04

Figure 1: Sophie's Kitchen. The color bar is the interactive reward controlled by the human.

towards early child development literature to inform our choice of agent-to-teacher signals, and stress the use of *gesture*, *gaze*, and *hesitation* as intuitive signals for an SG-ML agent (Kaye (1977); Argyle et al. (1973)).

3. 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. Sophie's World is also a web-based application, enabling the collection of a large amount of naïve-user data.

3.1 Sophie's MDP

A World $W = \langle L, O, \Sigma, T \rangle$ is a finite set of k locations $L = \{l_1, \ldots, l_k\}$ and n objects $O = \{o_1, \ldots, o_n\}$. Each object can be in one of an object-specific number of mutually exclusive states. We denote the set of states object o_i can be in as Ω_i. The complete object space can be described as $O^* = \langle \Omega_1 \times \ldots \times \Omega_n \rangle$. W is also defined by a set of legal states $\Sigma \subset \langle L \times L^O \times O^* \rangle$. Thus, a world state $s(l_a, l_{o_1} \ldots l_{o_n}, \omega)$ consists of the agent location, the location of each objects, and the object configuration in $\omega \in O^*$. Finally, W has a transition function $T : \Sigma \times A \mapsto \Sigma$, where A is fixed (see below).

The action space A has four atomic actions with arguments as follows: Assuming the locations L are arranged in a ring, at any time step, the agent can `GO left` or `right`; she can `PICK-UP` any object that is in her current location; she can `PUT-DOWN` any object currently in her possession; and she can `USE` any object in her possession on any object in her current location. Each action implements a transition function in T that advances the world state.

On top of this generic architecture we build task-specific implementations, which determine: the the spatial relationship between the locations, a limit on the number of objects the agent can possess, and the sub-transition $T_U \subset T$ accomplished by the `USE` action.

3.2 Interactive Rewards Interface

A central feature of Sophie's World is the interactive reward interface. Using the mouse, a human trainer can, at any point, award a scalar reward signal $r = [-1, 1]$. The user receives visual feedback enabling them to tune the reward signal value before sending it. The reward interface runs asynchronously, and thus does not halt the progress of the agent.

The version of Sophie's World for the initial study also let the user make a distinction between rewarding the whole state of the world or the state of a particular object (object specific rewards). An object specific reward is administered by clicking the mouse button down on a specific object. For visual feedback, the object is highlighted when moused over to indicate that any subsequent reward will be object specific. In the experiment, object specific rewards are used to learn about the human trainer's behavior and communicative intent; however, the learning algorithm treats all rewards equally and in the traditional sense of pertaining to the whole state.

3.3 Learning Algorithm

We believe the reinforcement learning paradigm lends itself naturally to human interaction, though there are few examples of interactive RL where a human issues reward/punishment in real-time. For the experiment presented below we implemented a standard Q-Learning algorithm (learning rate $\alpha = .3$ and discount factor $\gamma = .75$) (Sutton and Barto (1998)).

3.4 The Kitchen World

The task scenario used in this study is a Kitchen world (see Fig. 1). In it the agent is to learn to prepare batter for a cake and put the batter in the oven.

The object space has five objects: `Flour`, `Eggs`, a `Spoon` (each has one object state), a `Bowl` (with five object states: `empty`, `flour`, `eggs`, `unstirred`, `stirred`), and a `Tray` (with three object states: `empty`, `batter`, `baked`). The world has four locations: `Shelf`, `Table`, `Oven`, `Agent` (i.e., the agent in the center surrounded by a shelf, table and oven).

In this kitchen implementation, the agent can hold only one object at a time. The initial state, S_0, has all objects on the `Shelf`, and the agent faces the `Shelf`. A successful completion of the task will include putting flour and eggs in the bowl, stirring the ingredients using the spoon, then transferring the batter into the tray, and finally putting the tray in the oven. In order to encourage an efficient policy, an inherent negative reward signal of $\rho = -.04$ is placed in any non-goal state. Some states are so-called *disaster* states (e.g., putting the eggs in the oven), which result in a negative reward ($r = -1$), the termination of the current trial, and a transition to state S_0.

4. Experiment

4.1 Design

We had 18 participants play a video game, with the goal of getting Sophie to learn how to bake a cake on her own. They played the game as long as they felt necessary and then the experimenter tested the agent and their score was the degree to which Sophie finished baking the cake by herself. Participants received between $5 and $10 based on their score.

Participants were told they could not tell Sophie what actions to do, and could not do any actions directly. They were only able to give Sophie feedback messages with the mouse, explained as:

- Drag the mouse UP to make the box more GREEN, a POSITIVE message. Drag DOWN for RED/NEGATIVE. Lift the mouse button to send the message to Sohpie, she sees the color and size.
- If you click the mouse button down on an object, this tells Sophie your message is about that object. As in, “Hey Sophie, this is what I’m talking about...”. If you click anywhere else, Sophie assumes your feedback pertains to everything in general.

The system maintains an activity log with time step and real time of: state transitions, actions, human rewards, reward aboutness (if object specific), disasters, and goals. Participants completed a questionnaire and an informal interview after the game.

4.2 Results

4.2.1 Shifting Mental Models

We found evidence, in two behavioral examples, suggesting users adjust their behavior to the learner as they form and revise their mental model of the agent and how it learns.

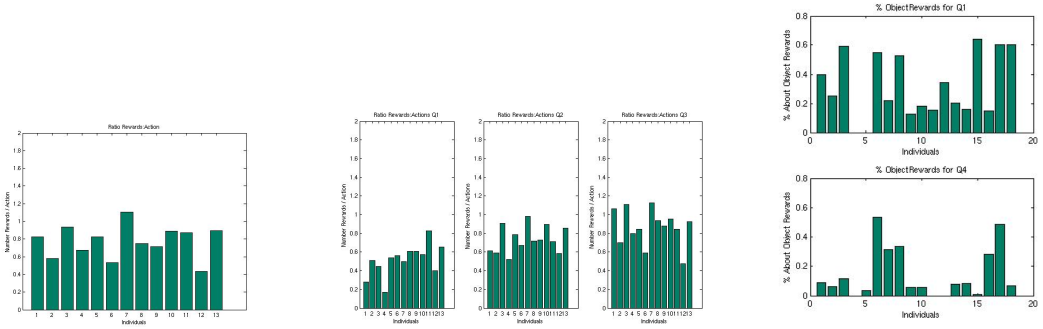
We hypothesized feedback would decrease over the training session (in related work Isbell et al. (2001) observed habituation in an interactive teaching task), but we found just the opposite. The ratio of rewards to actions over the entire training session had a mean of .77 and standard deviation of .18 (Fig. 2(a)), and, dividing individual sessions into four time quarters, we see an increasing trend in the ratio in the first three quarters (Fig. 2(b)). Based on the interview data, we believe this is a shift in mental model; as people realized the impact of their feedback they adjusted to fit this model of the learner. Many users said they gave more rewards once they concluded that their feedback was helping the learning process, adjusting their training behavior based on the learner’s behavior.

The second case concerns the object specific rewards, through which we hoped to measure whether people communicated rewards at different levels. Their usage varied greatly, and looking at the difference in the first and last quarters of training, we see many people tried the object specific rewards but stopped using them over time (Fig. 2(c)). Additionally, several users reported that the object rewards “did not seem to be working.” Thus, many participants were able to detect that an object specific reward did not have a different effect on the process than a general reward (see Sec. 3.2). Once they made this realization, they revised their mental model of the learner and stopped using the object rewards.

4.2.2 Guidance versus Feedback

A second finding pertains to expectations about the object specific rewards. Even with instructions stating that rewards were *feedback* messages, many people assumed the object specific rewards were guidance messages for the agent. When asked about their use of the object specific rewards, people had one of three responses. 1) They used the object reward to indicate desired/undesired object of attention or next object to use. 2) They tried to use the object reward in this way, but it “didn’t seem to be working.” 3) They did not understand what object rewards would mean.

Behavioral data quantifies this reported guidance behavior. Used in the traditional RL sense, object rewards should pertain to the last object the agent used. Many object rewards were highly correlated with the object used in the last action; however, a substantial number



(a) Ratio rewards to actions for entire training. Mean: .77; St. dev.: .18

(b) Ratio of rewards to actions over the first three quarters of the training sessions shows an increasing trend.

(c) % object specific in the first (top) and last (bottom) quarters.

Figure 2: Changes in the ratio of rewards to actions and use object specific rewards shows people's behavior adjustment as they developed a mental model of the learner.

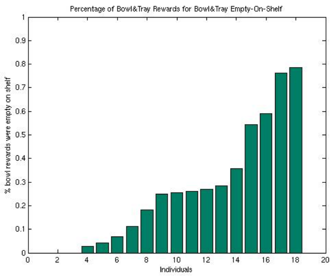


Figure 3: Percentage of bowl and tray rewards given when the bowl or tray was empty on the shelf. Assumed to be a guidance reward rather than a feedback reward.

of people had object rewards that were rarely correlated to the last object of attention. We hypothesized that rather than the last object of attention, some rewards may pertain to the future, what they want the agent to use next. To test roughly how many people used object rewards as a guidance mechanism, we consider one example case. A guidance reward could be used in a state where the agent is facing the shelf, (about which object to pick up). Further, a reward to the empty bowl or tray on the shelf can only be guidance since this cannot be part of a desired sequence (rewards to other objects on the shelf could be part of the sequence, some people had the robot put objects away after use).

We look at the bowl and tray rewards to learn about the guidance behavior. Fig. 3 shows the percentage of bowl and tray rewards that were given when they were empty on the shelf, for each individual. Only three people never gave a reward to the bowl or tray sitting empty on the shelf, and over half of the participants gave a substantial percentage of such rewards. Thus, many participants tried using rewards to guide the agent's behavior.

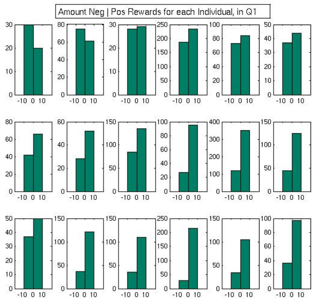


Figure 4: Reward histograms for each indiv. in first quarter, left=negative right=positive.

4.2.3 Positive Bias in Rewards

Another finding regarding the meaning of human rewards concerns the valence. Many people gave a majority of positive rewards. Even in the first quarter of training, well before the agent is behaving correctly, the majority of people show a positive bias (see Fig. 4).

One 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. People may feel bad giving negative rewards, or feel that it is important to be both instrumental and motivational with their communication. In the interviews a few participants did mentioned that they believed the agent would learn better from positive feedback.

Another hypothesis is that negative rewards did not have feedback for the user. A reinforcement learning agent does not have an instantaneous reaction to either positive or negative rewards, but particularly in the case of negative rewards, this could be interpreted as the agent 'ignoring' the feedback. In a similar fashion as the object specific rewards, the user may stop using them when they feel they are not being heard.

5. A Modified Reinforcement Learning Algorithm

These findings suggest specific recommendations for machine learning. The communication from the human teacher cannot be a single reward signal; we need to account for the various intentions people wish to convey to the machine. Additionally, people tune their behavior to match the needs of the machine, and this process can be augmented with more transparency of the internal state of the learner. We have made extensions to the system used in the initial experiment, and we are in the process of running a second set of studies with the Sophie video game to understand the effects of these changes on the machine learning process.

5.1 An embellished communication channel: Reward, Guidance, & Motivation

A reinforcement learning agent typically has feedback from the environment through a single reward signal. However, when this signal is administered by a human teacher, our study shows their communicative intentions go beyond simple feedback. Our initial findings suggest that the teacher be given three communication channels: 1) the standard feedback signal, 2) an attention direction or guidance signal, 3) a motivation signal.

While delayed rewards have been discussed in the Reinforcement Learning community, the concept of rewarding the *action the agent is about to do* is novel and requires new tools and attention. In our approach, the learning agent uses the reward signal in the standard way to update an action policy. However, with the guidance signal the human teacher is meaning to communicate with the action selection phase of the algorithm. Thus, we have modified the algorithm to let the human input distinctly affect both the policy update and the action selection phases with these two different channels of communication. The motivation channel has been added to test whether or not users desire a high level feedback signal (a hypothesis the initial study suggests), but the learning algorithm does not yet distinguish between this and a regular reward.

A standard reinforcement learning algorithm (e.g., Q-Learning) can be described as: `select-action`, `take-action`, `sense-world-reward`, `update-policy`. In our modified Q-Learning algorithm we have a pre-action and post-action phase. In the pre-action phase the agent registers guidance communication to bias action selection, and in the post-action phase the agent uses the reward channel in the standard way to evaluate that action and update a policy. With our modification the learning process becomes: `sense-world-guidance`, `select-action`, `take-action`, `sense-world-reward`, `update-policy`.

We implemented this in the Sophie's World framework. The human teacher communicates guidance with the mouse. With a right-click, a yellow square appears. Users are instructed that when there is an object enclosed in the yellow square this directs Sophie's attention to that object. The agent waits for guidance messages during the `sense-world-guidance` step (we introduce a short delay to allow the teacher time to administer guidance). A guidance message is in the form `<guidance><object>`; the agent saves this `object` as the `guidance-object-of-attention`.

During action selection, the default behavior (a standard approach) chooses randomly between the set of actions with the highest Q-values, within a bound ϵ. If any guidance messages were received, the agent will *instead* choose randomly between the set of actions that have the `guidance-object-of-attention` as their object of attention.

5.2 Soliciting Guidance through Transparency

In prior work, we have stressed that teaching and learning should be characterized as a *collaboration* (Breazeal et al. (2004a)). Teachers direct a learner's attention, structure experiences, support attempts, and regulate complexity. The learner contributes by revealing their internal state to help guide the teaching process. Each simplifies the task for each other, the adult is a more effective teacher and the child a more effective learner. The findings in this study support this notion of *partnership.* When everyday users are asked to interactively train a machine learning agent, we see them adjust their training behavior as the interaction proceeds, reacting to the behavior of the learner. This presents a huge opportunity for an interactive learning agent to *improve its own learning environment* by communicating more of its internal state to the human teacher.

The first element of transparency we are exploring is gaze, which requires that the learning agent have a physical embodiment (either real or virtual) that can be understood by the human as having a forward heading. Gaze precedes an action and communicates something about the action that is going to follow.

We have made the following extentions to Sophie's World to add gaze. During the sense-world-guidance phase, the learning agent finds the set of actions, A, with the highest Q-values, within a bound ϵ. For every action, a, in A, the learning agent gazes for 1 second at the object-of-attention of a (if any). This gazing behavior during the pre-action phase communicates a level of uncertainty through the amount of gazing that precedes an action. It introduces a delay (proportional to uncertainty) prior to select-action, soliciting and providing the opportunity for guidance messages. This also communicates overall task certainty as the agent will speed up when every set A has a single action. We expect this transparency to improve the teacher's model of the learner, creating a more understandable interaction for the human and a better learning environment for the machine.

5.3 Just-In-Time Error Correction

A final modification to the algorithm addresses how the learning agent should deal with negative feedback when it is being administered by a human teacher. In the standard Q-Learning framework, the effects of a reward on the action selection mechanism are not seen or manifested until that particular state from which the action was made is revisited (which can take a very long time). In many cases this is a benefit because being too responsive to any one reward would be detrimental to the exploration needed to learn. However, we argue that in the case where this reward signal is coming from a benevolent human teacher, the learning agent can and should be more responsive to negative feedback. Moreover, we expect that just-in-time error correction more closely resembles a natural human teaching interaction and will thus be more understandable for the human partner.

We have modified the Sophie agent to respond to negative feedback with an UNDO behavior (a natural correlate or opposite action) when possible. Thus a negative reward is handled in both the update-policy step and the subsequent select-action step. In the select-action step immediately following negative feedback, the action selection mechanism chooses the action that 'un-does' the last action if possible. The action selection mechanism needs only call an undo function on the last action, and the proper UNDO behavior is represented within each primitive action (e.g. the action GO <direction> returns GO <-direction>; PICK-UP <object> returns PUT-DOWN <object>; etc.).

6. Conclusions

The introduction of a human real-time reward signal brings about a range of new considerations for machine learning. We have presented our simulation framework, Sophie's World, used to study the impact of human interaction on a machine learning process. Our experiment with the Sophie's Kitchen video game indicates that people can and will adjust their training behavior to best fit the behavior of the learning agent, and people show various communicative intents in their rewarding behavior beyond feedback in the traditional sense.

Based on these findings, we have made modifications to a reinforcement learning algorithm that we hypothesize will be more successful in a real-time learning interaction with human partners. Our goal is Socially Guided Machine Learning systems that are specifically designed with human feedback and social guidance in mind, and we believe that this design methodology can benefit both the human teacher and the machine learner.

References

M. Argyle, R. Ingham, and M. McCallin. The different functions of gaze. *Semiotica*, pages 19–32, 1973.

C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A. Lockerd, and D. Mulanda. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1 (2), 2004a.

C. Breazeal, G. Hoffman, and A. Lockerd. Teaching and working with robots as collaboration. In *Proceedings of the AAMAS*, 2004b.

D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and J. Alspector, editors, *Advances in Neural Information Processing*, volume 7. Morgan Kaufmann, 1995.

C. Isbell, C. Shelton, M. Kearns, S. Singh, and P. Stone. Cobot: A social reinforcement learning agent. *5th Intern. Conf. on Autonomous Agents*, 2001.

F. Kaplan, P-Y. Oudeyer, E. Kubinyi, and A. Miklosi. Robotic clicker training. *Robotics and Autonomous Systems*, 38(3-4):197–206, 2002.

K. Kaye. Infants effects upon their mothers teaching strategies. In J. Glidewell, editor, *The Social Context of Learning and Development*. Gardner Press, New York, 1977.

G. Kuhlmann, P. Stone, R. J. Mooney, and J. W. Shavlik. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *Proceedings of the AAAI-2004 Workshop on Supervisory Control of Learning and Adaptive Systems*, San Jose, CA, July 2004.

S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein. Mobile robot programming using natural language. *Robotics and Autonomous Systems*, 38(3-4):171–181, 2002.

H. Lieberman, editor. *Your Wish is My Command: Programming by Example*. Morgan Kaufmann, San Francisco, 2001.

M. Mataric. Reinforcement learning in the multi-robot domain. *Autonomous Robots*, 4(1):73–83, 1997.

M. N. Nicolescu and M. J. Matarić. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proceedings of the 2nd Intl. Conf. AAMAS*, Melbourne, Australia, July 2003.

S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3: 233242, 1999.

G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *Proc. 17th ICML*, pages 839–846. Morgan Kaufmann, San Francisco, CA, 2000.

L. Steels and F. Kaplan. Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3–32, 2001.

R. S. Sutton and A. G. Barto. In *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 1998.

S. B. Thrun and T. M. Mitchell. Lifelong robot learning. Technical Report IAI-TR-93-7, 1, 1993.

R. Voyles and P. Khosla. A multi-agent system for programming robotic agents by human demonstration. In *Proceedings of AI and Manufacturing Research Planning Workshop*, 1998.