Connection Science Journal
Special Issue on Social Learning in Embodied Agents
Vol. X, No. X, Month 200X, 000–000

RESEARCH ARTICLE

Experiments in Socially Guided Exploration: Lessons Learned in Building Robots that Learn with and without Human Teachers

Abstract

We present a learning system, Socially Guided Exploration, in which a social robot learns new tasks through a combination of self-exploration and interpersonal interaction. The system's motivational drives (novelty, mastery), along with social scaffolding from a human partner, bias behavior to create learning opportunities for a Reinforcement Learning mechanism. The robot is able to learn on its own, but can flexibly use the guidance of a human teacher to improve performance. We report the results of a series of experiments where the robot learns on its own in addition to being taught by human subjects. We analyze these interactions to understand human teaching behavior and the social dynamics of the human-teacher/robot-learner system. With respect to learning performance, human guidance results in a task set that is significantly more focused and efficient, while self-exploration results in a broader set. Analysis of human teaching behavior reveals insights of social coupling between human teacher and robot learner, different teaching styles, strong consistency in the kinds and frequency of scaffolding acts across teachers, and nuance in the communicative intent behind positive and negative feedback.

Keywords*:* Human-Robot Interaction; Machine Learning; Artificial Intelligence; Social Scaffolding; Computational Models of Social Learning

NAMES WITHHELD[a*]

[a]*Department, University, City, Country;* [b]*Department, University, City, Country*

Connection Science Journal
Special Issue on Social Learning in Embodied Agents
Vol. X, No. X, Month 200X, 000–000

RESEARCH ARTICLE

Experiments in Socially Guided Exploration: Lessons Learned in Building Robots that Learn with and without Human Teachers

ANDREA L. THOMAZ,[a1*] CYNTHIA BREAZEAL [b]
[a]*Interactive Computing, Georgia Institute of Technology, Atlanta, GA USA;* [b]*Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA USA*

Abstract

We present a learning system, Socially Guided Exploration, in which a social robot learns new tasks through a combination of self-exploration and interpersonal interaction. The system's motivational drives (novelty, mastery), along with social scaffolding from a human partner, bias behavior to create learning opportunities for a Reinforcement Learning mechanism. The robot is able to learn on its own, but can flexibly use the guidance of a human teacher to improve performance. We report the results of a series of experiments where the robot learns on its own in addition to being taught by human subjects. We analyze these interactions to understand human teaching behavior and the social dynamics of the human-teacher/robot-learner system. With respect to learning performance, human guidance results in a task set that is significantly more focused and efficient, while self-exploration results in a broader set. Analysis of human teaching behavior reveals insights of social coupling between human teacher and robot learner, different teaching styles, strong consistency in the kinds and frequency of scaffolding acts across teachers, and nuance in the communicative intent behind positive and negative feedback.

Keywords*:* Human-Robot Interaction; Machine Learning; Artificial Intelligence; Social Scaffolding; Computational Models of Social Learning

[1] Corresponding Author, Email: athomaz@cc.gatech.edu
[*] This research was conducted at the MIT Media Lab.

1. Introduction

Enabling a human to efficiently transfer knowledge and skills to a robot has inspired decades of research. When much of this prior work is viewed along a *guidance-exploration* spectrum, an interesting dichotomy appears. Many prior systems are strongly dependent on human *guidance*, learning nothing without human interaction (e.g., learning by demonstration [Chernova07,Atkeson97] or by tutelage [Nicolescu03,Lockerd04]). In systems such as these, the learner does little if any exploration on its own to learn tasks or skills beyond what it has observed with a human. Furthermore, the teacher often must learn how to interact with the machine and know precisely how it needs to perform the task.

Other approaches are almost entirely *exploration* based. For example, many prior works have given a human trainer control a reinforcement learner's reward [Blumberg02, Kaplan02, Saksida98], allow a human to provide advice [Clouse92, Maclin05], or have the human tele-operate the agent during training [Smart02]. Exploration approaches have the benefit that learning does not require the human's undivided attention. However, they often give the human trainer a very restricted role to scaffold learning, and require the human to learn how to interact with the machine.

Our research is motivated by the promise of personal robots that operate in human environments to assist people on a daily basis. Personal robots will need to be able to learn new skills and knowledge while "on the job." Certainly, personal robots should be able to learn on their own – either discovering new skills and knowledge or mastering the familiar through practice. However, personal robots must also be able to learn from members of the general public who are *not* familiar with the technical details of robotic systems or Machine Learning algorithms. However, they do bring a lifetime of experience in learning from and teaching others. This is a collaborative process where the teacher guides the learner's exploration, and the learner's performance shapes further instruction through a large repertoire of social interactions. Therefore, personal robots should be designed to be social learners that can effectively leverage a broad repertoire of human scaffolding to be successful and efficient learners.

In sum, personal robots must be able to move flexibly along a *guidance-exploration* spectrum. They should be able to explore and learn on their own, but also take full advantage of a human teacher's guidance when available.

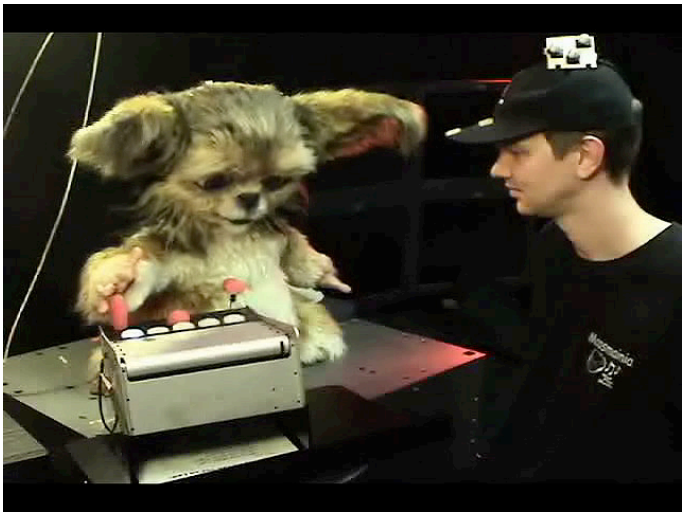

Figure 1. Our social robot, Leonardo, interacts with a human teacher to learn about a puzzle box.

In this paper, we present a novel learning architecture for a social robot that is inspired by theories in developmental psychology and is informed by recent advances in intrinsically motivated reinforcement learning (e.g., [Singh05, Oudeyer04, Schmidhuber05]). We call this approach *Socially Guided Exploration* to emphasize that the robot is designed to be an intrinsically motivated learner, but its exploration and learning process can readily take advantage of a broad repertoire of a human teacher's guidance and scaffolding.

Further, we approach this challenge from a Human-Robot Interaction perspective where we are interested in the human-teacher/robot-learner system. Hence, our analysis examines and compares the learning performance of the robot both when learning in isolation and with a human teacher. For the later, we conducted a human subjects experiment where we had 11 people (all previously unfamiliar with the robot) teach it to perform a number of tasks using a "smart" puzzle box. The puzzle box is pre-programmed with a suite of behaviors such as changing the color of its lights, opening or closing its lid, or playing a song when the correct sequence of button presses, switch flips, or slider toggles is performed. We analyze our human subjects' teaching behavior in relation to the robot's learning behavior to understand the dynamics of this coupled social process. Our findings reveal social constraints on how people teach socially interactive robots. These findings have important implications for how to design social robots that learn from everyday people.

2. Robot Platform

Our research platform is Leonardo ("Leo"), a 65 degree of freedom anthropomorphic robot specifically designed for human social interaction (Figure 1). Leo has speech and vision sensory inputs and uses gestures and facial expressions for social communication (the robot does not speak yet). Leo can visually detect objects in the workspace, humans and their head pose [Morency02], and hands pointing to objects. For highly accurate tracking of objects and people, we use a 10 camera VICON optical motion capture system. The speech understanding system is based on Sphinx, and has a limited grammar to facilitate accuracy.

The cognitive and learning system extends the C5M architecture [Blumberg02]. The Perception and Belief Systems are most relevant to the learning abilities described in this paper. Every time step, the robot has observations from its various sensory processes, $O=\{o_1,..,o_k\}$. The Perception System is a set of *percepts* $P = \{p_1,..., p_n\}$. Each $p \in P$ is a classification function, such that $p(o)=m$ where $m \in [0,1]$ is a match value. The Belief System maintains the belief set B by integrating these percepts into discrete object representations (based on spatial relationships and various similarity metrics). Figure 2 shows a simple example in which sensory data leads to five percepts with $m>0$, that result in two beliefs in B. In this paper, a "state" s refers to a snapshot of the belief set B at a particular time, and S refers to the theoretical set of all possible states. Let $A=\{a_1,...,a_i\}$ be the set of Leo's basic actions. For more details of the Perception and Belief Systems see [Breazeal05].

The Socially Guided Exploration system builds on these existing mechanisms --- adding capabilities for representing and learning goal-oriented tasks, self-motivated exploratory behavior, and expression/gesture capabilities to support a collaborative dialog with a human teacher.

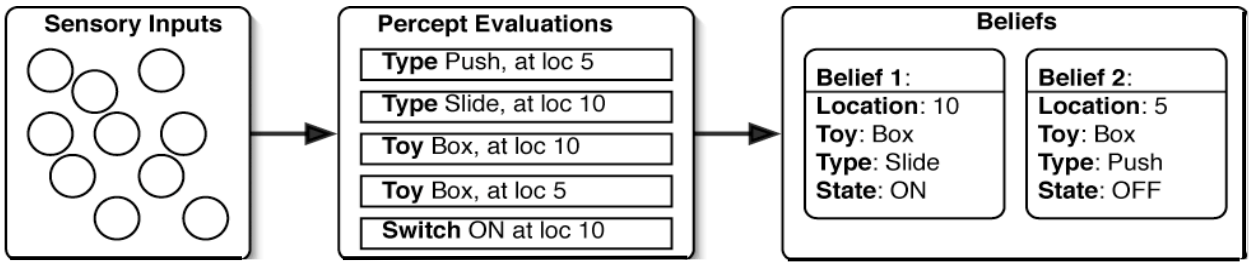

Figure 2. Sensory input is classified by percepts and then merged into discrete object representations. In this time step, five percepts yield two object beliefs.

3. Socially Guided Exploration

In most Machine Learning systems, learning is an explicit activity. Namely, the system is designed to learn a particular thing at a particular time. In human learning, on the other hand, learning is a part of all activity. There is a motivation for learning, a drive to know more about the environment, and an ability to seek out the expertise of others. Children explore and learn on their own, but in the presence of a teacher they can take advantage of the social cues and communicative acts provided to accomplish more (also known as *social scaffolding* [Vygotsky78]). A teacher often guides a learner by providing timely feedback, luring them to perform desired behaviors, and controlling the environment so the appropriate cues are salient, thereby making the learning process more effective. This is the primary inspiration for the Socially Guided Exploration system. This section highlights the key implementation details: the Motivation System, learning behaviors, goal-oriented task representation, transparency devices and social scaffolding mechanisms.

3.1 Motivational Drives for Learning

Living systems work to keep certain critical features within a bounded range through a process of behavioral homeostasis (e.g., food, water, temperature). If a parameter falls out of range, the animal becomes motivated to behave in a way that brings it back into the desired range.

Recently, this concept has inspired work on internal motivations for a Reinforcement Learning (RL) agent [Oudeyer04, Singh05, Shumidhuber05]. These works use a measure of novelty or certainty as intrinsic reward for a controller. Thus, an action that leads to a prediction error results in rewards that encourage focus on that portion of the space. Our approach is in a similar vein, but rather than contribute to the reward directly, Leo's internal motivations trigger learning behaviors that help the system arbitrate between learning a new task, practicing a learned task, and exploring the environment. Additionally, prior works in "motivated" RL have relied on a single drive (novelty/curiosity). In this work we introduce a mastery drive and demonstrate the benefits of the interplay between novelty and mastery in an agent's learning behavior.

Leo's Motivation System (based on prior work [Breazeal02]) is designed to guide a learning mechanism. Inspired by natural systems, it has two motivational drives, `Novelty` and `Mastery`. Each drive has a range [0,1], initial value of 0.5, a tendency to drift to 0.0, and a drift magnitude of 0.001 (max change in a time step). The Motivation System maintains the drive values based on the status of the internal and external environment:

The **`Novelty Drive.`** The `Novelty Drive` is an indication of the unfamiliarity of recent events. Every state transition will cause the `Novelty Drive` to rise for an amount of time related to the degree of the change, d_{chg}, based on the event's frequency: $d_{chg}(s_1,s_2) = 1/\text{frequency}(s_1,s_2)$. An event causes the `Novelty Drive` to drift towards its maximum value for a period, $t = d_{chg}(s_1,s_2)\, t_{max}$. The maximum effect time, t_{max}, is 30 seconds.

The **`Mastery Drive.`** The `Mastery Drive` reflects the current system confidence of the learned task set. `Mastery` is the average confidence of the tasks that are relevant in (i.e., can be initiated from) the current state, *s*. A task's confidence is the number of successful attempts over the total task attempts made.

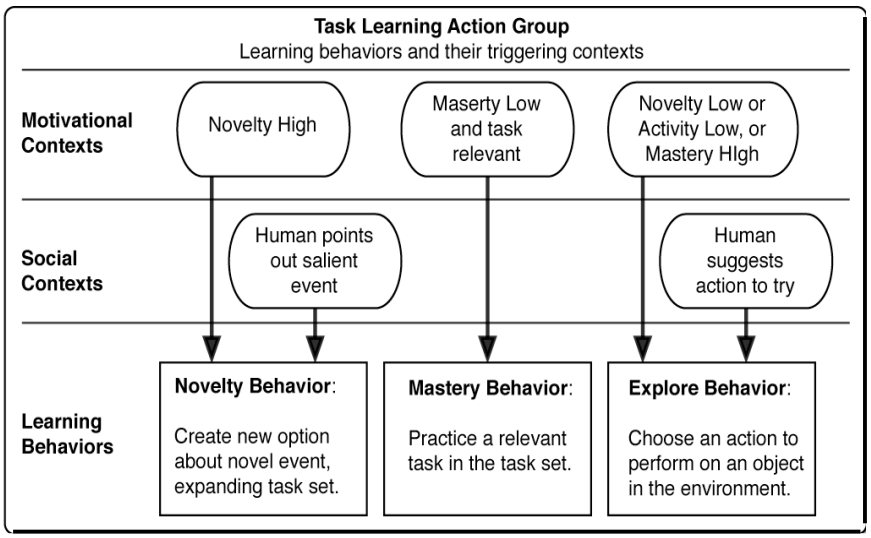

Figure 3. The three learning behaviors and their social/motivational contexts.

3.2 Learning Behaviors for Motivational & Social Contexts

The Task Learning Action Group is the piece of the Socially Guided Exploration system responsible for identifying and responding to learning opportunities in the environment. It maintains the set of known tasks (*Tasks*), and has three competing learning behaviors that respond to social and motivational learning contexts. Figure 3 is an overview of the behaviors and their internal/external triggering contexts.

The **`Novelty Behavior`**. One purpose of the `Novelty Drive` is to encourage the system to better understand new events, expanding the *Tasks* set. Thus, a significant rise in the `Novelty Drive` makes the `Novelty Behavior` available for activation. Additionally, this behavior may be activated due to a social context, when the human points out an event (e.g., "Look Leo, it's `TaskName-X`."). Once activated, the `Novelty Behavior` tries to create a new task. It makes a goal representation of the most recent state transition (s_1, *a*, s_2), and if there is not a $T \in Tasks$ with this goal, then a new task is created. Task creation, expansion, and generalization are covered below.

The **`Mastery Behavior`**. The purpose of the `Mastery Drive` is to cause the system to become confident in the environment, fleshing out the representations in the *Tasks* set. When the `Mastery Drive`

is low and any tasks are relevant in the current state, the `Mastery Behavior` may be activated. This behavior randomly selects a relevant task, executes it, and updates the confidence based on success in reaching the goal.

The **`Explore Behavior`**. Both motivational drives also work to encourage exploration. The `Explore Behavior` becomes available when `novelty` is low, encouraging the system to seek out the unexpected. Exploration is also triggered when `mastery` is high. Even if a known task is relevant, the system is biased to try to expand the *Tasks* set once confidence is high. Additionally, social interaction can trigger the `Explore Behavior` --- for example, if the human suggests an action (e.g., "Leo, try to `Act-X` the `Obj-Y`."). When the `Explore Behavior` is activated, it first tries to do any human-suggested action if possible. Otherwise, the `Explore Behavior` selects from the actions it can do in the current state, with a minimum frequency requirement. Once the action is completed, if it was a human-suggested action, the robot's attention is biased to look to the human in order to acknowledge the suggested action and provide the human with an opportunity for feedback.

3.3 Task and Goal Representation

These three behaviors result in a mechanism that learns object-oriented tasks. *Tasks* and their goals are represented with *Task Option Policies*. This name reflects its similarity to the Options approach in Reinforcement Learning [Sutton99].

Goals encode what must hold true to consider the task achieved. Specifically, a goal $G = x_1,..., x_y$ where every $x \in G$ represents a belief that changed over the task, grouping the belief's percepts into *expectation* percepts (indicating an expected feature value), and *criteria* percepts (indicating which beliefs to apply this expectation to).[2]

Each $T \in Tasks$ is a Task Option Policy, and is defined by a variation of the three Options constructs: I, π, β. To define these we use two subsets of states related to the task. Let $S_{task} \subset S$ be the states in which the task is relevant but not achieved, and $S_{goal} \subset S$ be the states in which the goal is achieved. Then, a Task Option Policy is defined by:

- π': $S_{task} \times A \rightarrow [0,1]$; estimates a value for (s, a) pairs in relation to achieving the task goal, G.
- β': S_{goal}; represents all of the states in which this task terminates because G is true.
- $I' = S_{task}$; represents the initiation set. The task can be initiated in any state for which it has a policy of action.

A task can be executed (is relevant) when the current state is in S_{task}. During execution, actions are chosen according to π' until the current state is in S_{goal} (with some probability of terminating early). A state s achieves the goal if: $\forall\ x \in G$, if any belief b in s matches all the *criteria* $\in x$, then b also matches all the *expectation* $\in x$.

3.4 Task Learning

The Socially Guided Exploration system learns a new Task Option Policy by creating a goal G about a state change and refining S_{task}, G, and π' over time through experience. The `Novelty Behavior` creates new tasks. First, it makes a potential goal state G from the most recent state change, (s_1, a, s_2), with a

2 This goal construct is also used in prior work, [Breazeal05,Lockerd04]

representation, x, for each belief in s_1 that changed in $s_1 \rightarrow s_2$. Any percept that changed is an expectation, the rest are criteria (e.g., see Figure 4).

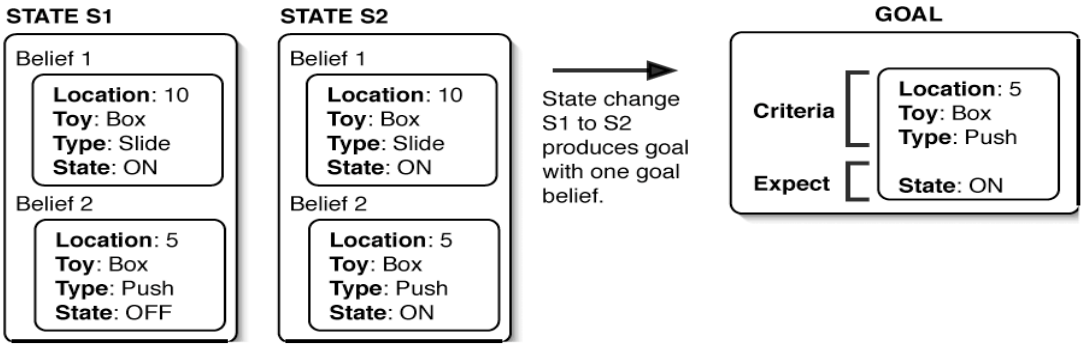

Figure 4. A simple example of creating a goal from a state change.

If there does not exist a $T \in Tasks$ with goal G, then a new Task Option Policy, T_{new}, is created. The S_{task} of T_{new} is initialized with the initiation state s_1, and π' is initialized with default values $q = .1$ for all actions from s_1. Then, the system takes into account the experience of (s_1, a, s_2), and (s_1, a) gets a higher value since s_2 is the goal.

Each $T \in Tasks$ can learn and expand from every experience (also referred to as intra-option learning [Sutton98]). Every action is an experience, (s_1, a, s_2); and each $T \in Tasks$ has the opportunity to extend its set S_{task} and update its π' based on this experience. To update π', rather than rely solely on external rewards from the environment, the system estimates the reward function based on the task's goal: $r = 1$ if the goal is true in s_2, otherwise $r = 0$.

3.5 Task Generalization

In addition to expanding initiation sets and updating value estimates for tasks, the system tries to generalize tasks over time. It works to generalize both the state representations in S_{task} and the goal representation G for all $T \in Tasks$.

Given two different tasks T_1 and T_2, the generalization mechanism attempts to combine them into a more general task T_{gen}. For example, if T_1 has the goal of turning ON a red button in location, `loc`(1, 2, 3), and T_2 has the goal of turning ON a red button in location, `loc`(4, 5, 6), then T_{gen} would have the goal of turning ON a red button without a location feature. When a feature is generalized from the goal, the system also tries to generalize the states in S_{task}, letting the task ignore that feature. Thus, T_{gen} can initiate in any location and any state with a red button ON achieves its goal.

This generalization is attempted each time a T_{new} is added to *Tasks*. If there exist two tasks T_1 and T_2 with similar goal states, then the system makes a general version of this task. Two goals are similar if they differ by no more than four percepts. In generalizing S_{task} and G for all $T \in Tasks$, the generalization mechanism expands the portion of the state space in which tasks can be initiated or considered achieved. This results in an efficient representation, as the system continually makes the state space representations more compact. Additionally, it is a goal-oriented approach to domain transfer, as the system is continually refining the *context* and the *goal* aspects of the activity representation.

In our red button example, the two tasks are similar since their expectations are the same, *expt* = {ON}, and their criteria differ only by the location feature. A new task is made with a goal that does not include location: G_{gen} = {expt = {ON}; crit = {`object,red,button`,....}} If the policies of the two tasks are similar, for example to do the `Press Action` in the state s ={b_1={`object,red,button,loc` = (x, y, z),...}}, then the new task will generalize location from all of S_{task}. On the other hand, if T_1 has the policy of doing the press action in state s={b_1={`object,red,button,loc`=(1, 2, 3),...}}, and T_2 has the policy of doing the flip action in state s={b_1={`object,red,button,loc`=(4, 5, 6),...}}, then the generalized task policy will maintain that in `loc`(1, 2, 3) a red button should be pressed to make it ON and in `loc`(4, 5, 6) a red button should be flipped to make it on.

3.6 Transparency Mechanisms

Leo has several expressive skills contributing to the robot's effectiveness as a social learner. Many are designed around theories of human joint activity [Clark96]. For example, consider principles of grounding. In general, humans look for evidence that their action has succeeded. This extends to joint activity where the ability to establish a mutual belief that a joint activity has succeeded is fundamental to a successful collaborative activity.

Table 1 highlights many of the social cues that Leo uses to facilitate the collaborative activity of learning. Eye gaze establishes joint attention, reassuring the teacher that the robot is attending to the right object. Subtle nods acknowledge task stages, e.g., confirming when the teacher labels a task goal.

Table 1. Social Cues for Transparency in a Socially Guided Exploration

Context	**Robot Behavior**	**Intention**
Human points to object	Looks at object	Shows object of attention
Human present in workspace	Gaze follows human	Shows social engagement
Executing an action	Looks at object	Shows object of attention
Human says "Look Leo, it's `TASK-X`"	Subtle head non and happy facial expression	Confirms goal state of task, `TASK-X`
Human says "Try to `ACT-Y` the `OBJ-Z`"	Look to human if suggestion is taken	Acknowledge partner's suggestion to perform specified action, `ACT-Y` on specified object, `OBJ-Z`
Speech did not parse; Unknown object request; Label without pointing gesture	Confusion expression	Communicates problem
Unconfident task execution	Glances to human more	Conveys uncertainly
Task is done and human says "Good"	Nods head	Positive feedback for current option
Human asks a yes/no question	Head Nod/shake	Communicates knowledge/ability
Intermittent	Eye blink, gaze shifts, posture shift	Conveys awareness and aliveness
Novel event	Surprise expression	Task model is created
Mastery triggers a task execution	Concentration expression	A known task is attempted

Completion of a successful task attempt	Happy expression	Expectation met
Completion of a failed task attempt	Sad expression	Expectation broken
Positive/Negative feedback from a human	Happy/Sad expression	Acknowledge type of feedback

Additionally, Leo uses its face for subtle expressions about the learning state. The robot's facial expression shifts to a particular pose for fleeting moments (2-3 seconds), indicating a state that pertains to its internal learning process, and then returns to a neutral pose. The expressions are chosen to communicate information to the human partner. They are inspired by research showing that different facial action units communicate specific meanings [Smith97] (Figure 5). For example, raised eyebrows and wide eyes indicate heightened attention, which is the desired communicative intent with Leo's surprised expression. This approach results in a dynamic and informative facial behavior.

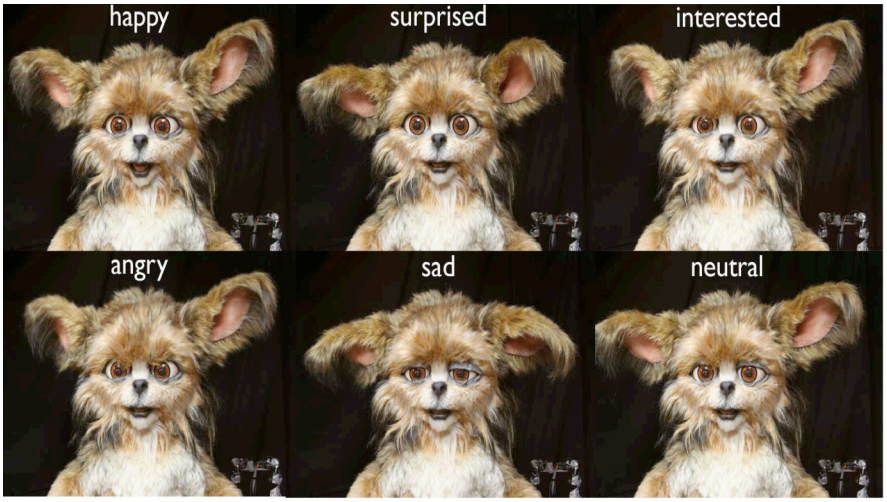

Figure 5. Leo can use several facial poses to express internal learning state.

Leonardo also communicates various learning contexts to the human partner with its facial expressions (Table 1). When the `Novelty Behavior` is triggered, a fleeting surprised expression lets the human know that a task is being created. When the `Mastery Behavior` causes a task to be practiced, Leo makes a concentrated facial expression and later a happy/sad expression upon the success/failure of the attempt. Throughout, if the human gives positive or negative feedback, Leo makes a happy or sad expression to acknowledge this feedback. When the human labels a goal state, Leonardo makes a happy expression and a head nod to acknowledge the labeling.

3.7 Scaffolding Mechanisms

The goal of our approach is for a robot learner to strike a balance between learning on it's own and benefiting from the social environment. The following are social scaffolding mechanisms at work on the Leonardo platform to enable Socially Guided Exploration.

Social attention: The attention of the robot is directed in ways that are intuitive for the human. Attention responds to socially salient stimuli and stimuli that are relevant to the current task. The robot tracks the pointing gestures and head pose of a human partner, which contribute to the saliency of objects and their likelihood for attention direction. For details on the robot's social attention system see [Thomaz05].

Guidance: Throughout the interaction, the human can suggest actions for Leo to try. The human's request is treated as a suggestion rather than an interrupt. The suggestion increases the likelihood that the `Explore Behavior` will trigger, but there is still some probability that Leo will decide to practice a relevant task or learn about a novel state change.

Recognizing goal states: Leo creates task representations of novelties in the environment. The human can facilitate this process by pointing out goal states with a variety of speech utterances (e.g., "Look Leo, it's `X`"). This serves to increase the likelihood that the `Novelty Behavior` will trigger, creating a task with the label "`X`".

Environmental structure: An implicit contribution of the human teacher is their ability to physically structure the learning environment, highlighting salient elements. They draw the robot learning system into new generalizations, link old information to new situations, and point out when a learned task is relevant in the current situation.

4. Experiment

To evaluate our Socially Guided Exploration system, we conducted a human subjects experiment where subjects interacted with the Leonardo robot. We solicited participation from the campus community, and had 11 participants complete the experiment over the course of two days (5 male, 6 female). Due to corrupted log files for two subjects, we only use data from 9 of the subjects in our analysis that depends on those log files. For the video analysis, we use the data from all 11 subjects.

4.1 Experimental Scenario

The experimental scenario is a shared workspace where Leo has a "smart" puzzle box (Figure 1). The puzzle box has three inputs (a switch, a slider, and a button), a lid that can open and close by activating an internal motor, five colored LEDs, and sound output. The box can be programmed with specific behaviors in response to actions on the input devices (e.g., the actions required to open the lid, or turn a colored LED on, etc.).

Leo has five primitive manual actions it can apply to the box (`Button-Press, Slider-Left, Slider-Right, Switch-Left, Switch-Right`), but no initial knowledge about the effects of these actions on the puzzle box. Leo uses the Socially Guided Exploration mechanism to build a *Tasks* set about the puzzle box.

In our experiment, the puzzle box is pre-programmed with the following input-output behavior:

- Pressing the button toggles through the five LED colors: white, red, yellow, green, and blue.
- If both the slider and the switch are flipped to the left when the color is white, then the box lid opens.
- If the slider and switch are flipped to the right when the color is yellow, then the box lid closes.
- If the lid is open and the color changes to blue, then the box will play a song.

4.2 Instructions to Human Subjects

Subjects are shown the functionality of the puzzle box and told that their goal is to help Leo learn about it. They are told the robot is able to do some simple actions on the toy puzzle box, and once turned on, it will start exploring what it can do with the box. Then the scaffolding mechanisms are explained. They were told they can help Leo learn tasks by making action suggestions, by naming aspects of the box, and by testing that these named aspects have been learned. The subjects were told that Leo understands the following kinds of utterances:

- "Leo, try to...[press the button, move the slider left/right, move the switch left/right]."
- "Look Leo, It's...[Open, Closed, A Song, Blue, White, Green, Red, Yellow]."
- "Leo, Try to make it...[Open, Closed, Play a Song, Blue, White, Green, Red, Yellow]."
- "Good Leo", "Good job", "Well done", "No", "Not quite."

Finally, the goal in this interaction was to make sure that Leo learns to do three things in particular:

- T_{Blue}--Make the light blue;
- T_{Open}--Make the lid open;
- T_{Song}--Make the song play.

5. Evaluation

5.1 Analysis of Guided Exploration versus Self Exploration

This first set of analyses examines how the teacher's social scaffolding influenced the robot's learning process. We compare data from the learning sessions in two conditions:

- GUIDED: The robot learns with a human teacher. As mentioned above, we have data from 9 participants in this condition.
- SELF: The robot learns by itself. For this condition, we collected data from 10 sessions of the Leonardo robot learning alone in the same environment.

All 9 participants succeeded in getting the robot to reach the T_{Blue} and T_{Open} tasks, but only four of the participants taught Leo the more complex T_{Song}. Everyone taught the T_{Blue} task first, and there was an average of 9 actions between first encountering the T_{Blue} and T_{Open} goals.

During the learning session, we logged several measures to analyze the effect of guidance on the learning process. In addition to collecting metrics during the learning session, the efficacy of the learned task sets was tested in simulation afterwards (detailed below). The differences between the Self Exploration and Socially Guided Exploration cases are summarized in Table 2.

Table 2. Summary of differences found between Guided exploration and Self exploration.

Measure	Means		1-tailed T-Tests	
	SELF	**GUIDE**	**t(19)**	**p**
Number actions to reach first goal in learning session	11.2	3.56	2.11	< .05
Size of resulting *Tasks* set	10.4	7.55	7.18	< .001
Number tasks for T_{Blue}	0.833	1.333	-2.58	< .01
Number tasks for T_{Open}	1	1.77	-1.83	< .05
Number Init States can reach T_{Open}	0.58	1.56	-2.88	< .01
Number actions to reach T_{Blue}	2.66	1.69	2.19	< .05

We found that the human teacher is able to guide the robot to the desired goal states *faster* than it can discover them on its own. This is seen in the difference between groups in the number of actions to the first encounter of any of the three experiment goal states. The average for GUIDE, 3.56, is significantly less than the average for the SELF condition, 11.2. Thus, people were able to utilize the social scaffolding mechanisms to focus the robot on aspects of the environment that they wanted it to learn. This is also supported by qualities of the resulting *Tasks* set that is learned. In the GUIDE condition, the resulting *Tasks* sets were *more related* to the experiment goals (i.e., T_{Blue}, T_{Open} or T_{Song} is true in a task's goal state). We see a significant difference in both the number of tasks related to T_{Blue} and T_{Open} (see Table 2).

Also, we found that the Socially Guided Exploration case *learns a better task set* for achieving the experiment goals. In the post analysis of the learned tasks, we tested each task set from a test suite of five initial states, looking for their ability to achieve the experimental goals. Each experiment goal has a different test suite of five initial states: three of which are very close to the goal (1 or 2 actions required), two of which are farther away (more than 2 actions required to achieve the goal). For each of the learned *Tasks* sets, we record the number of actions needed to reach each of the experimental task goals from each of the test states. We found some significant differences in the generality of the learned tasks. The average number of states that the GUIDE condition sets could reach the T_{Open} goal, 1.56, was significantly better than the average in the SELF condition, 0.58. And though we didn't find this particular difference for the T_{Blue} goal, we do see that the GUIDE condition is significantly faster at achieving T_{Blue} in the post analysis than the SELF condition, 1.69 versus 2.66. Thus, human guidance leads to task sets that are better at achieving the designated experimental goals.

5.2 Analysis of Human Scaffolding Behavior

Having learned about how human scaffolding changes the nature of what is learned during an exploration session, our next set of analyses focuses on understanding how people used the scaffolding mechanisms provided. We have video from each of the learning sessions, and we coded a transcript from each video that summarizes the following:

For each of the human's utterances, we coded the type of scaffolding: *suggestion, task label, task test, positive* or *negative feedback*. We also coded for three types of context for each utterance.

- Context 1: Did the person wait for the robot to make eye contact before they made the utterance?
- Context 2: Did the utterance happen after the robot made a facial expression?
- Context 3: Did the utterance happen while the robot was completing an action?

The transcript includes a recording of each action made by the robot, as well as each emotional expression. The camera was placed such that it was difficult to see every small facial expression made by the human, but we recorded all visible and audible emotional expressions (smiles, laughs, etc.).

Table 3. Relative amounts of the various scaffolding utterances by each participant.

Subject	Suggestions	Task Labels	Task Tests	Pos. Feedback	Neg. Feedback
1	44	8	7	16	10
2	49	26	18	30	0
3	82	36	15	34	23
4	7	2	2	3	0
5	75	31	8	11	9
6	35	10	14	4	4
7	40	20	5	20	9
8	14	7	4	6	4
9	32	15	12	6	2
10	29	14	3	15	4
11	33	16	1	8	0

Table 3 summarizes the frequency with which each of the subjects used each type of utterance. After normalizing by length of the learning session, we calculate the average frequency (i.e., number of utterances per action) for each scaffolding type. Interestingly, we found there was little variance in these frequencies across the participants.

- Action suggestions: average = 0.85, variance = 0.038
- Task labels: average = 0.36, variance = 0.022
- Task tests: average = 0.17, variance = 0.007
- Positive feedback: average = 0.31, variance = 0.027
- Negative feedback: average = 0.11, variance = 0.001

In addition to a characterization of how much and what kinds of scaffolding people used, we also looked at whether each type of scaffolding utterance happened in a particular context. Table 4 shows the average percentages for each of the three contexts for each type of scaffolding utterance. We see that nearly all (97%) task tests happen after eye contact, i.e., Context 1. Action suggestions are similar --- 89% are in Context 1. With task labels, 88% happen in Context 1. However, there is some tendency (14%) toward Context 3 where people label a state during Leo's action and before Leo looks up.

Table 4. Summary of the contexts of each type of scaffolding utterance.

Scaffolding	Context 1: eye contact		Context 2: expression		Context 3: action	
	average	variance	average	variance	average	variance
suggestions	0.892	0.011	0.084	0.003	0.075	0.007
task labels	0.879	0.012	0.051	0.003	0.136	0.027
task tests	0.971	0.006	0.049	0.008	0.000	0.000
positive feedback	0.469	0.049	0.050	0.005	0.582	0.059
negative feedback	0.538	0.155	0.000	0.000	0.631	0.091

Feedback is the most varied in terms of context, and like labeling it is seen in both Contexts 1 and 3, but people are more diverse in their behavior. Two of the 11 people issued most of their feedback (more than 65%) in Context 1; six people did most of their feedback in Context 3; and three people split their utterances nearly 50/50 between Contexts 1 and 3. In addition to the context of feedback, we looked at the relative amounts of positive and negative utterances (Figure 6). Again we have diversity among the 11 participants. We see that 3 people gave only positive feedback, 6 people had a positive bias to their feedback, and 2 people had fairly even amounts of positive and negative feedback. It is interesting that none of our subjects had a negative feedback bias.

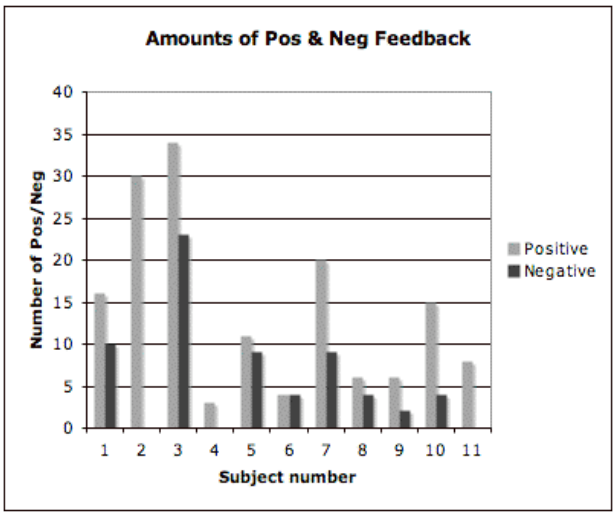

Figure 6. Relative amounts of positive and negative feedback issued by our human subjects.

The final question we looked at in the video analysis was "How often do people mirror the expressions/affect that Leo displayed?" Table 5 summarizes this data. There was a fairly wide range of matching behavior. Two people never did any visible or audible mirroring, while one person matched nearly 50% of Leo's expressions. On average people matched Leo's emotional expression 24% of the time.

Table 5. Summary of each person's mirroring behavior			
Subject	**Total amount of Expressions by Leo**	**Number of expressions that were matched**	**ratio**
1	10	0	0.000
2	16	2	0.125
3	13	3	0.231
4	4	0	0.000
5	14	5	0.357
6	17	8	0.471
7	12	4	0.333
8	7	1	0.143
9	10	4	0.400
10	6	2	0.333
11	9	2	0.222
Average			0.238

6. Discussion

In designing robotic agents that learn new skills and tasks "on the job" from everyday people, we recognize that the average person is not familiar with machine learning techniques, but they are intimately familiar with various forms of social learning (e.g., tutelage, imitation, etc.). This raises important research questions. For instance, "How do we design robots that learn effectively from human guidance?"; "What does human teaching behavior look like when interacting with a social robot?"; "How do we design robot learners to support human teaching behavior?" Also, there still remains the more traditional robot learning question, "How can robots be effective learners when a human teacher is not available?"

In prior works that incorporate a human into a machine learning process, the level of human interaction generally stays constant, remaining at one end of the guidance-exploration spectrum. Some are more guidance oriented, completely dependent on a human instruction. Others are more exploration based, using limited input from a teacher. In this work, we recognize that a social learner needs both, and the Socially Guided Exploration mechanism brings these together in one learning system. Motivations drive exploration of the environment and the creation of goal-oriented tasks about novel events. A human partner can influence learning through typical scaffolding acts such as directing attention, suggesting actions, highlighting and labeling goal states as interesting states to learn to achieve, testing task knowledge, and issuing positive/negative feedback.

Our experiments show that the Socially Guided Exploration mechanism is successful in allowing non-expert human teachers to guide the robot's learning process. People were able to focus the robot's learning to particular goals that they desired. And compared to self-learning in the same environment, the learning of these goals is accelerated, and the resulting representation of these tasks is more useful at a later time. The task sets resulting from guidance are smaller and more closely related to the specific tasks that the human was trying to teach. In self-learning on the other hand, the robot learned a broader task set, serendipitously learning aspects of the environment that the human was not focused on teaching. While not what the human had in mind today, this knowledge about the environment could be advantageous in the future. Clearly both types of learning are beneficial to a robot learner in different ways, supporting our approach of covering the full guidance-exploration spectrum.

In addition to illustrating the differences between guided and self learning, our experiment allows us to further explore a question that we have raised in prior work: "How do people naturally approach the task of teaching a machine?"

Our video analysis of the learning sessions lets us characterize key similarities and differences in how people use social scaffolding to help teach a physically embodied learner. First, we found that people exhibit consistent behavior in the relative frequency of the different types of scaffolding mechanisms available. Additionally we were able to learn something about the typical context for each of the scaffolding mechanisms. When making action suggestions or asking Leo to try to complete a task, people generally wait for eye contact. Presumably waiting for a signal that the robot is finished with its current action and ready to move on. Labeling a state (e.g., "Look, it's Blue") mostly happens after eye contact as well, but also happens during an action. Sometimes people want to give the label right as the state change happens. Feedback has an even greater split between the eye contact and action contexts. Either a feedback utterance is given right as an action is happening or the person waits until after the action completes and the robot looks up. This raises an important question for a social learning agent. Does a state label or a feedback utterance take on a different connotation when it is given in a different context? It is possible that a person means something different by an utterance given during an action versus one given at completion.

There is some anecdotal evidence that people have different interpretations of how task labeling should work. The current system assumes that the human might provide a label, and that it would pertain to the current state. Most participants did label in this way, but at least one participant gave 'pre-labels' for a given task. Saying, "Leo, now let's make it Blue." This is an interaction that the system is currently not designed to handle. Other people gave multiple labels for a state ("Look Leo, it's open, and it's green, and the switch is to the right..."). Over half of the participants did this multiple labeling behavior at least once. Again, the system is not designed to take advantage of this, but it is an interesting area for future work. These multiple labels could help the system more quickly learn to differentiate and generalize when a task is considered achieved.

The findings in this study support our previous data with teachable game characters regarding human feedback behavior. Previously, we studied people's interactions with a virtual robot game character that learns via interactive reinforcement learning [Thomaz07]. We found that people had a variety of intentions (guidance, motivation) that they communicated in addition to instrumental positive or negative feedback about the last action performed. We found a positive bias in the feedback an agent gets from a human teacher. Additionally, we showed that this asymmetry has a purpose. People mean qualitatively different things with positive and negative feedback. For instance, positive feedback was used to reinforce behavior but also to

motivate the agent. Negative feedback was used to "punish" but also to communicate, "undo and back up to your previous state."

In the study presented here we see more evidence of the varied nature of positive and negative feedback from a human partner. The split contexts of the feedback messages are an interesting area for future study. It is likely that the feedback takes on a different meaning dependent on the context. Again, we see a positive bias in people's feedback with 9 out of 11 people using more positive utterances (and in three of those cases the person only gave positive feedback).

The following is an interesting anecdote that highlights the complexity of feedback from a human partner. During a particular learning session, one teacher made the action suggestion to move the switch left when the box switch was in the left position. Leo did the suggested action, but since it was already left the action had no effect. The teacher gave positive feedback anyway, and then quickly corrected herself and suggested switch right. The true meaning of this positive feedback message is, "yes, you did what I asked, good job, but what I told you was wrong..." Thus, positive and negative feedback from a human partner is much more nuanced than a simple good/bad signal from the environment, and an embodied social learning agent will need the ability to discern these subtle meanings.

A final topic addressed this study is the extent to which the behavior of the robot influences the human teacher. In prior work, we showed that a virtual robot game character *can* influence the input from a human teacher with a simple gazing behavior [Thomaz06]. An embodied robotic agent like Leonardo has many more subtle ways in which to communicate its internal state to the human partner, and we see some evidence that people's behavior is influenced by the social cues of the robot. On average about 25% of the robot's facial expressions were mirrored by the human either with their own facial expression, tone of voice, or with a feedback utterance. Also, people waited for Leonardo's to make eye contact with them before they would say the next utterance. This has the nice property of a subtle cue the robot uses to slow down the human's input until the Leo is ready for it. In the future, one could imagine exploiting mutual gaze to elicit additional input "just in time." For instance, the robot might initiate and action, pause and look to the human if confidence is low, to elicit a confirmation or additional guidance before it executes the action.

We see additional anecdotal evidence of people shifting their teaching strategies based on the behavior of the robot. In one case, the person misunderstood the instructions and initially tried to demonstrate the task instead of guide an exploration. She would label a state, describe her actions, and then label the new state. But she quickly shifted into the guided exploration (after about four actions) once Leo started doing actions itself. In another case, the teacher's strategy was to 'pre-label' a task. She would say, "Leo's let's make it Blue", and then make the necessary action suggestions. Once Leo got to the desired state she'd say, "Good Job!" But she did not say the name of the state once they got there, so the label never got attached to that task representation. Then she would ask Leo to make it blue, and Leo would not know the name of that task. Finally, she did one post-labeling, saying the name of the task "Blue" after it was completed, and Leo demonstrated that he could do the blue task soon afterwards. At this point she stopped pre-labeling, and only did the post-labeling for the rest of the learning session.

7. Conclusion

This work acknowledges that a robot learning in a social environment needs the ability to both learn on its own and to take advantage of the social structure provided by a human partner. Our Socially Guided Exploration learning mechanism has motivations to explore its environment and is able to create goal-oriented

task representations of novel events. Additionally this process can be influenced by a human partner through attention direction, action suggestion, labeling goal states, and feedback using natural social cues. From our experiments, we found beneficial properties of the balance between intrinsically motivated learning and socially guided learning. Namely, self-exploration tended to result in a broader task repertoire from serendipitous learning opportunities. This broad task set can help to scaffold future learning with a human teacher. Guided-exploration with a human teacher tended to be more goal-driven, resulting in fewer tasks that were learned faster and generalized better to new starting states.

Our analysis of human teaching behavior revealed some interesting findings. First, we found that there was surprisingly little variance among human subjects with respect to how often they used specific types of scaffolding (action suggestions being the highest, negative feedback was the least). Our video analysis reveals different forms of behavior coupling between human teacher and robot learner through social cues. We found that most scaffolding was given to the robot after it made eye contact with the teacher. We also found that human teachers tended to mirror the expressive behavior of the robot (an average of 25%), but this varied by teaching style (some did not mirror at all, some mirrored more than 40%). In addition, we found that the communicative intent behind positive and negative feedback is subtle and varied – it is used in different contexts, sometimes before the robot takes action. Hence, it is not simply reinforcement of past actions. We also found that different teachers have different styles in how they use feedback – some have a positive bias, others are more balanced. Interestingly, none of our subjects had a negative bias.

These findings inform and motivate continued work in how to design robots that learn from human teachers with respect to the dynamic social coupling of teacher and learner to coordinate and improve the teaching/learning process, designing to support the frequency and kinds of scaffolding, understanding the subtlety of intention behind positive/negative feedback, and accommodating different teaching styles.

Acknowledgements

The work presented in this paper is a part of ongoing work of the graduate and undergraduate students in the Personal Robotics Group of the MIT Media Lab. The Leonardo robot was initially designed in collaboration with Stan Winston Studios and has been under development since 2002. This work is funded by the *Digital Life* and *Things That Think* consortia of the Media Lab, and in particular by the Toyota Motor Corporation.

References

[Atkeson97] Christopher G. Atkeson and Stefan Schaal. Robot learning from demonstration. In Proc. 14th International Conference on Machine Learning, pages 12–20. Morgan Kaufmann, 1997.

[Blumberg02] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M.P. Johnson, and B. Tomlinson. Integrated learning for interactive synthetic characters. In Proceedings of the ACM SIGGRAPH, 2002.

[Breazeal02] C. Breazeal. Designing Sociable Robots. MIT Press, Cambridge, MA, 2002.

[Breazeal05] C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. L. Thomaz. Using perspective taking to learn from ambiguous demonstrations. to appear in the Journal of Robotics and Autonomous Systems Special Issue on Robot Programming by Demonstration, 2005.

[Chernova07] S. Chernova and M. Veloso. Confidence-based policy learning from demonstration using gaussian mixture models. In Proc. of Autonomous Agents and Multi-Agent Systems (AAMAS), 2007.

[Clark96] H. H. Clark. Using Language. Cambridge University Press, Cam- bridge, 1996.

[Clouse92] J. Clouse and P. Utgoff. A teaching method for reinforcement learning. In Proc. of the Nineth International Conf. on Machine Learning (ICML), pages 92–101, 1992.

[Kaplan02] F. Kaplan, P-Y. Oudeyer, E. Kubinyi, and A. Miklosi. Robotic clicker training. Robotics and Autonomous Systems, 38(3-4):197–206, 2002.

[Vygotsky78] Ed. M. Cole L. S. Vygotsky. Mind in society: the development of higher psychological processes. Harvard University Press, Cambridge, MA, 1978.

[Lockerd04] A. Lockerd and C. Breazeal. Tutelage and socially guided robot learning. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2004.

[Maclin05] R. Maclin, J. Shavlik, L. Torrey, T. Walker, and E. Wild. Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression. In Proceedings of the The Twentieth National Conference on Artificial Intelligence (AAAI), Pittsburgh, PA, July 2005.

[Morency02] L-P. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast stereo-based head tracking for interactive environment. In Int. Conference on Automatic Face and Gesture Recognition, 2002.

[Nicolescu03] M. N. Nicolescu and M. J. Matari ´c. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In Proceedings of the 2nd Intl. Conf. AAMAS, Melbourne, Australia, July 2003.

[Oudeyer04] P-Y. Oudeyer and F. Kaplan. Intelligent adaptive curiosity: a source of self-development. In Proceedings of the 4th International Workshop on Epigenetic Robotics, volume 117, pages 127–130, 2004.

[Saksida98] L. M. Saksida, S. M. Raymond, and D. S. Touretzky. Shaping robot behavior using principles from instrumental conditioning. Robotics and Autonomous Systems, 22(3/4):231, 1998.

[Schmidhuber05] J. Schmidhuber. Self-motivated development through rewards for predictor errors/improvements. In D. Blank and L. Meeden, editors, Proc. Developmental Robotics 2005 AAAI Spring Symposium, 2005.

[Singh05] S. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In Proceedings of Advances in Neural Information Processing Systems 17 (NIPS), 2005.

[Smart02] W. D. Smart and L. P. Kaelbling. Effective reinforcement learning for mobile robots. In In Proceedings of the IEEE International Conference on Robotics and Automation, pages 3404–3410, 2002.

[Smith97] C. Smith and H. Scott. A componential approach to the meaning of facial expressions. In The Psychology of Facial Expression. Cambridge University Press, United Kingdom, 1997.

[Sutton98] R. Sutton, D. Precup, and S. Singh. Intra-option learning about temporally abstract actions. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML98), Masion, WI, 1998.

[Sutton99] R. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: Learning, planning and representing knowledge at multiple temporal scales. Journal of Artificial Intelligence Research, 1:139, 1999.

[Thomaz05] A. L. Thomaz, M. Berlin, and C. Breazeal. An embodied computational model of social referencing. In IEEE International Workshop on Human Robot Interaction (RO-MAN), 2005.

[Thomaz06] A. L. Thomaz and C. Breazeal. Transparency and Socially Guided Machine Learning. ICDL 2006.

[Thomaz07] A. L. Thomaz and C. Breazeal. Teachable Robots: Understanding human teaching behavior to build more effective robot learners, Artificial Intelligence Journal (AIJ). In Press.