Available online at www.sciencedirect.com
ScienceDirect
Artificial Intelligence 172 (2008) 716–737
Artificial Intelligence
www.elsevier.com/locate/artint

Teachable robots: Understanding human teaching behavior to build more effective robot learners ☆

Andrea L. Thomaz [a,*,1], Cynthia Breazeal [b,2]

[a] *Interactive Computing, Georgia Institute of Technology, USA*
[b] *MIT Media Laboratory, USA*

Received 9 August 2006; received in revised form 25 September 2007; accepted 27 September 2007

Available online 20 December 2007

Abstract

While Reinforcement Learning (RL) is not traditionally designed for interactive supervisory input from a human teacher, several works in both robot and software agents have adapted it for human input by letting a human trainer control the reward signal. In this work, we experimentally examine the assumption underlying these works, namely that the human-given reward is compatible with the traditional RL reward signal. We describe an experimental platform with a simulated RL robot and present an analysis of real-time human teaching behavior found in a study in which untrained subjects taught the robot to perform a new task. We report three main observations on how people administer feedback when teaching a Reinforcement Learning agent: (a) they use the reward channel not only for feedback, but also for future-directed guidance; (b) they have a positive bias to their feedback, possibly using the signal as a motivational channel; and (c) they change their behavior as they develop a mental model of the robotic learner. Given this, we made specific modifications to the simulated RL robot, and analyzed and evaluated its learning behavior in four follow-up experiments with human trainers. We report significant improvements on several learning measures. This work demonstrates the importance of understanding the human-teacher/robot-learner partnership in order to design algorithms that support how people want to teach and simultaneously improve the robot's learning behavior.

Keywords: Human–robot interaction; Reinforcement learning; User studies

1. Introduction

As robots enter the human environment to assist people in their daily lives, the ability for ordinary people to easily teach them new tasks will be key to their success. Various works have addressed some of the hard problems robots face when learning in the real-world, e.g., real-time learning in environments that are partially observable, dynamic, continuous [22,33,34]. However, learning quickly from interactions with a human teacher poses additional challenges (e.g., limited human patience, ambiguous human input) as well as opportunities for Machine Learning systems.

☆ This work was funded by the MIT Media Laboratory.
* Corresponding author.
E-mail addresses: athomaz@cc.gatech.edu (A.L. Thomaz), cynthiab@media.mit.edu (C. Breazeal).
1 Assistant Professor.
2 Associate Professor.

0004-3702/$ – see front matter
doi:10.1016/j.artint.2007.09.009

Several examples of agents learning interactively with a human teacher are based on Reinforcement Learning (RL). Many question RL as a viable technique for complex real-world environments due to practical problems; but it has certain desirable qualities, like exploring and learning from experience, prompting its use for robots and game characters. A popular approach incorporates real-time human feedback by having a person supply reward and/or punishment as an additional input to the reward function [5,10,13,14,31].

Most of this work models the human input as indistinguishable from any other feedback in the environment, and implicitly assumes people will correctly communicate feedback as expected by the algorithm. We question these assumptions and argue that reinforcement-based learning approaches should be reformulated to more effectively incorporate a human teacher. To address this, we advocate an approach that integrates Machine Learning into a Human–Robot Interaction (HRI) framework. Our first goal is to understand the nature of a teacher's input. We want to understand *how people want to teach* and *what they try to communicate to the robot learner*. Our second goal is to incorporate these insights into standard Machine Learning techniques, to adequately support a human teacher's contribution in guiding a robot's learning behavior.

This paper presents a series of five experiments analyzing the scenario of a human teaching a virtual robot to perform a novel task within a reinforcement-based learning framework. Our experimental system, *Sophie's Kitchen*, is a computer game that allows a Q-Learning agent to be trained interactively.[3]

In the first experiment (Section 5) we study 18 people's interactions with the agent and present an analysis of their teaching behavior. We found several prominent characteristics for how people approach the task of explicitly teaching a RL agent with direct control of the reward signal. To our knowledge, this work is the first to explicitly address and report such results, which are relevant to any interactive learning algorithm:

- People want to direct the agent's attention to guide the exploration process.
- People have a positive bias in their rewarding behavior, suggesting both instrumental and motivational intents with their communication channel.
- People adapt their teaching strategy as they develop a mental model of how the agent learns.

The second contribution of this work is to incorporate these findings into specific modifications of the agent's graphical interface and RL algorithm. We had over 200 people play the game in four follow-up experiments, showing that these modifications significantly improve the learning behavior of the agent and make the agent's exploratory behavior more appropriately responsive to the human's instruction.

- *Leveraging human guidance*: In the second experiment (Section 8), we show the positive effects of adding a guidance channel of communication. Human players are able to direct the agents attention to yield a faster and more efficient learning process.
- *Transparency to guide a human teacher*: In the third experiment (Section 9), we show that transparency behaviors, such as gaze, that reveal the internal state of the agent can be utilized to improve the human's input.
- *The Asymmetry of human feedback*: In the fourth and fifth experiments (Section 10), we show beneficial asymmetric interpretations of feedback from a human partner. The fourth experiment shows that giving human players a separate motivational communication channel decreases the positive rewards bias. The fifth experiment shows the benefits of treating negative feedback from the human as both feedback for the last action and a suggestion to reverse the action if possible.

This work contributes to the design of real-time learning agents that are better matched to human teaching. These experiments lay the foundation for designing robots that both learn more effectively and are easier to teach. We demonstrate that an understanding of the coupled human-teacher/robot-learner system allows for the design of algorithms that support how people want to teach and simultaneously improve the machine's ability to learn.

[3] Q-Learning is used in this work because it is a widely understood RL algorithm, affording the transfer of these lessons to other reinforcement-based approaches.

2. Background: Related works in human-trainable systems

A review of related works in Machine Learning yields several dimensions upon which human-trainable systems can be characterized. One is implicit verses explicit training. For instance, personalization agents and adaptive user interfaces rely on the human as an implicit teacher, modeling human preferences or activities through passive observation of the user's behavior [12,17,23]. In contrast, our work addresses explicit training, where the human teaches the learner through interaction.

In systems that learn via interaction, another salient dimension is whether the human or the machine leads the interaction. Active learning or learning with queries is an approach that explicitly acknowledges an interactive supervisor [9,27]. Through queries, the algorithm controls the interaction without regard for what a human could provide in a real scenario. Alternatively, our work addresses the human-side of the interaction and specifically asks *how do humans want to teach machines?*

A third interesting dimension is the balance between relying on guidance verses exploration to learn new tasks. Several works have focused on how a machine can learn from human instruction, and a number of these rely heavily on a human guidance. The learning problem is essentially reduced to programming through natural interfaces—with little if any exploration on the part of the machine. For example: learning by demonstration [24,36], learning by imitation [26], programming by example [19], learning via tutelage [20], programming by natural language [18]. This results in a dependence on having a human present to learn, but allows the human complete control over the learning process.

On the other hand, there have been several reinforcement-based approaches positioned strongly along the exploration dimension. For example, several works allow the human to contribute to the reward function [5,10,13,14,31]. An exploration approach has the benefit that the human need not know exactly how the agent should perform the task, and learning does not require their undivided attention.

Our long-term goal is to create learning systems that can dynamically slide along this exploration-guidance spectrum, to leverage a human teacher when present as well as learn effectively on its own. While there are known practical issues with RL (training time requirements, representations of state and hidden state, practical and safe exploration strategies), we believe that an appropriate reformulation of RL-based approaches to include input from a human teacher could alleviate these shortcomings. To do this properly, we must first understand the human teacher as a unique contribution that is distinct from other forms of feedback coming from the environment.

3. Approach: A HRI framework for machine learning

Our approach is based on a *Social Learner Hypothesis*, namely that humans will naturally want to teach robots as social learners. As such, our work draws inspiration from Situated Learning Theory—a field of study that looks at the social world of children and how it contributes to their development. A key concept is *scaffolding*, where a teacher provides support such that a learner can achieve something they would not be able to accomplish independently [11,16].

In a situated learning interaction, the teaching and learning processes are intimately coupled. A good instructor maintains a mental model of the learner's state (e.g., what is understood, what remains confusing or unknown, etc.) in order to appropriately support the learner's needs. Attention direction is one of the essential mechanisms that contribute to structuring the learning process [38]. Other scaffolding acts include providing feedback, structuring successive experiences, regulating the complexity of information, and otherwise guiding the learner's exploration. This scaffolding is a complex process where the teacher dynamically adjusts their support based on the learner's demonstrated skill level and success.

The learner, in turn, helps the instructor by making their learning process *transparent* to the teacher through communicative acts (such as facial expressions, gestures, gaze, or vocalizations that reveal understanding, confusion, attention), and by demonstrating their current knowledge and mastery of the task [1,15]. Through this reciprocal and tightly coupled interaction, the learner and instructor cooperate to simplify the task for each other—making each a more effective partner.

This situated learning process stands in dramatic contrast to typical Machine Learning scenarios, that have traditionally ignored "teachability issues" such as how to make the teaching-learning process interactive and intuitive for a non-expert human partner. We advocate a new perspective that reframes the Machine Learning problem as an inter-

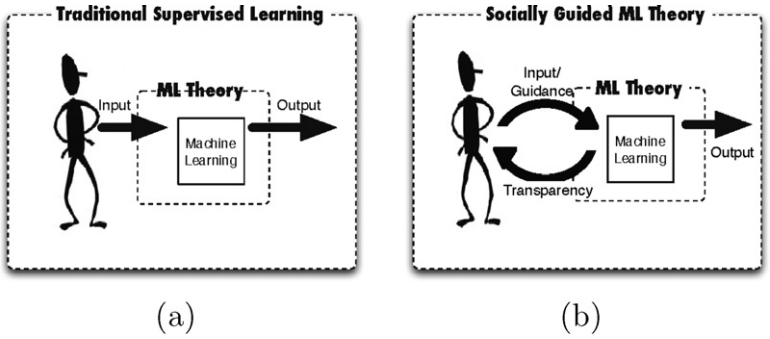


Fig. 1. (a) is a standard view of supervised learning: analyze input then output a model or classifier. Our approach includes the human teacher, (b), emphasizing that teaching/learning is a two-way process. We add transparency, where the machine learner provides feedback to the human teacher during learning; and we augment the human input with guidance. We aim to enhance the performance of the tightly coupled partnership of a machine learner with a human teacher.

action between the human and the machine. This allows us to take advantage of human teaching behavior to construct a Machine Learning process that is more amenable to an everyday human partner.

Fig. 1(a) is a high level view of a supervised Machine Learning process. A human provides input to the learning mechanism, which performs its task and provides the output. Alternatively, an HRI perspective of Machine Learning models the complete human-machine system, characterized in Fig. 1(b). This diagram highlights the key aspects of a social learning system. This interaction approach to Machine Learning challenges the research community to consider new questions, which we begin to explore in this paper. We need a principled theory of the content and dynamics of this tightly coupled process in order to design systems that can learn effectively from ordinary users.

Input channels: A social interaction approach asks: "How do humans want to teach?" In addition to designing the interaction based on what the machine needs for success, we also need to understand what kinds of intentions people will try to communicate in their everyday teaching behavior. We can then change the input portion of the Machine Learning training process to better accommodate a human partner. It is important to understand the many ways that natural human social cues (e.g. referencing, attention direction, etc.) can frame the learning problem for a standard Machine Learning process. This paper explicitly examines the effect of allowing the human to guide the attention of a learner as well as provide feedback during its exploration process.

Output channels: A social interaction approach asks: "How can the output provided by the agent improve the teaching-learning system?" In a tightly coupled interaction, a 'black box' learning process does nothing to improve the quality and relevance of the human's instruction. However, transparency about the process could greatly improve the learning experience. By communicating its internal state, revealing what is known and what is unclear, the robot can guide the teaching process. To be most effective, the robot should reveal its internal state in a manner that is intuitive for the human partner [2,6]. Facial expression, eye gaze, and behavior choices are a significant part of this output channel.

Input/output dynamics: Combining the previous two, a social interaction approach recognizes that these input and output channels interact over time. Furthermore, this dynamic can change the nature of the human's input. An incremental on-line learning system creates a very different experience for the human than a system that needs a full set of training examples before its performance can be evaluated. In an incremental system the human can provide more examples or correct mistakes right away instead of waiting to evaluate the results at the end of the training process. Moreover, the sense of progress may keep the human engaged with the training process for a longer time, which in turn benefits the learning system.

4. Experimental platform: *Sophie's Kitchen*

To investigate how social interaction can impact Machine Learning for robots, we have implemented a Java-based simulation platform, *"Sophie's Kitchen"*, to experiment with learning algorithms and enhancements. *Sophie's Kitchen* is an object-based state-action MDP space for a single agent, Sophie, with a fixed set of actions on a fixed set of stateful objects.

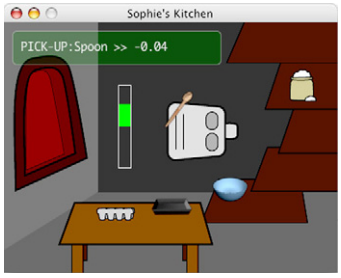


Fig. 2. *Sophie's Kitchen*. The agent is in the center, with a shelf on the right, oven on the left, a table in between, and five cake baking objects. The vertical bar is the interactive reward and is controlled by the human.

4.1. Sophie's Kitchen MDP

The task scenario is a kitchen world (see Fig. 2), where the agent, Sophie, learns to bake a cake. This system is defined by (L, O, Σ, T, A).

- There are a finite set of k locations $L = \{l_1, \ldots, l_k\}$. In our kitchen task, $k = 4$; $L = \{\texttt{Shelf}, \texttt{Table}, \texttt{Oven}, \texttt{Agent}\}$. As shown in Fig. 2, the agent is surrounded by a shelf, table and oven; and the location *Agent* is available to objects (i.e., when the agent picks up an object, then it has location *Agent*).
- There is a finite set of n objects $O = \{o_1, \ldots, o_n\}$. Each object can be in one of an object-specific number of mutually exclusive object states. Thus, Ω_i is the set of states for object o_i, and $O^* = (\Omega_1 \times \cdots \times \Omega_n)$ is the entire object configuration space. In the kitchen task scenario $n = 5$: the objects `Flour`, `Eggs`, and `Spoon` each have only one object state; the object `Bowl` has five object states: `empty`, `flour`, `eggs`, `both`, `mixed`; and the object `Tray` has three object states: `empty`, `batter`, `baked`.
- Let L^A be the possible agent locations: $L^A = \{\texttt{Shelf}, \texttt{Table}, \texttt{Oven}\}$; and let L^O be the possible object locations: $L^O = \{\texttt{Shelf}, \texttt{Table}, \texttt{Oven}, \texttt{Agent}\}$. Then the legal set of states is $\Sigma \subset (L^A \times L^O \times O^*)$, and a specific state is defined by $(l_a, l_{o_1} \ldots l_{o_n}, \omega)$: the agent's location, $l_a \in L^A$, and each object's location, $l_{o_i} \in L^O$, and the object configuration, $\omega \in O^*$.
- T is a transition function: $\Sigma \times A \mapsto \Sigma$. The action space A is expanded from four atomic actions (`GO<x>`, `PUT-DOWN<x>`, `PICK-UP<x>`, `USE<x><y>`): Assuming the locations L^A are arranged in a ring, the agent can always `GO left` or `right` to change location; she can `PICK-UP` any object in her current location; she can `PUT-DOWN` any object in her possession; and she can `USE` any object in her possession on any object in her current location. The agent can hold only one object at a time. Thus the set of actions available at a particular time is dependent on the particular state, and is a subset of the entire action space, A. Executing an action advances the world state in a deterministic way defined by T. For example, executing `PICK-UP <Flour>` advances the state of the world such that the `Flour` has location `Agent`. `USE`ing an ingredient on the `Bowl` puts that ingredient in it; using the `Spoon` on the `both-Bowl` transitions its state to the `mixed-Bowl`, etc.

In the initial state, s_0, all objects and the agent are at location `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. Some end 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. In order to encourage short sequences, an inherent negative reward ($r = -.04$) is placed in any non-goal state.

This is a novel task domain that, seemingly simple, has sufficient complexity to experiment with RL agents. The kitchen task has on the order of 10,000 states, and between 2 and 7 actions available in each state. Additionally, the cake task has the realistic and desirable quality of being hierarchical, thus there are many different ways to reach the goal. To have an idea of how difficult this task is for an RL agent, we ran tests where the agent learns only by itself with ($r = -1$) for disaster states, ($r = 1$) for goal states, and ($r = -.04$) for any other state. The agent starts in random states rather than s_0 after a disaster/goal is reached. We found that it took the agent a few thousand actions to reach the goal for the first time (on average, over 5 such self-learning experiments). This serves as a useful baseline to keep in mind. In all of our experiments with a human partner described in this paper, the additional feedback and support

1: s = last state, s' = current state, a = last action, r = reward
2: **while** learning **do**
3: a = random select weighted by $Q[s, a]$ values
4: execute a, and transition to s'
(small delay to allow for human reward)
5: sense reward, r
6: update Q-value:

$$Q[s, a] \leftarrow Q[s, a] + \alpha(r + \gamma(\max_{a'} Q[s', a']) - Q[s, a])$$

7: **end while**

Algorithm 1. Q-Learning with interactive rewards from a human partner.

from the human allows the agent reach the goal state for the first time an order of magnitude faster (on the order of 100 actions).

Due to the flexibility of the task, there are many trajectories that can lead to the desired goal. Here is one such action sequence: `PICK-UP Bowl`; `GO right`; `PUT-DOWN Bowl`; `GO left`; `PICK-UP Flour`; `GO right`; `USE Flour,Bowl`; `PUT-DOWN Flour`; `GO left`; `PICK-UP Eggs`; `GO right`; `USE Eggs,Bowl`; `PUT-DOWN Eggs`; `GO left`; `PICK-UP Spoon`; `GO right`; `USE Spoon,Bowl`; `PUT-DOWN Spoon`; `GO left`; `PICK-UP Tray`; `GO right`; `PUT-DOWN Tray`; `PICK-UP Bowl`; `USE Bowl,Tray`; `PUT-DOWN Bowl`; `PICK-UP Tray`; `GO right`; `PUT-DOWN Tray`.

4.2. Learning algorithm

The algorithm implemented for the experiments in this paper is a standard Q-Learning algorithm (learning rate $\alpha = .3$ and discount factor $\gamma = .75$) [37], shown above in Algorithm 1. A slight delay happens in line 4 as the agent's action is animated. This also allows the human time to issue interactive rewards. Q-Learning is used as the instrument for this work because it is a widely understood RL algorithm, thus affording the transfer of these lessons to other reinforcement-based approaches.

4.3. Interactive rewards interface

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

The interface also lets 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 doing a feedback message on a particular object (objects are highlighted when the mouse is over them to indicate that any subsequent reward will be object specific). This distinction exists to test a hypothesis that people will prefer to communicate feedback about particular aspects of a state rather than the entire state. However, object specific rewards are used only to learn about the human trainer's behavior and communicative intent; the learning algorithm treats all rewards in the traditional sense of pertaining to a whole state and action pair.

5. Experiment: How people teach RL agents

Some may note that restricting the human to a reinforcement signal is not the most efficient mechanism for our baking task, but it is important to note that we are using *Sophie's Kitchen* as a tool to better understand human interaction with an exploratory learner. In many problems or tasks, the human teacher may not know precisely what actions the agent needs to take, but they may have enough intuition to guide the learner's exploration. This is the scenario that our research aims to inform.

It may seem obvious that a standard Reinforcement Learning agent is not ideal for learning a task using interactive reward training as described above—if only due to the vast number of trials necessary to form a reasonable policy.

However, the details of what exactly needs adjustment, and what human factors are dominant in such an interaction, are largely unexplored. It is these components that we wish to uncover and enumerate. The purpose of this initial experiment with *Sophie's Kitchen* is to understand, when given a single reward channel (as in prior works), how do people use it to teach the agent?

5.1. Experiment design

In the experiment, 18 volunteers from the campus community and came to our research lab. After a brief introduction, each participant played the game. The system maintains an activity log and records time step and real time of each of the following: state transitions, actions, human rewards, reward aboutness (if object specific), disasters, and goals. Afterwards they answered a brief survey, and completed an informal interview with the experimenter. Participants were asked to rate their expertise with Machine Learning software and systems, (1 = none, 7 = very experienced), and we found it was an above average but reasonably diverse population (mean = 3.7; standard deviation = 2.3).[4] The following are the full instructions participants were given about the task:

The Game Setup: In this study you play a video game. This game has one character, Sophie, a robot in a kitchen. Sophie begins facing a shelf that has objects that can be picked up, put down, or used on other things (a bowl, a spoon, a tray, flour, and eggs). In the center of the screen is a table, the workspace for preparing foods before they go in the brick oven.

Baking a Cake: In this game your goal is for Sophie to bake a cake, but she does not know how to do the task yet. Your job is to get Sophie to learn how to do it by playing this training game. The robot character has 'a mind of its own' and when you press the "Start" button on the bottom of the screen, Sophie will try to start guessing how to do the task. Overall steps include: 1) make batter by putting both the flour and eggs in the bowl and 2) mix them with the spoon. 3) then put the batter into the tray 4) then put the tray in the oven.

Feedback Messages: You can't tell Sophie what actions to do, and you can't do any actions directly, you're only allowed to give Sophie feedback by using the mouse. When you click the mouse anywhere on the kitchen image, a rectangular box will appear. This box shows the message that you are going to send to Sophie.

- Drag the mouse UP to make the box GREEN, a POSITIVE message.
- Drag the mouse DOWN to make the box RED, a NEGATIVE message.
- By lifting the mouse button, the message is sent to Sohpie, she sees the color and size of the message and it disappears.
- Clicking the mouse button down on an object tells Sophie that your message is about that object. As in, "Hey Sophie, this is what I'm talking about..." (the object lights up to let you know you're sending an object specific message).
- If you click the mouse button down anywhere else, Sophie assumes that your feedback pertains to everything in general.

Disasters and Goals: Sometimes Sophie will accidentally do actions that lead to the Disaster state. (Like putting the spoon in the oven!) When this happens "Disaster" will flash on the screen, the kitchen gets cleaned up and Sophie starts a new practice round. Additionally, if Sophie successfully bakes the cake, "Goal!" will flash on the screen, the kitchen gets cleaned up and Sophie starts a new practice round. For the disaster state, Sophie is automatically sent a negative message. For the goal state, Sophie is automatically sent a positive message.

Completing the Study: Play the training game with Sophie until you believe that she can get the cake baked all by herself (or you've had enough fun with the training game, whichever happens first!). Note that she may need your help baking the cake more than once before she can do it herself. When you think she's got it, press the 'Finish' button and notify the experimenter.

5.2. Results of the teaching study

Of the 18 participants only one person did not succeed in teaching Sophie the task. During the first day of testing, four participants had to interrupt their trial due to a software error. As a result, some of the analysis below includes

[4] We had both male and female participants, but did not keep gender statistics.

only the 13 individuals that finished the complete task. However, since participants who experienced this error still spent a significant amount of time training the agent, their data is included in those parts of the analysis that relate to overall reward behavior. In this section we present three main findings about how people approach the task of teaching an RL agent with an interactive reward signal. 1) They assume the ability to guide the agent. 2) Their teaching behavior changes as they develop a mental model for the learning agent. 3) There is a positive bias in rewards.

5.2.1. Guidance intentions

Even though the instructions clearly stated that communication of both general and object specific rewards were *feedback* messages, many people assumed that object specific rewards were future directed messages or guidance for the agent. Several people mentioned this in the interview, and this is also suggested through behavioral evidence in the game logs.

An object specific reward used in a standard RL sense, should pertain to the last object the agent used. Fig. 3 has a mark for each player, indicating the percentage of object rewards that were about the last object: 100% would indicate that the player always used object rewards in a feedback connotation, and 0% would mean they never used object rewards as feedback. We can see that several players had object rewards that were rarely correlated to the last object (i.e., for 8 people less than 50% of their object rewards were feedback about the last object).

Interview responses suggested these people's rewards actually pertain to the future, indicating what they want (or do not want) the agent to use next. We look at a single test case to show how many people used object rewards as a guidance mechanism: When the agent is facing the shelf, a guidance reward could be administered about what to pick up. A positive reward given to either the empty bowl or empty tray on the shelf could *only* be interpreted as guidance since this state would not be part of any desired sequence of the task (only the initial state). Thus, rewards to empty bowls and trays in this configuration serve to measure the prevalence of guidance behavior.

Fig. 4 indicates how many people tried giving rewards to the empty bowl or empty tray on the shelf. Nearly all of the participants, 15 of 18, gave rewards to these objects sitting empty on the shelf. Thus, many participants tried using the reward channel to guide the agent's behavior to particular objects, giving rewards for actions the agent was *about to do* in addition to the traditional rewards for what the agent had just done.

These *anticipatory* rewards observed from everyday human trainers will require new attention in learning systems in order for agents to correctly interpret their human partners. Section 8 covers the design, implementation, and evaluation of algorithm and interface modifications for guidance.

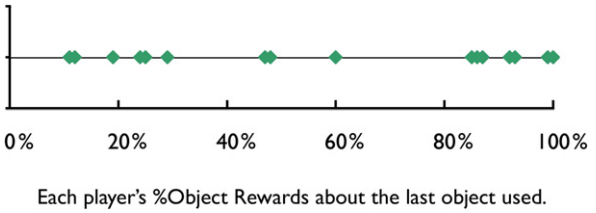


Fig. 3. There is one mark for each player indicating the percentage of object rewards that were about the last object of attention. Many people's object rewards were rarely about the last object, rarely used in a feedback connotation.

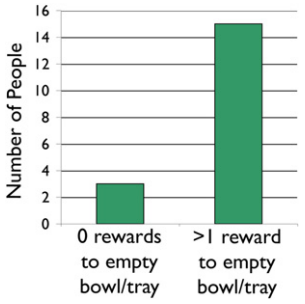


Fig. 4. A reward to the empty bowl or tray on the shelf is assumed to be meant as guidance instead of feedback. This graph shows that 15 of the 18 players gave rewards to the bowl/tray empty on the shelf.

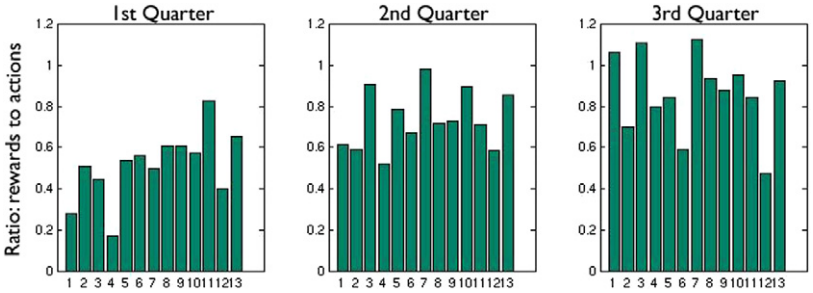


Fig. 5. Ratio of rewards to actions over the first three quarters of the training sessions shows an increasing trend.

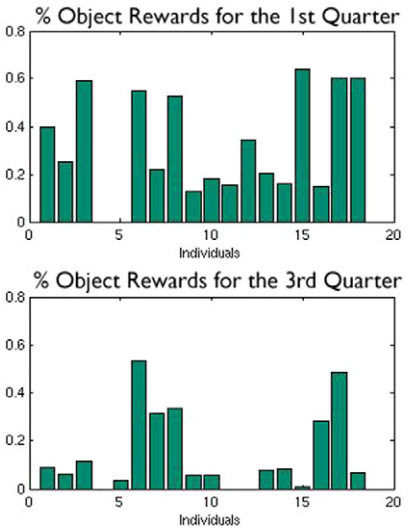


Fig. 6. Each bar represents an individual; the height is the percentage of object rewards. The difference in the first and last training quarters shows a drop in usage.

5.2.2. Inferring a model of the learner

Informed by related work [13], it is reasonable to expect people would habituate to the activity and that feedback would decrease over the training session. However, the opposite was found: the ratio of rewards to actions over the entire training session had a mean of .77 and standard deviation of .18. Additionally, there is an increasing trend in the rewards-to-actions ratio over the first three quarters of training. Fig. 5 shows data for the first three quarters for training, each graph has one bar for each individual indicating the ratio of rewards to actions. A 1:1 ratio in this case means that the human teacher gives a reward after every action taken by the agent. By the third quarter more bars are approaching or surpassing a ratio of 1.

One explanation for this increasing trend is a shift in mental model; as people realize the impact of their feedback they adjusted their reward schedule to fit this model of the learner. This 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 agent learn and they subsequently gave more rewards. Many users described the agent as a "stage" learner, that it would seem to make large improvements all at once. This is precisely the behavior one sees with a Q-Learning agent: fairly random exploration initially, and the results of learning are not seen until the agent restarts after a failure. Without any particular understanding of the algorithm, participants were quickly able to develop a reasonable mental model of the agent. They were encouraged by the learning progress, and subsequently gave more rewards.

A second expectation was that people would naturally use goal-oriented and intentional communication (which we attempted to measure by allowing people to specify object specific rewards, see Section 4.3). The difference between the first and last quarters of training shows that many people tried the object specific rewards at first but stopped using them over time (Fig. 6). In the interview, many users reported that the object rewards "did not seem to be working".

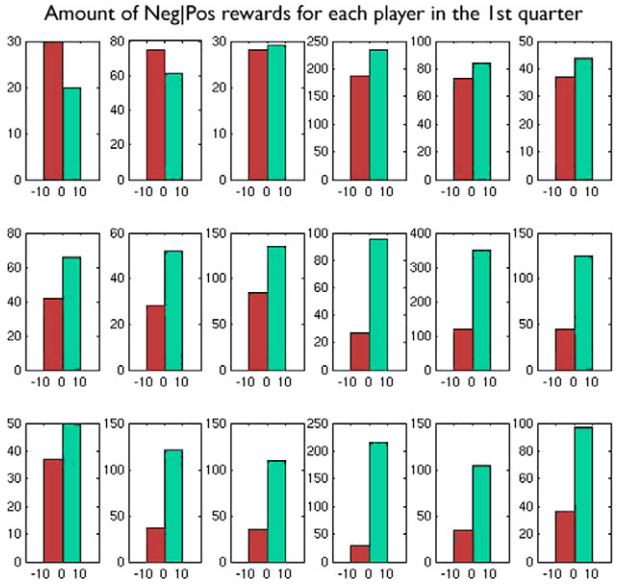


Fig. 7. Histograms of rewards for each individual in the first quarter of their session. The left column is negative rewards and the right is positive rewards. Most people even in the first quarter of training have a much higher bar on the right.

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.

These are concrete examples of the human trainer's propensity to learn from the agent how they can best impact the process. This presents a huge opportunity for an interactive learning agent to *improve its own learning environment* by communicating more internal state to the human teacher, making the learning process transparent. Section 9 details the implementation and evaluation of a transparent gazing behavior to improve the learning environment.

5.2.3. An asymmetric use of rewards

For many people, a large majority of rewards given were positive, the mean percentage of positive rewards for all players was 69.8%. This was thought at first to be due to the agent improving and exhibiting more correct behavior over time (soliciting more positive rewards); however, the data from the first quarter of training shows that well before the agent is behaving correctly, the majority of participants still show a positive bias. Fig. 7 shows reward histograms for each participant's first quarter of training; the number of negative rewards on the left and positive rewards on the right, most participants have a much larger bar on the right. A plausible hypothesis is that people are falling into a natural teaching interaction with the agent, treating it as a social entity that needs encouragement. Some people specifically mentioned in the interview that they felt positive feedback would be better for learning. Section 10 details the implementation and evaluation of *Sophie's Kitchen* with asymmetric use of human rewards.

6. Lessons learned from the teaching study

The findings in this study offer empirical evidence to support our Social Learner Hypothesis and the *partnership* of humans teaching artificial agents. When untrained users are asked to interactively train a RL agent, we see them treat the agent in a social way, tending towards positive feedback, guiding the robot, and adjusting their training behavior in reaction to the learner. Importantly, we see this tendency even without specifically adding any behavior to the robot to elicit this attitude. This demonstrates the human propensity to treat other entities as intentional agents. To date, RL does not account for the teacher's commitment to adapt to the learner, presenting an opportunity for an interactive learning agent to improve its own learning environment by communicating more of its internal state.

Additionally, our findings indicate that the learning agent can take better advantage of the different kinds of messages a human teacher is trying to communicate. In common RL, a reward signal is stationary and is some function

of the environment. It is usually a symmetrical scalar value indicating positive or negative feedback for being in the current state or for a particular state-action pair. Introducing human-derived real-time reward prompts us to reconsider these assumptions. We find that with a single communication channel people have various communicative intents—feedback, guidance, and motivation. Augmenting the human reward channel will likely be helpful to both the human teacher and the learning algorithm.

Finally, timing of rewards has been a topic in the RL community, particularly the credit assignment problem associated with delayed rewards. As opposed to delayed rewards, however, we saw that many human teachers administered anticipatory or guidance rewards to the agent. While delayed rewards have been discussed, the concept of rewarding the *action the agent is about to do* is novel and will require new tools and attention in the RL community.

7. Next steps: Modifications and follow-up experiments

The results of our first experiment suggest a few specific recommendations for interactive Machine Learning. One is that the communication from the human teaching partner cannot be merged into one single reward signal. We need to embellish the communication channel to account for the various intentions people wish to convey to the machine, particularly guidance intentions. 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.

In order to more deeply understand the impact social guidance and transparency behaviors can have on a Machine Learning process, we examine the following extensions in four follow-up versions of the Sophie game:

Guidance: Having found people try to communicate both guidance and feedback in their reward messages, a follow-up version of Sophie distinguishes between these two inputs. Users can still send a normal feedback message, but can also communicate attention direction or guidance. The learning algorithm is biased to select actions based on this attention direction signal.

Gaze as a transparency behavior: A second modification explores the effect of gazing between the objects of attention of candidate actions during action selection. The amount of gazing that precedes action communicates uncertainty. We expect this transparency behavior will improve the teacher's mental model of the learner, creating a more understandable interaction for the human and a better learning environment for the machine.

Undo: A third modification has the Sophie agent respond to negative feedback with an `UNDO` behavior (natural correlate or opposite action) when possible. This is expected to increase the responsiveness and transparency of the agent and could balance the amount of positive and negative rewards seen. The algorithm changes such that after negative feedback, the action selection mechanism chooses the action that 'un-does' the last action if possible.

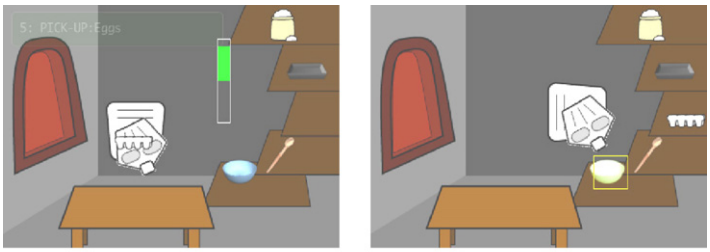
Motivation: One hypothesis about the positive rewards bias is that people were using the reward channel for motivation. A fourth modification of the Sophie game allows explicit encouragement or discouragement by administering a reward *on* Sophie. This will allow people to distinguish specific feedback about the task (e.g., "That was good!") from general motivational feedback (e.g., "Doing good Sophie!").

8. Leveraging human guidance

Theoretically, it is known that supervision can improve an RL process. Prior works have shown this by allowing a trainer to influence action selection with domain-specific advice [8,21] or by directly controlling the agent's actions during training [28]. These approaches have been tested with experts (often the algorithm designer), and lead us to expect that a guidance signal will improve learning (which we confirm in an experiment with an expert trainer). The contribution of our work is the focus on non-expert trainers. In an experiment we show that everyday human teachers can use attention direction as a form of guidance, to improve the learning behavior of an RL agent.

8.1. Modification to game interface

The guidance intentions identified in our teaching experiment suggest that people want to speak directly to the action selection part of the algorithm, to influence the exploration strategy. To accomplish this, we added a guidance channel of communication to distinguish this intention from feedback. Clicking the right mouse button draws an outline of a yellow square. When the yellow square is administered on top of an object, this communicates a guidance message to the learning agent and the content of the message is the object. Fig. 8(b) shows the player guiding Sophie to



(a) Feedback message. (b) Guidance message.

Fig. 8. The embellished communication channel includes the feedback messages as well as guidance messages. In (a), feedback is given by left-clicking and dragging the mouse up to make a green box (positive) and down for red (negative). In (b), guidance is given by right-clicking on an object, selecting it with the yellow square. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

```
1:  while learning do
2:      while waiting for guidance do
3:          if receive human guidance message then
4:              g = guide-object
5:          end if
6:      end while
7:      if received guidance then
8:          a = random selection of actions containing g
9:      else
10:         a = random selection weighted by Q[s, a] values
11:     end if
12:     execute a, and transition to s'
        (small delay to allow for human reward)
13:     sense reward, r
14:     update Q-value:
            Q[s, a] ← Q[s, a] + α(r + γ(max_a' Q[s', a']) − Q[s, a])
15: end while
```

Algorithm 2. Interactive Q-Learning modified to incorporate interactive human guidance in addition to feedback.

pay attention to the bowl. Note, the left mouse button still allows the player to give feedback as described previously, but there are no longer object rewards.

8.2. *Modification to learning algorithm*

Conceptually, our modified version gives the learning algorithm a pre-action and post-action phase in order to incorporate the new guidance input. 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. The modified Q-Learning process is shown in Algorithm 2 (see Algorithm 1 for the process used for the initial experiment with *Sophie's Kitchen*).

The agent begins each iteration of the learning loop by pausing to allow the teacher time to administer guidance (1.5 seconds). The agent saves the object of a guidance messages as g. During the action selection step, the default behavior chooses randomly between the set of actions with the highest Q-values, within a bound β. However, if any guidance messages were received, the agent will *instead* choose randomly between the set of actions that have to do with the object g. In this way the human's guidance messages bias the action selection mechanism, narrowing the set of actions the agent considers.

Table 1
An *expert* user trained 20 agents, with and without guidance, following a strict best-case protocol in each condition; this yields theoretical best-case effects of guidance on learning. (F = failed trials, G = first success.) Results from 1-tailed t-tests

Measure	Mean no guide	Mean guide	chg	t(18)	p
# trials	6.4	4.5	30%	2.48	.01
# actions	151.5	92.6	39%	4.9	<.01
# F	4.4	2.3	48%	2.65	<.01
# F before G	4.2	2.3	45%	2.37	.01
# states	43.5	25.9	40%	6.27	<.01

8.3. Evaluation: Guidance improves learning

The first experiment with the guidance modification evaluates the effects of guidance from an expert trainer. This is analogous to prior works, and serves to confirm that supervision is beneficial to the agent in *Sophie's Kitchen*. We collected data from expert[5] training sessions, in two conditions:

(1) `No guidance`: has feedback only and the trainer gives one positive or negative reward after every action.
(2) `Guidance`: has both guidance and feedback available; the trainer uses the same feedback behavior and additionally guides to the desired object at every opportunity.

For the user's benefit, we limited the task for this testing (e.g., taking out the spoon/stirring step, among other things). We had one user follow the above expert protocol for 10 training sessions in each condition. The results of this experiment are summarized in Table 1, showing that guidance improves several learning metrics. The number of training trials needed to learn the task was significantly less, 30%; as was the number actions needed to learn the task, 39% less. In the `guidance` condition the number of unique states visited was significantly less, 40%; thus the task was learned more efficiently. And finally the `guidance` condition was more successful, the number of trials ending in failure was 48% less, and the number of failed trials before the first successful trial was 45% less.

Having confirmed that guidance has the potential to drastically improve several metrics of the agent's learning behavior, our next evaluation of the guidance modification evaluates performance with everyday human trainers.

We solicited 28 people to come to our research lab to play the *Sophie's Kitchen* game, people were assigned to one of two groups. One group played the game with only the feedback channel (the `no guidance` condition). The other group had both feedback and guidance messages (the `guidance` condition). We added the following instructions about the guidance messages to the instructions from the previous experiment (and took out object specific rewards):

> *Guidance Messages*: *You can direct Sophie's attention to particular objects with guidance messages. Click the right mouse button to make a yellow square, and use it to guide Sophie to objects, as in 'Pay attention to this!'*

The comparison of these two groups is summarized in Table 2. The ability for the human teacher to guide the agent's attention to appropriate objects at appropriate times creates a significantly faster learning interaction. The number of training trials needed to learn the task was 48.8% less in the `guidance` condition, and the number actions needed was 54.9% less.

The `guidance` condition provided a significantly more successful training experience. The number of trials ending in failure was 37.5% less, and the number of failed trials before the first successful trial was 41.2% less. A more successful training experience is particularly desirable for robot learning agents that may not be able to withstand many failures. A successful interaction, especially reaching the first successful attempt sooner, may also help the human feel that progress is being made and prolong their engagement in the process.

Finally, agents in the `guidance` condition learned the task by visiting a significantly smaller number of unique states, 49.6% less. Additionally, we analyze the time spent in a good portion of the state space, defined as $G = \{$every

[5] One of the authors.

Table 2
Non-expert human players trained Sophie with and without guidance communication and also show positive effects of guidance on the learning. (F = failed trials, G = first success). Results from 1-tailed t-tests

Measure	Mean no guide	Mean guide	chg	t(26)	p
# trials	28.52	14.6	49%	2.68	<.01
# actions	816.44	368	55%	2.91	<.01
# F	18.89	11.8	38%	2.61	<.01
# F before G	18.7	11	41%	2.82	<.01
# states	124.44	62.7	50%	5.64	<.001
% good states	60.3	72.4		−5.02	<.001

unique state in X}, where $X = \{$all non-cyclic sequences, $s_0, \ldots, s_n$, such that $n \leqslant 1.25(\textit{min\_sequence\_length})$, and $s_n =$ a goal state$\}$. The average percentage of time that `guidance` agents spent in G was 72.4%; significantly higher than the 60.3% average of `no guidance` agents. Thus, attention direction helps the human teacher keep the exploration of the agent within a smaller and more useful portion of the state space. This is a particularly important result since the ability to deal with large state spaces has long been a criticism of RL. A human partner may help the algorithm overcome this challenge.

9. Transparency to guide a human teacher

In the previous section, we saw that the ability for the human teacher to direct the Sophie agent's attention has significant positive effects on several learning metrics. This section reports a related result—that the ability of the agent to use gaze as a transparency behavior results in measurably better human guidance instruction.

Gaze requires that the learning agent have a physical/graphical embodiment that can be understood by the human as having a forward heading. In general, gaze precedes an action and communicates something about the action that is going to follow. In this way gaze serves as a transparency device, allowing an onlooker to make inferences about what the agent is likely to do next, their level of confidence and certainty about the environment, and perhaps whether or not guidance is necessary. A gaze behavior was added to the *Sophie's Kitchen* game. More than 50 people played the modified game over the World Wide Web, and data collected allows for a concrete analysis of the effect that gaze had on a human teacher's behavior.

9.1. Modification to game interface

Recall the interactive Q-Learning algorithm modified for guidance (Algorithm 2). The gaze behavior modification makes one alteration to the stage at which the agent is waiting for guidance, shown in Algorithm 3. When the agent is waiting for guidance, it finds the set of actions, A^+, with the highest Q-values, within a bound β. $\forall a \in A^+$, the learning agent gazes for 1 second at the `object-of-attention` of a (if it has one). For an example of how the Sophie agent orients towards an object to communicate gazing, see Fig. 9. This gazing behavior during the pre-action phase communicates a level of uncertainty through the amount of gazing that precedes an action. It introduces an additional delay (proportional to uncertainty) prior to the action selection step, both soliciting and providing the opportunity for guidance messages from the human. This also communicates overall task certainty or confidence as the agent will stop looking around when every set, A^+, has a single action. The hypothesis is that this transparency will improve the teacher's model of the learner, creating a more understandable interaction for the human and a better learning environment for the agent.

9.2. Evaluation: Gaze improves guidance

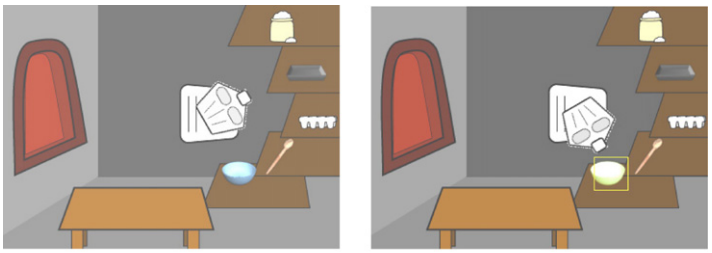
To evaluate the use of transparency, we deployed the *Sophie's Kitchen* game on the World Wide Web. Participants were asked to play the computer game and were given instructions on administering feedback and guidance. Each of the 52 participants played the game in one of the following test conditions:

while learning **do**
  $A^+ = [a_1 \ldots a_n]$, the n actions from s with the highest Q values within a bound β
  for $i = 1 \ldots n$ **do**
    $o =$ the object of attention of a_i
    if $o \neq null$ **then**
      set gaze of the agent to be o for 1 sec.
    end if
  end for
  if receive human guidance message **then**
    $g = guide\text{-}object$
    $a =$ random selection of actions containing g
  else
    $a =$ random selection weighted by $Q[s, a]$ values
  end if
  execute a, and transition to s'
  (small delay to allow for human reward)
  sense reward, r
  update policy:

$$Q[s, a] \leftarrow Q[s, a] + \alpha(r + \gamma(\max_{a'} Q[s', a']) - Q[s, a])$$

end while

Algorithm 3. Interactive Q-Learning with guidance and a gazing transparency behavior.



(a) (b)

Fig. 9. Sophie's gaze transparency behavior. In (a) Sophie is facing the shelf, gazing at the tray prior to selecting an action; in (b) at the bowl.

- `Guidance`: Players had both feedback and guidance communication.
- `Gaze-guide`: Players had feedback and guidance channels. Additionally, the agent used the gaze behavior.

The system maintained an activity log and recorded time step and real time of each of the following: state transitions, actions, human rewards, guidance messages and objects, gaze actions, disasters, and goals. These logs were analyzed to test the transparency hypothesis: Learners can help shape their learning environment by communicating aspects of the internal process. In particular, the gaze behavior will improve a teacher's guidance instruction.

To evaluate this we compare players in the `guidance` condition to those in the `gaze-guide` condition; these results are summarized in Fig. 10. Note that the players without the gaze behavior still had ample opportunity to administer guidance; however, the time that the agent waits is uniform throughout.

Looking at the timing of each player's guidance instruction, their communication can be separated into two segments: the percentage of guidance that was given when the number of action choices was $\geqslant 3$ (high uncertainty), and when choices were $\leqslant 3$ (low uncertainty), note that these are overlapping classes. Three is chosen as the midpoint because the number of action choices available to the agent at any time in the web-based version of *Sophie's Kitchen* is at most 5. Thus we describe a situation where the number of equally valued action choices is $\geqslant 3$ as high uncertainty, and $\leqslant 3$ as low uncertainty.

Players in the `gaze-guide` condition had a significantly lower percentage of guidance when the agent had low uncertainty compared to the players in the `guidance` condition, $t(51) = -2.22$, $p = .015$. And conversely the percentage of guidance when the agent had high uncertainty increased from the `guidance` to the `gaze-guide`

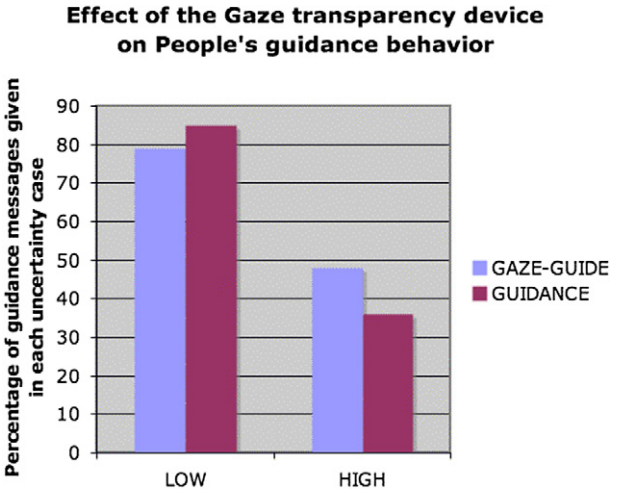


Fig. 10. Results of a 1-tailed t-test of the effect of gaze on guidance. The distribution of guidance messages was significantly different in the gaze vs. gaze-guide conditions. The gaze transparency device caused people to give less guidance when uncertainty was low, and more when uncertainty was high (uncertainty low = number of action choices $\leqslant 3$, high = number of choices $\geqslant 3$). T-test results for the difference seen here: in the low uncertainty case $t(51) = -2.22$, $p < .05$; and in the high uncertainty case $t(51) = 1.96$, $p < .05$.

condition, $t(51) = 1.96$, $p = .027$. Thus, when the agent uses the gaze behavior to indicate which actions it is considering, the human trainers do a better job matching their instruction to the needs of the agent throughout the training session. They give more guidance when it is needed and less when it is not.

We also looked at the speed and efficiency metrics between the `guidance` and `gaze-guide` groups, but did not find significant difference. We presume that this indicates that people are able to achieve the large performance gains seen in Section 8 anytime they are given the guidance channel. However, what the results from this experiment indicate is that people without the gaze indication seem to be 'overusing' the guidance channel, giving guidance whenever possible. With the gaze transparency behavior, on the other hand, people exhibit guidance communication that is less redundant and better matches the needs of the agent.

10. The asymmetry of human feedback

One of the main findings in our initial experiment concerned the biased nature of positive and negative feedback from a human partner. Clearly, people have different intentions they are communicating with their positive and negative feedback messages. In this section we present two modifications to the game interface that address the asymmetric meanings of human feedback.

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 to the agent, or feel that it is important to be both instrumental and motivational with their communication channel. In interviews a number of participants mentioned that they believed the agent would learn better from positive feedback.

Another hypothesis is that negative rewards did not produce the expected reaction from the robot. A typical RL agent does not have an instantaneous reaction to either positive or negative rewards, but in the case of negative rewards, this could be interpreted as the agent "ignoring" the human's feedback. In that case, the user may stop using them when they feel the agent is not taking their input into account. One way to address this is to introduce an `UNDO` behavior. Many actions (`PICK-UP`, `PUT-DOWN`, `TURN`) have a natural correlate or opposite action that can be performed in response to negative feedback. This could add to the responsiveness and transparency of the agent and balance the amount of positive and negative rewards seen.

We explore both of these hypotheses in this section. First, we look at adding a motivation channel of communication, to test if the positive bias was in part due to motivational intentions. Second, we add the `UNDO` behavior and show that this reaction to a person's negative feedback produces a significantly better learning behavior for the RL agent.

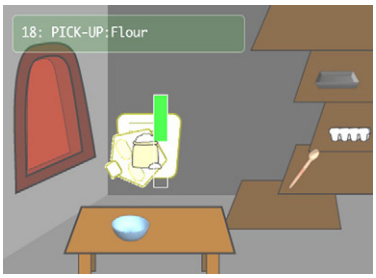


Fig. 11. A reward is considered motivational rather than instrumental if it is administered *on* Sophie, as pictured. Instructions about this input channel indicate that it is for general feedback (e.g. "Doing good Sophie!" or "Doing bad!") as opposed to feedback about a particular action.

10.1. Motivational communication

In this experiment, we add a motivation communication channel. Our hypothesis is that we should see the positive bias decrease when the players have a separate channel for motivational versus instrumental communication.

10.1.1. Modification to the game interface

For this experiment we have the original feedback channel of communication, and a dedicated motivational input. This is done by considering a reward motivational if it is administered *on* Sophie. For visual feedback the agent is shaded yellow to let the user know that a subsequent reward will be motivational. Fig. 11 shows a positive motivational message to Sophie. The game instructions given to players indicate that this input channel is for general feedback about the task (e.g. "Doing good Sophie!" or "Doing bad!") as opposed to feedback about a particular action.

10.1.2. Evaluation: Motivation intentions confirmed

To test our hypothesis about people's motivational intents, we deployed the *Sophie's Kitchen* game on the World Wide Web and had 98 people play the game. Players that had the motivation signal had a significantly more balanced feedback valance than the players that did not have it. Players that did not have a motivational channel had a mean ratio ($\frac{\#positive}{\#negative}$) of 2.07; whereas those with the motivational channel had a mean ratio of 1.688. This is a significant effect, $t(96) = -2.02$, $p = .022$. Thus, we conclude that motivation is a separate intention that was folded into the players' positive feedback in the initial study. Future work is to understand how an agent can utilize this signal in a different way to improve the learning interaction.

10.2. `UNDO` behavior

The `UNDO` modification addresses a second asymmetric meaning of human feedback. The intuition is that positive feedback tells a learner undeniably, "what you did was good". However, negative feedback has multiple meanings: 1) that the last action was bad, and 2) that the current state is bad and future actions should correct that. Thus, negative feedback is about both the past and about future intentions for action. In the final modification to *Sophie's Kitchen*, the algorithm assumes that a negatively reinforced action should be reversed if possible. This `UNDO` interpretation of negative feedback shows significant improvements in several metrics of learning.

10.2.1. Modification to the algorithm

This baseline algorithm is modified to respond to negative feedback with an `UNDO` behavior (a natural correlate or opposite action) when possible. Thus a negative reward affects the policy in the normal fashion, but also alters the subsequent action selection if possible. The proper `UNDO` behavior is represented within each primitive action and is accessed with an *undo* function: `GO [direction]` returns `GO [-direction]`; `PICK-UP [object]` returns `PUT-DOWN [object]`; `PUT-DOWN [object]` returns `PICK-UP [object]`; `USE` actions are not reversible. Algorithm 4 shows how this is implemented with the changes in lines 2–6, as compared to the baseline Algorithm 1.

1: **while** learning **do**
2: **if** (reward last cycle $< -.25$) and (can undo last action, a_{last}) **then**
3: $a = undo(a_{last})$
4: **else**
5: $a =$ random select weighted by $Q[s, a]$ values
6: **end if**
7: execute a, and transition to s'
(small delay to allow for human reward)
8: sense reward, r
9: update policy:

$$Q[s, a] \leftarrow Q[s, a] + \alpha(r + \gamma(\max_{a'} Q[s', a']) - Q[s, a])$$

10: **end while**

Algorithm 4. Interactive Q-Learning with the addition of the UNDO behavior.

Table 3
1-tailed t-test: Significant differences were found between the baseline and undo conditions, in training sessions with nearly 100 non-expert human subjects playing the *Sophie's Kitchen* game online

Measure	Mean baseline	Mean undo	chg	t(96)	p
# states	48.3	42	13%	−2.26	=.01
# F	6.94	4.37	37%	−3.76	<.001
# F before G	6.4	3.87	40%	−3.7	<.001
# actions to G	208.86	164.93	21%	−2.25	=.01
# actions	255.68	224.2	12%	−1.32	=.095

10.2.2. Evaluation of UNDO behavior

We found the UNDO response to negative feedback from the human trainer significantly improves the learning performance of the agent in a number of ways. Data was collected from 97 participants by deploying the *Sophie's Kitchen* game on the World Wide Web.

In this experiment, each participant played the game in one of two groups, offering a measurable comparison between two conditions of the learning algorithm. In the baseline case the algorithm handles both positive and negative feedback in a standard way, feedback is incorporated into the value function (Algorithm 1). In the undo case the algorithm uses feedback to update the value function but then also uses negative feedback in the action selection stage as an indication that the best action to perform next is the reverse of the negatively reinforced action (Algorithm 4). Statistically significant differences were found between the baseline and undo conditions on a number of learning metrics (summarized in Table 3).

The UNDO behavior helps the agent avoid failure. The total number of failures during learning was 37% less in the undo case. This is particularly interesting for robotic agents that need to learn in the real world, where learning from failure may not be a viable option. The undo case also had 40% less failures before the first success. This is especially important when the agent is learning with a human partner, who will have limited patience and will need to see progress quickly in order to remain engaged in the task. The undo behavior seems to be a good technique for reaching the first success faster.

The UNDO behavior seems to lead to a more efficient exploration. There was a nearly significant effect ($p = .095$) for the number of actions required to learn the task, with the undo condition requiring 12% fewer steps (the high degree of variance in the number of steps needed to learn the task leads to the higher p value). Another indication of the efficiency of the undo case compared to the baseline is in the state space needed to learn the task. The number of unique states visited is 13% less in the undo case. This indicates that when the algorithm interprets negative feedback as a directive for reversing the previous action, or returning to the previous state, the resulting behavior is more efficient in its use of the state space to learn the desired task.

11. Discussion

Robotic and software agents that operate in human environments will need the ability to learn new skills and tasks 'on the job' from everyday people. It is important for designers of learning systems to recognize that while the average person is not familiar with Machine Learning techniques, they are intimately familiar with various forms of social learning (e.g., tutelage, imitation, etc.). This raises two important and related research questions for the Machine Learning community. 1) How do people want to teach machines? 2) How do we design machines that learn effectively from natural human interaction?

In this article we have demonstrated the utility of a *Socially Guided Machine Learning* approach, exploring the ways machines can be designed to more fully take advantage of a natural human teaching interaction. Our work emphasizes the *interactive* elements in teaching. There are inherently two sides to an interaction, in this case the human teacher and the machine learner. Our approach aims to enhance standard Machine Learning algorithms from both perspectives of this interaction: modifying the algorithm to build a better learning agent, and modifying the interaction techniques to provide a better experience for the human teacher. Understanding how humans want to teach is an important part of this process.

Using human input with a learning agent has received some attention in the Machine Learning community (see Section 2). Many prior works have addressed how human input can theoretically impact a learning algorithm or interaction. In contrast, our work addresses the nature of *real* people as teachers; our ground truth evaluation is the performance of the machine learner with non-expert human teachers. Whereas prior works typically lend control either to the machine or the human, our contribution is the focus on how a machine learner can use transparency behaviors to steer the instruction it receives from a human, creating more reciprocal control of the interaction.

Several prior works that utilize a human teacher are inspired by animal or human learning. For instance, game characters that the human player can shape through interaction have been successfully incorporated into a few computer games [10,29,31]. Breazeal et al. have demonstrated aspects of collaboration and social learning on a humanoid robot, using social cues to guide instruction [7]. Animal training techniques and human tutelage have been explored in several robotic agents [14,20,25,30]. As a software agent example, Blumberg's virtual dog character can be taught via clicker training, and behavior can be shaped by a human teacher [5].

Many of these works agree with our situated learning paradigm for machines, and emphasize that an agent should use social techniques to create a better interface for a human partner. Our work goes beyond gleaning inspiration from natural forms of social learning/teaching to formalize and empirically ground it in observed human teaching behavior through extensive user studies. One of the main contributions of this work is empirical evidence that social interaction not only creates a good interface for a human partner, but also creates a better learning environment and significant learning benefits for an agent.

Our findings indicate that a learning agent can take better advantage of the different kinds of messages a human teacher tries to communicate. Given a single communication channel, people have various communicative intents.

In addition to common instrumental feedback, *people assume they can guide the agent*, even when they are explicitly told that only feedback messages are supported. In their guidance communication, people mean to bias the action selection mechanism of the RL algorithm. When we allow this, introducing a separate interaction channel for attention direction and modifying the action selection mechanism of the algorithm, we see a significant improvement in the agent's learning. The agent is able to learn tasks using fewer actions over fewer trials. It has a more efficient exploration strategy that wasted less time in irrelevant states. We argue that a less random and more sensible exploration will lead to more understandable and teachable agents. Guidance also led to fewer failed trials and less time to the first successful trial. This is particularly important since it implies a less frustrating teaching experience, which in turn creates a more engaging interaction for the human.

We also see that players treat the agent as a social entity and want a *motivational channel of communication* to encourage it. This is seen despite the fact that the learning agent in this work is very mechanistic and simplistic. One can assume that this effect will only be more prominent with characters that are explicitly designed to be socially and emotionally appealing. We argue that to build successful agents that learn from people, attention of the research community should focus on understanding and supporting the psychology and social expectations of the human teacher. It remains future work to explore how this motivational or encouragement channel of communication should influence the learning algorithm in a different way than the ordinary positive and negative feedback. Our hypothesis is that this communication is intended to influence the internal motivations, drives and goals of the agent.

This work offers a concrete example that the *transparency of the agent's behavior to the human can improve its learning environment*. In a social learning interaction both learner and teacher influence the performance of the tutorial dyad. While this observation seems straightforward in the human literature, little attention has been paid to the communication between human teacher and artificial agent in the traditional Machine Learning literature. Particularly, we believe that the transparency of the learner's internal process is paramount to the success of the tutorial dialog. Specifically, this work has shown that when the learning agent uses gazing behaviors to reveal its uncertainties and potential next actions, people were significantly better at providing more guidance when it was needed and less when it was not. Thus the agent, through its own behavior, was able to shape the human's input to be more appropriate. Gaze is just one such transparency device, the exploration of other transparency devices and their relation to the learning process is part of our future work.

Additionally, as designers we add these transparency behaviors to boost the overall realism and believability of the agent, thereby making it more engaging for the human. The creation of believable characters that people find emotionally appealing and engaging has long been a challenge [3,32]. Autonomy complicates this goal further, since the character has to continually make action choices that are reasonable and useful as well as believable and engaging. Blumberg et al. has some of the most extensive work in this domain [4,35] within a dog learning context. Thus another challenge for teachable robots is to be appropriately responsive to the human's instruction.

In this work we have studied one aspect of such responsiveness, informed by our initial user study. Negative feedback from a human teacher can be treated as both feedback for the action and suggestion to perform an `UNDO` behavior and *reverse the action if possible*. When this is part of the agent's behavior, learning is improved in both speed and efficiency.

We chose to use the Q-Learning algorithm for this work because it is standard and widely understood. This affords the transfer of these lessons and modifications to other reinforcement-based approaches. We have shown significant improvements in an RL domain, showing that learning in a situated interaction with a human partner can help overcome some of the well recognized problems of RL. Furthermore, these improvements in learning will contribute to higher quality interactive robotic and software agents that are better equipped to take advantage of the ways that people naturally approach teaching.

12. Conclusion

This work shows that designing for the complete human–machine learning system creates a more successful learning agent. Our initial experiment with an interactive computer game lead to three main findings: people assume they can guide the agent, they dynamically adjust their behavior as they develop a model of the agent, and they have a positive bias in their rewards.

We addressed these findings in four follow-up experiments with modified versions of the *Sophie's Kitchen* game. Our modifications include: an embellished channel of communication that distinguishes between guidance, feedback, and motivational intents; endowing the character with transparency behaviors that reveal specific aspects of the agent's learning process; and providing a more natural reaction to negative feedback. A series of user studies show that these empirically informed modifications result in learning improvements across several dimensions including the speed of task learning, the efficiency of state exploration, the understandability of the agent's learning process for the human, and a significant drop in the number of failures encountered during learning.

Importantly, in this work we acknowledge that the ground truth evaluation, for systems meant to learn from people, is performance with non-expert humans. This topic deserves more attention from the Machine Learning community, as it will be important for progress towards a social learning scenario for machines. This series of experiments with an interactive learning agent illustrates the effectiveness of this approach for building machines that can learn from ordinary people. The ability to utilize and leverage social skills is far more than a nice interface technique. It can positively impact the dynamics of underlying learning mechanisms to show significant improvements in a real-time interactive learning session with non-expert human teachers.

Acknowledgements

This work is funded in part by the Media Lab corporate sponsors of the Things that Think and Digital Life consortia.

References

[1] M. Argyle, R. Ingham, M. McCallin, The different functions of gaze, Semiotica (1973) 19–32.

[2] R. Arkin, M. Fujita, T. Takagi, R. Hasegawa, An ethological and emotional basis for human–robot interaction, in: Proceedings of the Conference on Robotics and Autonomous Systems, 2003.

[3] J. Bates, The role of emotion in believable agents, Communications of the ACM 37 (7) (1997) 122–125, http://citeseer.ist.psu.edu/bates94role.html.

[4] B. Blumberg, Old tricks, new dogs: Ethology and interactive creatures, Ph.D. thesis, Massachusetts Institute of Technology, 1997.

[5] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. Johnson, B. Tomlinson, Integrated learning for interactive synthetic characters, in: Proceedings of the ACM SIGGRAPH, 2002.

[6] C. Breazeal, Designing Sociable Robots, MIT Press, Cambridge, MA, 2002.

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

[8] J. Clouse, P. Utgoff, A teaching method for reinforcement learning, in: Proc. of the Ninth International Conf. on Machine Learning (ICML), 1992, pp. 92–101.

[9] D. Cohn, Z. Ghahramani, M. Jordan, Active learning with statistical models, in: G. Tesauro, D. Touretzky, J. Alspector (Eds.), Advances in Neural Information Processing, vol. 7, Morgan Kaufmann, 1995.

[10] R. Evans, Varieties of learning, in: S. Rabin (Ed.), AI Game Programming Wisdom, Charles River Media, Hingham, MA, 2002, pp. 567–578.

[11] P.M. Greenfield, Theory of the teacher in learning activities of everyday life, in: B. Rogoff, J. Lave (Eds.), Everyday Cognition: Its Development in Social Context, Harvard University Press, Cambridge, MA, 1984.

[12] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, K. Rommelse, The lumiere project: Bayesian user modeling for inferring the goals and needs of software users, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July 1998, pp. 256–265.

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

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

[15] R.M. Krauss, Y. Chen, P. Chawla, Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? in: M. Zanna (Ed.), Advances in Experimental Social Psychology, Academic Press, Tampa, 1996, pp. 389–450.

[16] L.S. Vygotsky, in: M. Cole (Ed.), Mind in Society: The Development of Higher Psychological Processes, Harvard University Press, Cambridge, MA, 1978.

[17] Y. Lashkari, M. Metral, P. Maes, Collaborative interface agents, in: Proceedings of the Twelfth National Conference on Artificial Intelligence, vol. 1, AAAI Press, Seattle, WA, 1994, http://citeseer.ist.psu.edu/lashkari94collaborative.html.

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

[19] H. Lieberman (Ed.), Your Wish is My Command: Programming by Example, Morgan Kaufmann, San Francisco, CA, 2001.

[20] A. Lockerd, C. Breazeal, Tutelage and socially guided robot learning, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2004.

[21] R. Maclin, J. Shavlik, L. Torrey, T. Walker, E. Wild, Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression, in: Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI), Pittsburgh, PA, July 2005.

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

[23] T.M. Mitchell, S. Wang, Y. Huang, Extracting knowledge about users' activities from raw workstation contents, in: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI), 2006.

[24] M.N. Nicolescu, 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.

[25] L.M. Saksida, S.M. Raymond, D.S. Touretzky, Shaping robot behavior using principles from instrumental conditioning, Robotics and Autonomous Systems 22 (3/4) (1998) 231.

[26] S. Schaal, Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences 3 (1999) 233–242.

[27] G. Schohn, D. Cohn, Less is more: Active learning with support vector machines, in: Proc. 17th ICML, Morgan Kaufmann, San Francisco, CA, 2000, pp. 839–846.

[28] W.D. Smart, L.P. Kaelbling, Effective reinforcement learning for mobile robots, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2002, pp. 3404–3410.

[29] K.O. Stanley, B.D. Bryant, R. Miikkulainen, Evolving neural network agents in the nero video game, in: Proceedings of IEEE 2005 Symposium on Computational Intelligence and Games (CIG'05), 2005.

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

[31] A. Stern, A. Frank, B. Resner, Virtual petz (video session): A hybrid approach to creating autonomous, lifelike dogz and catz, in: AGENTS '98: Proceedings of the Second International Conference on Autonomous Agents, ACM Press, New York, 1998, pp. 334–335.

[32] F. Thomas, O. Johnson, Disney Animation: The Illusion of Life, Abbeville Press, New York, 1981.

[33] S. Thrun, Robotics, in: S. Russell, P. Norvig (Eds.), Artificial Intelligence: A Modern Approach, second ed., Prentice Hall, 2002.

[34] S.B. Thrun, T.M. Mitchell, Lifelong robot learning, Tech. Rep. IAI-TR-93-7, 1993.

[35] B. Tomlinson, B. Blumberg, Social synthetic characters, Computer Graphics 26 (2) (2002).

[36] R. Voyles, P. Khosla, A multi-agent system for programming robotic agents by human demonstration, in: Proceedings of AI and Manufacturing Research Planning Workshop, 1998.
[37] C. Watkins, P. Dayan, Q-learning, Machine Learning 8 (3) (1992) 279–292.
[38] J.V. Wertsch, N. Minick, F.J. Arns, Creation of context in joint problem solving, in: B. Rogoff, J. Lave (Eds.), Everyday Cognition: Its Development in Social Context, Harvard University Press, Cambridge, MA, 1984.