Perspective Taking: An Organizing Principle for Learning in Human-Robot Interaction

Matt Berlin, Jesse Gray, Andrea L. Thomaz, Cynthia Breazeal
MIT Media Lab
20 Ames St., Cambridge, MA 02139
{mattb, jg, alockerd, cynthiab}@media.mit.edu

Abstract

The ability to interpret demonstrations from the perspective of the teacher plays a critical role in human learning. Robotic systems that aim to learn effectively from human teachers must similarly be able to engage in perspective taking. We present an integrated architecture wherein the robot's cognitive functionality is organized around the ability to understand the environment from the perspective of a social partner as well as its own. The performance of this architecture on a set of learning tasks is evaluated against human data derived from a novel study examining the importance of perspective taking in human learning. Perspective taking, both in humans and in our architecture, focuses the agent's attention on the subset of the problem space that is important to the teacher. This constrained attention allows the agent to overcome ambiguity and incompleteness that can often be present in human demonstrations and thus learn what the teacher intends to teach.

Introduction

This paper addresses an important issue in building robots that can successfully learn from demonstrations that are provided by "naïve" human teachers who do not have expertise in the learning algorithms used by the robot. As a result, the teacher may provide sensible demonstrations from a human's perspective; however, these same demonstrations may be insufficient, incomplete, ambiguous, or otherwise "flawed" in terms of providing a correct and sufficiently complete training set in order for the learning algorithm to generalize properly.

To address this issue, we believe that socially situated robots will need to be designed as socially cognitive learners that can infer the intention of the human's instruction, even if the teacher's demonstrations are less than perfect for the robot. Our approach to endowing machines with socially-cognitive learning abilities is inspired by leading psychological theories and recent neuroscientific evidence for how human brains might infer the mental states of others. Specifically, *Simulation Theory* holds that certain parts of the brain have dual use; they are used to not only generate behavior and mental states, but also to predict and infer the same in others (Davies & Stone 1995; Barsalou *et al.* 2003; Sebanz, Bekkering, & Knoblich 2006).

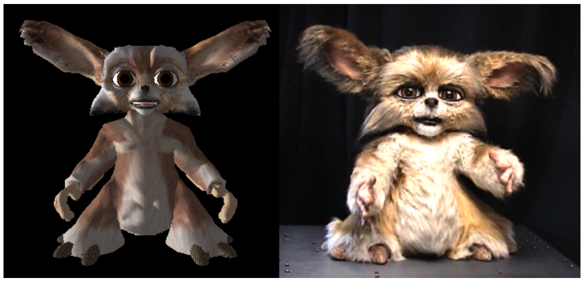

Figure 1: The Leonardo robot and graphical simulator

In this paper, we present an integrated architecture wherein the robot's cognitive functionality is organized around the ability to understand the environment from the perspective of the teacher using simulation-theoretic mechanisms. Perspective taking focuses the agent's attention on the subset of the problem space that is important to the teacher. Focusing on a subset of the input/problem space directly affects the set of hypotheses entertained by the learning algorithm, and thus directly affects the skill transferred to the agent via the interaction with the teacher. This constrained attention allows the agent to overcome ambiguity and incompleteness that can often be present in human demonstrations.

The outline of this paper is as follows. First, we present an overview of our integrated architecture which runs on a 65 degree of freedom humanoid robot and its graphical simulator (Fig. 1). We then detail the perceptual-belief pipeline, describe our tutelage-inspired approach to task learning, and present how perspective taking integrates with these mechanisms. This architectural integration allows the robot to infer the goal and belief states of the teacher, and thus to more accurately model the intent of their demonstrations. Finally, we present the results of a novel study examining the importance of perspective taking in human learning. Data derived from the suite of learning tasks in this study are used to create a benchmark suite to evaluate the performance of our architecture and to illustrate the robot's ability to learn in a human compatible way from ambiguous demonstrations presented by a human teacher.

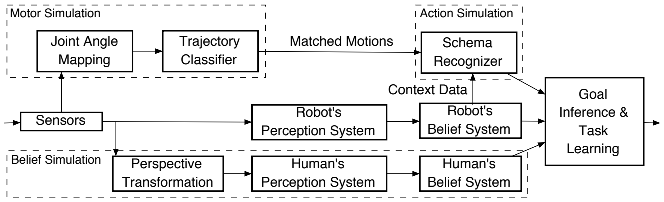

Figure 2: System Architecture Overview

Architecture Overview

Our architecture, based on (Blumberg *et al.* 2002), incorporates simulation-theoretic mechanisms as a foundational and organizational principle to support collaborative forms of human-robot interaction, such as the tutelage-based learning examined in this paper. An overview of the architecture is shown in Figure 2. Our implementation enables a humanoid robot to monitor an adjacent human teacher by simulating his or her behavior within the robot's own generative mechanisms on the motor, goal-directed action, and perceptual-belief levels (Gray, J. *et al.* 2005).

While others have identified that visual perspective taking coupled with spatial reasoning are critical for human-robot collaboration on a shared task within a physical space (Trafton *et al.* 2005), and collaborative dialog systems have investigated the role of plan recognition in identifying and resolving misconceptions (see (Carberry 2001) for a review), this is the first work to examine the role of perspective taking for introceptive states (e.g., beliefs and goals) in a human-robot learning task.

Belief Modeling

In order to convey how the robot interprets the environment from the teacher's perspective, we must first describe how the robot understands the world from its own perspective. This section presents a technical description of two important components of our cognitive architecture: the Perception System and the Belief System. The Perception System is responsible for extracting perceptual features from raw sensory information, while the Belief System is responsible for integrating this information into discrete object representations. The Belief System represents our approach to sensor fusion, object tracking and persistence, and short-term memory.

On every time step, the robot receives a set of sensory observations $O = \{o_1, o_2, ..., o_N\}$ from its various sensory processes. Information is extracted from these observations by the Perception System. The Perception System consists of a set of *percepts* $P = \{p_1, p_2, ..., p_K\}$, where each $p \in P$ is a classification function defined such that

$$p(o) = (m, c, d), \tag{1}$$

where $m, c \in [0, 1]$ are match and confidence values and d is an optional derived feature value. For each observation $o_i \in O$, the Perception System produces a *percept snapshot*

$$s_i = \{(p, m, c, d) | p \in P, p(o_i) = (m, c, d), m * c > k\}, \tag{2}$$

where $k \in [0, 1]$ is a threshold value, typically 0.5.

These snapshots are then clustered into discrete object representations called *beliefs* by the Belief System. This clustering is typically based on the spatial relationships between the various observations, in conjunction with other metrics of similarity. The Belief System maintains a set of beliefs B, where each belief $b \in B$ is a set mapping percepts to history functions: $b = \{(p_x, h_x), (p_y, h_y), ...\}$. For each $(p, h) \in b$, h is a history function defined such that

$$h(t) = (m'_t, c'_t, d'_t) \tag{3}$$

represents the "remembered" evaluation for percept p at time t. History functions may be lossless, but they are often implemented using compression schemes such as low-pass filtering or logarithmic timescale memory structures.

A Belief System is fully described by the tuple (B, G, M, d, q, w, c), where

- B is the current set of beliefs,
- G is a generator function map, $G : P \rightarrow \mathcal{G}$, where each $g \in \mathcal{G}$ is a history generator function where $g(m, c, d) = h$ is a history function as above,
- M is the belief merge function, where $M(b_1, b_2) = b'$ represents the "merge" of the history information contained within b_1 and b_2,
- $d = d_1, d_2, ..., d_L$ is a vector of belief distance functions, $d_i : B \times B \rightarrow \mathcal{R}$,
- $q = q_1, q_2, ..., q_L$ is a vector of indicator functions where each element q_i denotes the applicability of d_i, $q_i : B \times B \rightarrow \{0, 1\}$,
- $w = w_1, w_2, ..., w_L$ is a vector of weights, $w_i \in \mathcal{R}$, and
- $c = c_1, c_2, ..., c_J$ is a vector of culling functions, $c_j : B \times B \rightarrow \{0, 1\}$.

Using the above, we define the Belief Distance Function, D, and the Belief Culling Function, C:

$$D(b_1, b_2) = \sum_{i=1}^{L} w_i q_i(b_1, b_2) d_i(b_1, b_2) \tag{4}$$

$$C(b) = \prod_{j=1}^{J} c_j(b) \tag{5}$$

The Belief System manages three key processes: creating new beliefs from incoming percept snapshots, merging these new beliefs into existing beliefs, and culling stale beliefs. For the first of these processes, we define the function N, which creates a new belief b_i from a percept snapshot s_i:

$$\begin{aligned} b_i = N(s_i) = \{(p, h) | &(p, m, c, d) \in s_i, \\ &g = G(p), h = g(m, c, d)\} \end{aligned} \tag{6}$$

For the second process, the Belief System merges new beliefs into existing ones by clustering proximal beliefs, assumed to represent different observations of the same object. This is accomplished via bottom-up, agglomerative clustering as follows.

For a set of beliefs B:

```
while $\exists b_x, b_y \in B$ such that $D(b_x, b_y) < thresh$ do
    find $b_1, b_2 \in B$ such that $D(b_1, b_2)$ is minimal
    $B \leftarrow B \cup \{M(b_1, b_2)\} \setminus \{b_1, b_2\}$
end while
```

Finally, the Belief System culls stale beliefs by removing all beliefs from the current set for which $C(b) = 1$.

Task and Goal Learning

We believe that flexible, goal-oriented, hierarchical task learning is imperative for learning in a collaborative setting from a human partner, due to the human's propensity to communicate in goal-oriented and intentional terms. Hence, we have a hierarchical, goal-oriented task representation, wherein a task is represented by a set, S, of schema hypotheses: one primary hypothesis and n others. A schema hypothesis has x executables, E, (each either a primitive action a or another schema), a goal, G, and a tally, c, of how many seen examples have been consistent with this hypothesis.

Goals for actions and schemas are a set of y goal *beliefs* about what must hold true in order to consider this schema or action achieved. A goal belief represents a desired change during the action or schema by grouping a belief's percepts into i criteria percepts (indicating features that holds constant over the action or schema) and j expectation percepts (indicating an expected feature change). This yields straightforward goal evaluation during execution: for each goal belief, all objects with the criteria features must match the expectation features.

Schema Representation:

$$S = \{[(E_1...E_x), G, c]_P, [(E_1...E_x), G, c]_{1...n}\}$$
$$E = a|S$$
$$G = \{B_1...B_y\}$$
$$B = p_{C_1}...p_{C_i} \cup p_{E_1}...p_{E_j}$$

For the purpose of task learning, the robot can take a snapshot of the world (i.e. the state of the Belief System) at time t, $Snp(t)$, in order to later reason about world state changes. Learning is mixed-initiative such that the robot pays attention to both its own and its partner's actions during a learning episode. When the learning process begins, the robot creates a new schema representation, S, and saves belief snapshot $Snp(t_0)$. From time, t_0, until the human indicates that the task is finished, t_{end}, if either the robot or the human completes an action, act, the robot makes an action representation, $a = [act, G]$ for S:

```
For action $act$ at time $t_b$ given last action at $t_a$
$G =$ belief changes from $Snp(t_a)$ to $Snp(t_b)$
append $[act, G]$ to executables of $S$
$t_a = t_b$
```

At time t_{end}, this same process works to infer the goal for the schema, S, making the goal inference from the differences in $Snp(t_0)$ and $Snp(t_{end})$. The goal inference mechanism notes all changes that occurred over the task; however, there may still be ambiguity around which aspects of the state change are the goal (the change to an object, a class of objects, the whole world state, etc.). Our approach uses hypothesis testing coupled with human interaction to disambiguate the overall task goal over a few examples.

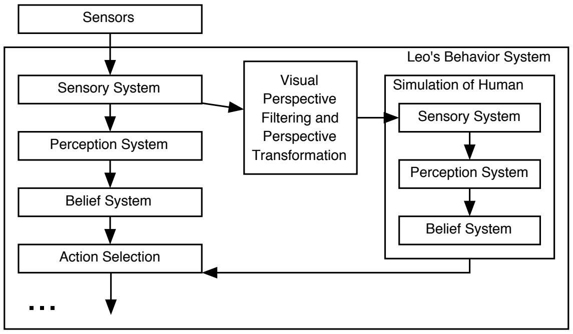

Figure 3: Architecture for modeling the human's beliefs reuses the robot's own architecture for belief maintenance.

Once the human indicates that the current task is done, S contains the representation of the seen example $([(E_1...E_x), G, 1])$. The system uses S to expand other hypotheses about the desired goal state to yield a hypothesis of all goal representations, G, consistent with the current demonstration (for details of this expansion process see (Lockerd & Breazeal 2004); to accommodate the tasks described here we additionally expand hypotheses whose goal is a state change across a simple disjunction of object classes). The current best schema candidate (the primary hypothesis) is chosen through a Bayesian likelihood method: $P(h|D) \propto P(D|h)P(h)$. The data, *D*, is the set of all examples seen for this task. $P(D|h)$ is the percentage of the examples in which the state change seen in the example is consistent with the goal representation in *h*. For priors, $P(h)$, hypotheses whose goal states apply to the broadest object classes with the most specific class descriptions are preferred (determined by number of classes and criteria/expectation features, respectively). Thus, when a task is first learned, every hypothesis schema is equally represented in the data, and the algorithm chooses the most specific schema for the next execution.

Perspective Taking

In this section, we describe how perspective taking integrates with the cognitive mechanisms discussed above: belief modeling and task learning. Inferring the beliefs of the teacher allows the robot to build task models which capture the intent behind human demonstrations.

Perspective Taking and Belief Inference

When demonstrating a task to be learned, it is important that the context within which that demonstration is performed be the same for the teacher as it is for the learner. However, in complex and dynamic environments, it is possible for the instructor's beliefs about the context surrounding the demonstration to diverge from those of the learner. For example, a visual occlusion could block the teacher's viewpoint of a region of a shared workspace (but not that of the learner) and consequently lead to ambiguous demonstrations where the teacher does not realize that the visual information of the scene differs between them.

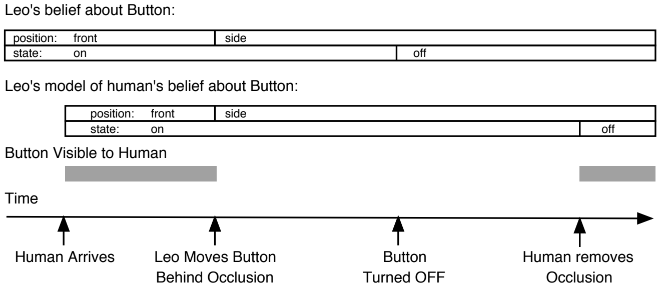

Figure 4: Timeline following the progress of the robot's beliefs for one button. The robot updates its belief about the button with any sensor data available - however, the robot only integrates new data into its model of the human's belief if the data is available when the human is able to perceive it.

To address this issue, the robot must establish and maintain mutual beliefs with the human instructor about the shared context surrounding demonstrations. The robot keeps track of its own beliefs about object state using its Belief System, described above. In order to model the beliefs of the human instructor as separate and potentially different from its own, the robot re-uses the mechanism of its own Belief System. These beliefs that represent the robot's model of the human's beliefs are in the same format as its own, but are maintained separately so the robot can compare differences between its beliefs and the human's beliefs.

As described above, belief maintenance consists of incorporating new sensor data into existing knowledge of the world. The robot's sensors are all in its reference frame, so objects in the world are perceived relative to the robot's position and orientation. In order to model the beliefs of the human, the robot re-uses the same mechanisms used for its own belief modeling, but first transforms the data into the reference frame of the human (see Fig. 3).

The robot can also filter out incoming data that it believes is not perceivable to the human, thereby preventing that new data from updating the model of the human's beliefs. As you recall, the sensory observations $O = \{o_1, o_2, ..., o_N\}$ are the input to the robot's belief system. The inputs to the secondary belief system that models the human's beliefs are O', where:

$$O' = \{P(o') | o' \in O, V(o') = 1\} \tag{7}$$

where:

$$V(x) = \begin{cases} 1 & \text{if } x \text{ is visible to human} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

and:

$$\begin{aligned} P :& \{\text{robot local observations}\} \\ & \rightarrow \{\text{person local observations}\} \end{aligned} \tag{9}$$

Visibility can be determined by a cone calculated from the human's position and orientation, and objects on the opposite side of known occlusions from the human can be marked invisible.

Maintaining this parallel set of beliefs is different from simply adding metadata to the robot's original beliefs because it reuses the entire architecture which has mechanisms for object permanence, history of properties, etc. This allows for a more sophisticated model of the human's beliefs. For instance, Fig. 4 shows an example where this approach keeps track of the human's incorrect beliefs about objects that have changed state while out of the human's view. This is important for establishing and maintaining mutual beliefs in time-varying situations where beliefs of individuals can diverge over time.

Perspective Taking and Task Learning

In a similar fashion, in order to model the task from the demonstrator's perspective, the robot runs a parallel copy of its task learning engine that operates on its simulated representation of the human's beliefs. In essence, this focuses the hypothesis generation mechanism on the subset of the input space that matters to the human teacher.

At the beginning of a learning episode, the robot can take a snapshot of the world in order to later reason about world state changes. The integration of perspective taking means that this snapshot can either be taken from the robot's (R) or the human's (H) belief perspective. Thus when the learning process begins, the robot creates two distinct schema representations, S_{Robot} and S_{Hum}, and saves belief snapshots $Snp(t_0, R)$ and $Snp(t_0, H)$. Learning proceeds as before, but operating on these two parallel schemas.

Once the human indicates that the current task is done, S_{Robot} and S_{Hum} both contain the representation of the seen example. Having been created from the same demonstration, the executables will be equivalent, but the goals may not be equal since they are from differing perspectives. Maintaining parallel schema representations gives the robot three options when faced with inconsistent goal hypotheses: assume that the human's schema is correct, assume that its own schema is correct, or attempt to resolve the conflicts between the schemas. Our evaluation in the following section focuses on the simplest approach: take the perspective of the teacher, and assume that their schema is correct.

Human Subjects Study

We conducted a human subjects study to evaluate our approach. The study had two purposes. First, to gather human performance data on a set of learning tasks that were well matched to our robot's existing perceptual and inferential capabilities, creating a benchmark suite for our perspective taking architecture. Second, the study served to highlight the role of perspective taking in human learning.

Study participants were asked to engage in four different learning tasks involving foam building blocks. We gathered data from 41 participants, divided into two groups. 20 participants observed demonstrations provided by a human teacher sitting opposite them (the social condition), while 21 participants were shown static images of the same demonstrations, with the teacher absent from the scene (the nonsocial condition). Participants were asked to show their understanding of the presented skill either by re-performing the skill on a novel set of blocks (in the social context) or by selecting the best matching image from a set of possible images (in the nonsocial context).

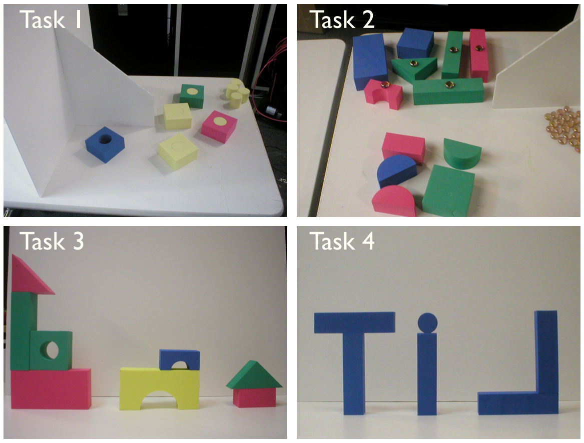

Figure 5: The four tasks demonstrated to participants in the study (photos taken from the participant's perspective). Tasks 1 and 2 were demonstrated twice with blocks in different configurations. Tasks 3 and 4 were demonstrated only once.

Table 1: Differential rule acquisition for study participants in social vs. nonsocial conditions. ***: $p < 0.001$

Task	Condition	PT Rule	NPT Rule	Other	p
Task 1	social	**6**	1	13	***
	nonsocial	1	**12**	8	
Task 2	social	**16**	0	4	***
	nonsocial	7	**12**	2	
Task 3	social	**12**	8	-	***
	nonsocial	0	**21**	-	
Task 4	social	**14**	6	-	***
	nonsocial	0	**21**	-	

Fig. 5 illustrates sample demonstrations of each of the four tasks. The tasks were designed to be highly ambiguous, providing the opportunity to investigate how different types of perspective taking might be used to resolve these ambiguities. The subjects' demonstrated rules can be divided into three categories: perspective taking (PT) rules, non-perspective taking (NPT) rules, and rules that did not clearly support either hypothesis (Other).

Task 1 focused on visual perspective taking during the demonstration. Participants were shown two demonstrations with blocks in different configurations. In both demonstrations, the teacher attempted to fill all of the holes in the square blocks with the available pegs. Critically, in both demonstrations, a blue block lay within clear view of the participant but was occluded from the view of the teacher by a barrier. The hole of this blue block was never filled by the teacher. Thus, an appropriate (NPT) rule might be "fill all but blue," or "fill all but this one," but if the teacher's perspective is taken into account, a more parsimonious (PT) rule might be "fill all of the holes" (see Fig. 6).

Task 2 focused on resource perspective taking during the demonstration. Again, participants were shown two demonstrations with blocks in different configurations. Various manipulations were performed to encourage the idea that some of the blocks "belonged" to the teacher, whereas the others "belonged" to the participant, including spatial separation in the arrangement of the two sets of blocks. In both demonstrations, the teacher placed markers on only "his" red and green blocks, ignoring his blue blocks and all of the participant's blocks. Because of the way that the blocks were arranged, however, the teacher's markers were only ever placed on triangular blocks, long, skinny, rectangular blocks, and bridge-shaped blocks, and marked all such blocks in the workspace. Thus, if the blocks' "ownership" is taken into account, a simple (PT) rule might be "mark only red and green blocks," but a more complicated (NPT) rule involving shape preference could account for the marking and non-marking of all of the blocks in the workspace (see Fig. 6).

Task 3 and 4 investigated whether or not visual perspective is factored into the understanding of task goals. In both tasks, participants were shown a single construction demonstration, and then were asked to construct "the same thing" using a similar set of blocks. Fig. 5 shows the examples that were constructed by the teacher. In both tasks, the teacher assembled the examples from left to right. In task 4, the teacher assembled the word "LiT" so that it read correctly from their own perspective. Our question was, would the participants rotate the demonstration (the PT rule) so that it read correctly for themselves, or would they mirror the figure (the NPT rule) so that it looked exactly the same as the demonstration (and thus read backwards from their perspective). Task 3, in which the teacher assembled a sequence of building-like forms, was essentially included as a control, to see if people would perform any such perspective flipping in a non-linguistic scenario.

The results of the study are summarized in Table 1 where participant behavior was recorded and classified according to the exhibited rule. For every task, differences in rule choice between the social and nonsocial conditions were highly significant (chi-square, $p < 0.001$). The most popular rule for each condition is highlighted in bold (note that, while many partic-

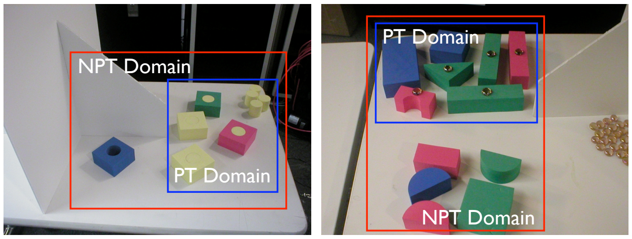

Figure 6: Input domains consistent with the perspective taking (PT) vs. non-perspective taking (NPT) hypotheses. In visual perspective taking (left image), the student's attention is focused on just the blocks that the teacher can see, excluding the occluded block. In resource perspective taking (right image), attention is focused on just the blocks that are considered to be "the teacher's," excluding the other blocks.

Table 2: High-likelihood hypotheses entertained by the robot at the conclusion of benchmark task demonstrations. The highest likelihood (winning) hypotheses are highlighted in bold.

Task	Condition	High-Likelihood Hypotheses
Task 1	with PT	***all****; all but blue*
	without PT	***all but blue***
Task 2	with PT	***all red and green****; shape preference*
	without PT	***shape preference***
Task 3 & 4	with PT	***rotate figure****; mirror figure*
	without PT	***mirror figure***

ipants fell into the "Other" category for Task 1, there was very little rule agreement between these participants). These results strongly support the intuition that perspective taking plays an important role in human learning in socially situated contexts.

Robot Evaluation and Discussion

The tasks from our study were used to create a benchmark suite for our architecture. In our graphical simulation environment, the robot was presented with the same task demonstrations as were provided to the study participants (Fig. 7). The learning performance of the robot was analyzed in two conditions: with the perspective taking mechanisms intact, and with them disabled.

The robot was instructed in real-time by a human teacher. The teacher delineated task demonstrations using verbal commands: "Leo, I can teach you to do task 1," "Task 1 is done," etc. The teacher could select and move graphical building blocks within the robot's 3D workspace via a mouse interface. This interface allowed the teacher to demonstrate a wide range of tasks involving complex block arrangements and dynamics. For our benchmark suite, the teacher followed the same task protocol that was used in the study, featuring identical block configurations and movements. For the purposes of perspective taking, the teacher's visual perspective was assumed to be that of the virtual camera through which the scene was rendered.

As the teacher manipulated the blocks, the robot attended to the teacher's movements. The robot's task learning mechanisms parsed these movements into discrete actions and assembled a schema representation for the task at hand, as detailed in previous sections. At the conclusion of each demonstration, the robot expanded and revised a set of hypotheses about the intended goal of the task. After the final demonstration, the robot was instructed to perform the task using a novel set of blocks arranged in accordance with the human study protocol. The robot's behavior was recorded, along with all of the task hypotheses considered to be valid by the robot's learning mechanism.

Table 2 shows the highest-likelihood hypotheses entertained by the robot in the various task conditions at the conclusion of the demonstrations. In the perspective taking condition, likely hypotheses included both those constructed from the teacher's perspective as well as those constructed from the robot's own perspective; however, as described above, the robot preferred hypotheses constructed from the teacher's perspective. The hypotheses favored by the learning mechanism (and thus executed by the robot) are highlighted in bold. For comparison, Table 3 displays the rules selected by study participants, with the most popular rules for each task highlighted in bold.

Table 3: Hypotheses selected by study participants following task demonstrations. The most popular rules are highlighted in bold.

Task	Condition	Hypotheses Selected
Task 1	social	***all****; number; spatial arrangement*
	nonsocial	***all but blue****; spatial arrangement; all but one*
Task 2	social	***all red and green****; shape preference; spatial arrangement*
	nonsocial	***shape preference****; all red and green*
Task 3 & 4	social	***rotate figure****; mirror figure*
	nonsocial	***mirror figure***

For every task and condition, the rule learned by the robot matches the most popular rule selected by the humans. This strongly suggests that the robot's perspective taking mechanisms focus its attention on a region of the input space similar to that attended to by study participants in the presence of a human teacher. It should also be noted, as evident in the tables, that participants generally seemed to entertain a more varied set of hypotheses than the robot. In particular, participants often demonstrated rules based on spatial or numeric relationships between the objects — relationships which are not yet represented by the robot. Thus, the differences in behavior between the humans and the robot can largely be understood as a difference in the scope of the relationships considered between the objects in the example space, rather than as a difference in this underlying space. The robot's perspective taking mechanisms are successful at bringing the agent's focus of attention into alignment with the humans' in the presence of a social teacher.

This is the first work to examine the role of perspective taking for introceptive states (e.g., beliefs and goals) in a human-robot learning task. It builds upon and integrates two important areas of research: (1) ambiguity resolution and perspective taking, and (2) learning from humans. Ambiguity has been a topic of interest in dialog systems (Grosz & Sidner 1990; Gorniak 2005). Others have looked at the use of visual perspective taking in collaborative settings (Trafton *et al.* 2005; Jones & Hinds 2002). We also draw inspiration from research into learning from humans, which typically focuses on either modeling a human via observation (Horvitz *et al.* 1998; Lashkari, Metral, & Maes 1994) or on learning in an interactive setting (Lieberman 2001; Atkeson & Schaal 1997; Nicolescu & Matarić 2003). The contribution of our work is in combining and extending these thrusts into a novel, integrated approach where perspective taking is used as an organizing principle for learning in human-robot interaction.

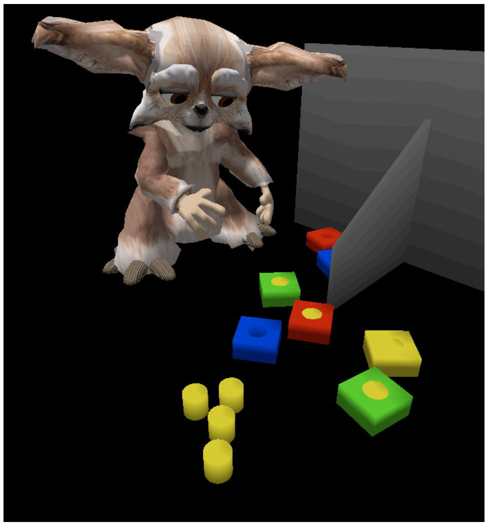

Figure 7: The robot was presented with similar learning tasks in a simulated environment.

Conclusion

This paper makes the following contributions. First, in a novel human subjects study, we show the important role that perspective taking plays in learning within a socially situated context. People use perspective taking to entertain a different set of hypotheses when demonstrations are presented by another person, verses when they are presented in a nonsocial context. Thus, perspective taking abilities shall be critical for robots that interact with and learn from people in social contexts.

Second, we present a novel architecture for collaborative human-robot interaction, informed by recent scientific findings for how people are able to take the perspective of others, where simulation-theoretic mechanisms serve as the organizational principle for the robot's perspective taking skills over multiple system levels (e.g., perceptual-belief, action-goal, task learning, etc.).

Finally, we evaluated our architecture on a benchmark suite drawn from the human subjects study and show that our humanoid robot can apply perspective taking to draw the same conclusions as humans under conditions of high ambiguity. Perspective taking, both in humans and in our architecture, focuses the agent's attention on the subset of the problem space that is important to the teacher. This constrained attention allows the agent to overcome ambiguity and incompleteness that can often be present in human demonstrations.

Acknowledgments

The work presented in this paper is a result of ongoing efforts of the graduate and undergraduate students of the MIT Media Lab Robotic Life Group. This work is funded by the *Digital Life* and *Things That Think* consortia of the MIT Media Lab. Jesse Gray would like to thank Samsung Electronics for their support.

References

Atkeson, C. G., and Schaal, S. 1997. Robot learning from demonstration. In *Proc. 14th International Conference on Machine Learning*, 12–20. Morgan Kaufmann.

Barsalou, L. W.; Niedenthal, P. M.; Barbey, A.; and Ruppert, J. 2003. Social embodiment. *The Psychology of Learning and Motivation* 43.

Blumberg, B.; Downie, M.; Ivanov, Y.; Berlin, M.; Johnson, M. P.; and Tomlinson, B. 2002. Integrated learning for interactive synthetic characters. *ACM Transactions on Graphics* 21(3: Proceedings of ACM SIGGRAPH 2002).

Carberry, S. 2001. Techniques for plan recognition. *User Modeling and User-Adapted Interaction* 11(1-2):31–48.

Davies, M., and Stone, T. 1995. Introduction. In Davies, M., and Stone, T., eds., *Folk Psychology: The Theory of Mind Debate*. Cambridge: Blackwell.

Gorniak, P. 2005. *The Affordance-Based Concept*. Phd thesis, MIT.

Gray, J.; Breazeal, C.; Berlin, M.; Brooks, A.; and Lieberman, J. 2005. Action parsing and goal inference using self as simulator. In *14th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*. Nashville, Tennessee: IEEE.

Grosz, B. J., and Sidner, C. L. 1990. Plans for discourse. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions in communication*. Cambridge, MA: MIT Press. chapter 20, 417–444.

Horvitz, E.; Breese, J.; Heckerman, D.; Hovel, D.; and Rommelse, K. 1998. The lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 256–265.

Jones, H., and Hinds, P. 2002. Extreme work teams: using swat teams as a model for coordinating distributed robots. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, 372–381. ACM Press.

Lashkari, Y.; Metral, M.; and Maes, P. 1994. Collaborative Interface Agents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume 1. Seattle, WA: AAAI Press.

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

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

Nicolescu, M. N., and Matarić, M. J. 2003. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

Sebanz, N.; Bekkering, H.; and Knoblich, G. 2006. Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences* 10(2):70–76.

Trafton, J. G.; Cassimatis, N. L.; Bugajska, M. D.; Brock, D. P.; Mintz, F. E.; and Schultz, A. C. 2005. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics* 35(4):460–470.