Reinforcement Learning with Human Teachers: Understanding How People Want to Teach Robots

Andrea L. Thomaz, Guy Hoffman, Cynthia Breazeal

***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 robot a task through Reinforcement Learning: (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. In conclusion, we discuss future extensions to RL to accommodate these lessons.

I. Introduction

Machine learning shall play a significant role in the development of robotic assistants that operate in human environments (e.g., homes, schools, hospitals, offices). Considering the difficulty of hard-coding all the information needed for the robot to play a long term role in a dynamic world, human users will need to be able to easily teach such robots. Various works have addressed some of the hard problems robots face when learning in the real-world [1], [2], [3]. However, learning *from a human teacher* poses additional challenges (and benefits) for Machine Learning systems.

Several examples of agents that learn interactively with a human teacher are based on Reinforcement Learning (RL). RL has certain desirable qualities, such as the possibility to explore and learn from unsupervised experience. However, many also question RL as a viable technique for learning in complex real-world environments because of practical problems such as long training time requirements, non-scaling state representations, sparse rewards (resulting in slow utility propagation), and safe exploration strategies. Many of these considerations are particularly pertinent to robots using RL, prompting the use of human guidance. As a result, Reinforcement Learning has been utilized for teaching robots and game characters, incorporating real-time human feedback by having a person supply reward and/or punishment as an additional input to the reward function [4], [5], [6], [7], [8].

Most of this work models the human input as indistinguishable from any other feedback coming from 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 do this properly, we must understand the human teacher's contribution: *how* does the human teach, and *what* do they try to communicate to a robot learner? In this work we aim to understand a human teacher's contribution in guiding a robot's active exploration.

Sophie's Kitchen is a video game framework for studying the impact of social interaction for RL. We report results from a user study with the game, which records a social interaction with a Q-Learning agent[1]. We present observations of the teaching strategies that the human instructors employed in training the game agent. To our knowledge, this paper is the first to explicitly address and report such results, relevant to any interactive learning algorithm.

In our experiment we find that people try to relay to the robot information that the algorithm has no means to interpret, suggesting specific ways in which RL algorithms should change to accommodate real-time interaction with human users. Our main findings are threefold:

- In addition to administering feedback, users want to *guide* the agent towards an action and give anticipatory rewards. While delayed rewards have been discussed in the machine learning literature [9], anticipatory rewards and guidance are not part of the RL model.
- We find that users give *more positive than negative* feedback, possibly reflecting their opinions on motivation and human learning, or feeling that their negative feedback is ignored by the robot. This is at odds with RL which is usually symmetrical with regard to the valence of reward.
- We find that users read the behavior of the learner and adjust their training strategies as their mental model of the agent changes. Viewing the human input as a traditional RL reward signal does not take advantage of the fact that a benevolent teacher adjusts their training behavior to best suit the learner.

II. Related Work

In addition to the related RL works mentioned above, several works address the topic of human input for machine learning systems. Personalization agents and adaptive user interfaces are examples of software that learns by observing

[1]Q-Learning is used as the instrument for this work because it is a simple and widely understood formulation of RL, thus affording the transfer of these lessons to any reinforcement-based approach.

human behavior, modeling human preferences or activities [10]. However, our work is concerned with explicit training where the human teaches the learner through interaction.

Various works address 'trainable' software and robotic agents, exploring explicit human input: learning classification tasks [11] and navigation tasks [12] via natural language, robots that learn by demonstration or example [13], [14], [15], and software agents that learn by example or training [16], [17]. While somewhat social and natural, many of these approaches constrain the teacher to a special interaction or language. Also, in many, the learning problem is essentially equivalent to programing new tasks through natural interfaces, leaving little exploration on the part of the machine.

Active learning [18], [19] is an approach that explicitly acknowledges a human in the loop. In contrast to the above works, in active learning it is the algorithm that drives the interaction by issuing sparse queries to the human. However, work in that field has not addressed social aspects, or investigated how humans would want to teach learning machines, which is the focus of this paper.

III. The Sophie's Kitchen Platform

To investigate the ways in which social interaction can impact machine learning for robots, we have implemented a Java-based simulation platform, *"Sophie's Kitchen"*. Sophie's Kitchen is an object-based state-action MDP space for a single agent, Sophie, using a fixed set of actions on a fixed set of stateful objects.

A. Sophie's Kitchen MDP

An object-based state-action world $W = \langle L, O, \Sigma, A, T \rangle$ 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 object states, and in one of the locations in L. If Ω_i denotes the set of states for object o_i, $O^* = \langle \Omega_1 \times \ldots \times \Omega_n \rangle$ is the entire object configuration space. Similarly, L^* is the object location space $L^* = \langle L \times \ldots \times L \rangle$. W is also defined by a set of legal states $\Sigma \subset \langle L \times L^* \times O^* \rangle$. A world state $s = (l_a, l_{o_1} \ldots l_{o_n}, \omega)$ consists of the agent's location, and the location and configuration, $\omega \in O^*$, of each object. Finally, W has a set of actions A, and a transition function $T : \Sigma \times A \mapsto \Sigma$.

In our experiments, we used a "kitchen" world (see Fig. 1), where the agent (Sophie) learns to bake a cake. This world has five objects: `Flour`, `Eggs`, a `Spoon`, a `Bowl` (with five states: `empty`, `flour`, `eggs`, `both`, `mixed`), and a `Tray` (with three 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).

The action space A is fixed and is defined by four atomic actions: Assuming the locations L 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. Each action implements a transition function in T that advances the world state. 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 `mixed`, etc.

Fig. 1. *Sophie's Kitchen*. The robot is facing the shelf, holding the spoon. The vertical green bar is controlled by the human and here shows a positive interactive reward of circa +.6 .

In the initial state, 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 (for example—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. The kitchen task has on the order of 10,000 states with between 2 and 7 actions available in each state.

The algorithm implemented for these experiments is a standard Q-Learning algorithm (learning rate $\alpha = .3$ and discount factor $\gamma = .75$) [20].

B. 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 $r = [-1, 1]$. The user receives visual feedback enabling them to tune the reward signal to a certain value before sending it to the agent. Choosing and sending the reward value does not halt the progress of the agent, which runs asynchronously to the interactive human reward.

Additionally, the interface lets the user make a distinction between rewarding the whole state of the world or the state of a particular object (object specific rewards). An object specific reward is administered by clicking the mouse button down on a specific object. As visual feedback to the user, the object is highlighted when pointed on, to indicate that any subsequent reward will be object specific. In the experiment, object specific rewards are used only to learn about the human trainer's behavior and communicative intent; the learning algorithm treats all rewards equally, in the traditional sense of pertaining to the whole state.

IV. Experiment

We had 18 participants play a game, in which their goal was to get the virtual robot, Sophie, to learn how to bake a cake on her own. Participants were asked on a scale of 1 to 7 how experienced they were with machine learning software and systems (1=no experience, 7=very experienced). We had an above average (mean=3.7), but reasonably diverse (standard deviation=2.3) population with respect to machine learning expertise.

Participants played the training game as long as they felt necessary. At this point the experimenter tested the agent and their game score was the degree to which Sophie finished baking the cake by herself. Participants received between $5 and $10 based on their game score.

Participants were told they could not tell Sophie what actions to do, nor could they do any actions directly. They were only able to give Sophie feedback messages with the mouse, according to the following instructions:

- Drag the mouse UP to make the box more GREEN, a POSITIVE message. Drag DOWN for RED/NEGATIVE.
- By lifting the mouse button, the message is sent to Sohpie, she sees the color and size of the message.
- If you click the mouse button down on an object, this tells Sophie your message is about that object (as in: "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, recording time step and real time of each of the following: state transitions, actions, human rewards, reward aboutness (if object specific), disasters, and goals. In addition to this behavioral data, participants completed a short questionnaire and an informal interview after the game.

V. Results

Of the 18 participants only one person did not succeed 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 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.

A. Guidance vs. Feedback

In Reinforcement Learning, rewards always pertain to the current state or previous action. In contrast, even though the instructions clearly stated that communication and rewards were *feedback* messages, we found that many people assumed that the object specific rewards were future directed messages or guidance for the agent. In the interview, subjects had one of two types of responses regarding their use of object specific rewards: 1) Many said that they used the object rewards to indicate the desired/undesired object of attention or next object to use. Similarly, some said they tried to use the object reward in the above way, but that it "didn't seem to work." 2) A second group of subjects said they did not understand what the object rewards meant for the agent. In other words, many subjects reportedly used or tried to use their reward feedback to *guide* the agent. The following behavioral data quantifies this self-reported guidance behavior.

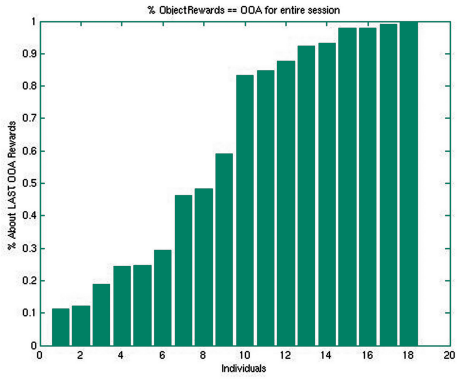

Fig. 2. Percentage of object rewards that were about the last object of attention (OOA). Sorted to show that many people's object rewards rarely correlated with the last OOA.

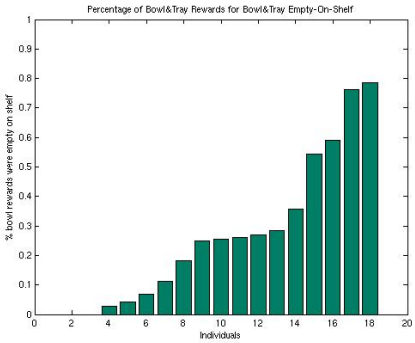

Fig. 3. Percentage of bowl and tray rewards given when the bowl or tray was empty on the shelf. We can assume that this is a guidance reward rather than a feedback reward.

If people were using the object rewards in the traditional RL sense, these rewards should always pertain to the last object the agent used. Figure 2 shows the percentage of object specific rewards that were given to the last object the agent used. The graph shows one bar per player, sorted, indicating that nearly half of the subjects gave object rewards that were rarely correlated to the last object. Specifically, for 8 people less than 50% of their object rewards pertained to the last object.

To further examine whether subjects used these rewards in a future directed way (as a guidance mechanism) we looked at a single test case: when the agent is facing the shelf, 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. Thus, rewards to empty bowls and trays in this configuration can serve to measure the prevalence of guidance behavior.

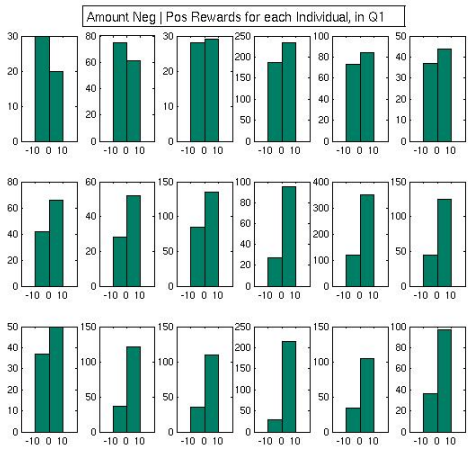

Fig. 4. Histograms of rewards for each individual during the first quarter of their session. Left bar corresponds to negative rewards; right bar to positive rewards.

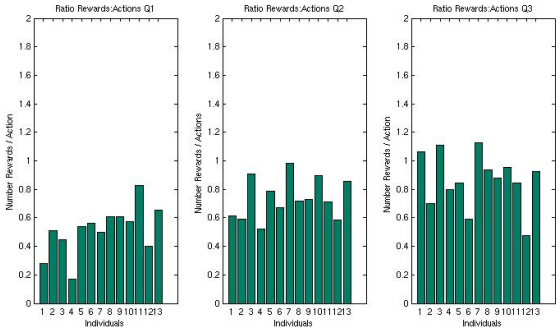

Fig. 5. Ratio of rewards to actions over the first three quarters of the training sessions shows an increasing trend.

Figure 3 shows a graph, with one bar per subject. The Y-axis indicates the percentage of bowl and tray rewards that were given when these objects were empty on the shelf. Well over half of the participants gave a substantial percentage of bowl and tray rewards to the objects sitting empty on the shelf, with very few never engaging in this behavior. This leads to the conclusion that 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 RL rewards for the last action.

B. Positive Bias in Rewards

Another finding concerns the valence of human rewards. We found that for many subjects, a majority of rewards given were positive. The mean percentage of positive rewards was 69.8%. Initially, we thought that this was due to the agent improving and exhibiting more correct behavior over time (soliciting more positive rewards). However, examining the data from the first quarter of training we find that well before the agent is behaving correctly, the majority of participants show a positive bias. Fig. 4 shows reward histograms for each participant's first quarter of training; the left bar indicates the number of negative rewards and the right the number of positive rewards. As can be clearly seen, most participants tend towards positive rewards.

This positive bias is an interesting area for follow-up study. 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. In future studies we attempt to remedy this problem by introducing 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 will explore both hypotheses presented in this section in our future work.

C. Shifting Mental Models

A third set of findings relates to the subjects' adaptation to the learner. Informed by related work [6], we expected people to habituate to the teaching activity and that, as a result, feedback would decrease over the training session. However, we found just the opposite to be true: not only did the ratio of rewards to actions over the entire training session have a mean of .77 and standard deviation of .18, there was also 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, with each graph including one bar per individual, indicating the ratio of rewards to actions. The graphs show a clear rising tendency between the first and the third quarter.

We explain this as a shift in mental model; as people realize the impact of their feedback they adjust their reward schedule to fit this model of the learner. This hypothesis finds anecdotal support in the interview responses. Subjects reported that at some point they realized that their feedback was helping the agent learn, even if there was no immediate response from the robot, and subsequently gave more rewards. Many users described the agent as a "stage" learner, one that seems to make large improvements all at once. This is precisely the behavior one sees with a Q-Learning agent: fairly random exploration to update a policy, with the results of learning not seen until the agent restarts after a failure. Without a-priori understanding of the algorithm, many participants were quickly able to develop the right mental model of the agent through the interaction. As a

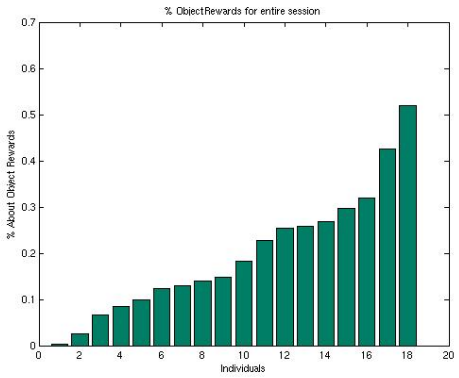

Fig. 6. Percentage of object specific rewards per subject.

result, they were encouraged by the learning progress, and subsequently gave more rewards. It is noteworthy that this behavior is contrary to the requirements of traditional RL, which benefits more from early rewards than from rewards given later in the process.

A second instance of adaptation can be seen in the delivery of object-specific rewards. Subjects varied greatly in the usage of these object rewards (Fig. 6). Further, in examining the difference between the first and last quarters of training, we see that many people tried the object specific rewards at first but stopped using them over time (Fig. 7). In the interview, many users reported that the object rewards "did not seem to be working." Thus, many participants tried to give object specific rewards in the beginning, 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, see Sec. III-B). Once they saw that the object rewards did not meet their expectations or did not seem to affect the agent's behavior they revised their mental model of the learner and stopped using the object rewards.

VI. Discussion

In this work we have attempted to investigate how Reinforcement Learning can be adapted to better suit human-robot interaction, specifically looking at how humans will want to teach robots. In humans, teaching and learning is a collaboration. Teachers direct a learner's attention, structure experiences, support learning attempts, and regulate complexity. The learner contributes to the coupling by revealing their internal state to help guide the teaching process. Teacher and learner read and respond to each other, to more effectively guide the learner's exploration. We believe that this view can also inform social learning in robots, using social cues and gestures to achieve transparency and guide instruction [21], [22].

The findings in this study offer empirical evidence to support this concept of *partnership* when humans teach 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 as the interaction proceeds,

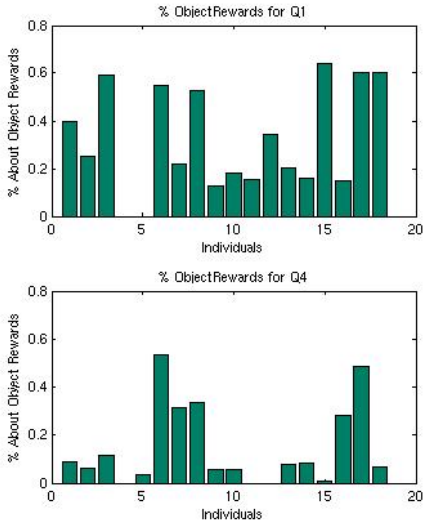

Fig. 7. Each bar represents an individual and the height is the percentage of object rewards. The difference in the first (top) and last (bottom) training quarters shows a drop off in usage over time.

reacting to the behavior of the learner. Importantly, we see this tendency even without specifically adding any behavior to the robot to elicit this attitude. This suggests that there is a human propensity to treat and understand other entities as intentional agents, and to adapt to them.

To date, RL does not accommodate 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 to the human teacher.

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 machine 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.

VII. Future Work

Our findings suggest recommendations for designing Reinforcement Learning with human interaction in mind: (1) We

need to embellish the communication channel to account for the various intentions people wish to convey to the machine, particularly guidance intentions and motivational messages. (2) People tune their behavior to match the needs of the machine learner, and this process should be augmented with more transparency of the internal state on the part of the learner. To further understand the impact of social guidance on a machine learning process, we are running follow-up studies with a second version of the Sophie video game that includes the following:

Gaze as a Transparency Behavior: The second version of our study explores the effect of gazing between the objects of attention for equally valuable candidate actions during the action selection phase. This communicates a level of uncertainty through the amount of gazing that precedes action. Additionally, this communicates overall task certainty as the process will speed up when the gazing between actions is no longer necessary. We expect this transparency behavior to 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. Specifically, we expect more rewards to be administered when the agent seems uncertain.

Guidance: Having found people try to communicate both guidance and feedback in their reward message, the next version of Sophie distinguishes between these two inputs. Users can still send a normal feedback message using the left mouse button (Sec. III-B), but they can also use the right mouse button to communicate attention direction or guidance. The learning algorithm changes such that if, during the pre-action phase, the human teacher administers any guidance input, the algorithm tries to select an action that contains this object as its object of attention.

Undo: In the next version, the Sophie agent responds 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 in the step following 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. The next version of the Sophie game allows explicit encouragement or discouragement of Sophie. This allows people to distinguish specific feedback about the task (e.g., "That was good!") from general motivational feedback (e.g., "Doing good Sophie!").

VIII. Conclusions

The introduction of a human real-time reward signal brings about a range of new considerations for robot learning. We have presented a simulation framework 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. This work calls for Machine Learning systems that are specifically designed with human feedback and social guidance in mind, for the benefit of both human teacher and machine learner.

References

[1] M. Mataric, "Reinforcement learning in the multi-robot domain," *Autonomous Robots*, vol. 4, no. 1, pp. 73–83, 1997.

[2] S. B. Thrun and T. M. Mitchell, "Lifelong robot learning," University of Bonn, Institut fuer Informatik III, Tech. Rep. IAI-TR-93-7, 1, 1993.

[3] S. Thrun, "Robotics," in *Artificial Intelligence: A Modern Approach (2nd edition)*, S. Russell and P. Norvig, Eds. Prentice Hall, 2002.

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

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

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

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

[8] A. Stern, A. Frank, and 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*. New York, NY, USA: ACM Press, 1998, pp. 334–335.

[9] L. P. Kaelbling, M. L. Littman, and A. P. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[10] Y. Lashkari, M. Metral, and P. Maes, "Collaborative Interface Agents," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*. Seattle, WA: AAAI Press, 1994, vol. 1. [Online]. Available: citeseer.ist.psu.edu/lashkari94collaborative.html

[11] L. Steels and F. Kaplan, "Aibo's first words: The social learning of language and meaning," *Evolution of Communication*, vol. 4, no. 1, pp. 3–32, 2001.

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

[13] 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.

[14] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in Cognitive Sciences*, vol. 3, p. 233242, 1999.

[15] 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.

[16] H. Lieberman, Ed., *Your Wish is My Command: Programming by Example*. San Francisco: Morgan Kaufmann, 2001.

[17] J. Blythe, "Task learning by instruction in tailor," in *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, 2005.

[18] D. Cohn, Z. Ghahramani, and M. Jordan., "Active learning with statistical models," in *Advances in Neural Information Processing*, G. Tesauro, D. Touretzky, and J. Alspector, Eds. Morgan Kaufmann, 1995, vol. 7.

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

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

[21] 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*, vol. 1, no. 2, 2004.

[22] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as collaboration," in *Proceedings of the AAMAS*, 2004.