An Embodied Cognition Approach to Mindreading Skills for Socially Intelligent Robots

Cynthia Breazeal, Jesse Gray and Matt Berlin

The International Journal of Robotics Research 2009; 28; 656
DOI: 10.1177/0278364909102796

Cynthia Breazeal
Jesse Gray
Matt Berlin
The Media Laboratory,
MIT,
Wiesner Building, E15,
20 Ames Street,
Cambridge, MA 02139-4307,
USA
{cynthiab, jg, mattb}@media.mit.edu

An Embodied Cognition Approach to Mindreading Skills for Socially Intelligent Robots

Abstract

Future applications for personal robots motivate research into developing robots that are intelligent in their interactions with people. Toward this goal, in this paper we present an integrated socio-cognitive architecture to endow an anthropomorphic robot with the ability to infer mental states such as beliefs, intents, and desires from the observable behavior of its human partner. The design of our architecture is informed by recent findings from neuroscience and embodies cognition that reveals how living systems leverage their physical and cognitive embodiment through simulation-theoretic mechanisms to infer the mental states of others. We assess the robot's mindreading skills on a suite of benchmark tasks where the robot interacts with a human partner in a cooperative scenario and a learning scenario. In addition, we have conducted human subjects experiments using the same task scenarios to assess human performance on these tasks and to compare the robot's performance with that of people. In the process, our human subject studies also reveal some interesting insights into human behavior.

KEY WORDS—Human-Robot interaction, social robot, cognitive architecture, social cognition, human-robot teamwork, learning from demonstration, perspective taking, mental models.

1. Introduction

The promise of personal robots motivates new applications for robotic technologies that interact with people to help realize human goals at home, at work, in hospitals, in schools, and beyond (see Tapus et al. (2007) for a review). In particular, whereas much of the research into robotics has emphasized developing systems that are intelligent in their interactions with inanimate objects and physical environments, social robotics focuses on developing systems that are intelligent in their interactions with people in human environments.

In psychology, *Theory of mind* (ToM)—also called *Mindreading*—is the ability to attribute mental states (e.g. beliefs, intents, desires, feelings, knowledge, etc.) to oneself and to others, and to understand that these mental states can be the cause of and thus can be used to explain and predict the behavior of others (Premack and Woodruff 1978). In practice, this competence enables people to coordinate minds and bodies to achieve skillful social behavior across numerous domains and contexts, including collaborative (or adversarial) teamwork, conversation, learning from others, and more.

Similarly, personal robots need an analogous competence to be able to skillfully think about, relate to, and coordinate their behavior with humans over a wide range of real-time, real-world social scenarios. Toward this long-term goal, this paper presents an integrated socio-cognitive architecture to endow social robots with mindreading skills.

1.1. A Simulation-theoretic Approach to Mindreading

The design of our socio-cognitive architecture is inspired by recent findings from neuroscience (e.g. Gallese and Goldman (1998)), embodied cognition (e.g. Barsalou et al. (2003)), and developmental psychology (e.g. Meltzoff and Decety (2003)) that reveal how living systems leverage their physical and cognitive embodiment through simulation-theoretic mechanisms to infer the mental states of others. Specifically, *simulation theory* holds that certain parts of the brain have a dual use: they are used to not only generate our own behavior and mental states, but also to predict and infer the same in others. To

The International Journal of Robotics Research
Vol. 28, No. 5, May 2009, pp. 656–680
DOI: 10.1177/0278364909102796

Figures 1–4, 8–16 appear in color online: http://ijr.sagepub.com

Fig. 1. Leonardo, operating a remote control panel with a slider, a button and a switch used in our cooperative behavior experiment. The robot can track objects and people tagged with reflective markers. The simulated Leonardo can perform more dexterous tasks than the physical robot, such as inserting pegs into blocks used in our social learning experiments. People demonstrate the task to be learned using a computer mouse to move objects in the virtual world.

understand another person's mental process, we use our own similar cognitive processes and body structure to simulate the mental states of the other person Gordon 1986; Davies and Stone 1995, in effect, taking the mental perspective of another.

There is growing scientific evidence that early ToM abilities and critical precursors develop from more embodied processes (e.g. mirror neurons (Rizzolatti et al. 1996)) and other embodied cognition mechanisms such as perspective-taking and simulation (e.g. Barsalou et al. (2003)) rather than only by symbolic processes (e.g. language). Our benchmark tasks allow us to investigate computational models of these embodied processes where we explicitly consider tasks that do not require language and only depend on human non-verbal behavior.

We have developed a benchmark suite of tasks that are inspired by those used in psychology to probe children's developing ToM competence. In particular, one of the most important milestones in ToM development is gaining the ability to attribute *false belief*: to recognize that others can have beliefs about the world that are wrong or different from one's own. The canonical test for this developmental milestone is the false-belief task (originally formulated by Wimmer and Perner (1983)). Inspired by these methods, several of the robot's assessment tasks are adapted from false-belief tasks to probe the robot's ability to ascribe knowledge to an agent based on perceptual experience, attribute false beliefs, take visual perspective, and to infer intents and desires to anticipate an agent's actions. These abilities are exercised in the context of two different domains: assisting a human to attain what they desire and learning from ambiguous human demonstrations. Finally, we have run a parallel set of human subject studies on the same benchmark suite. This allows us to assess human performance on these tasks, and to compare the robot's performance directly with human data. Through this cross-domain and "cross-species" analysis, our objective is to advance the state-of-the-art in endowing social robots with a flexible repertoire of mindreading skills that can be skillfully demonstrated in diverse tasks involving human partners, as well as to learn about human performance.

We assess the performance of our integrated system on *Leonardo*, a 65-degree-of-freedom anthropomorphic robot (and its simulated counterpart) that interacts in real-time with a human partner (see Figure 1). The same socio-cognitive architecture generates the behavior of the physical and simulated robots.

Similarly, our socio-cognitive architecture has been designed to interpret human behavior and the underlying mental states in real-time by simulating them within the robot's own generative mechanisms on the perceptual, motor, belief, and intentional levels. This grounds and constrains the robot's information about the human in terms of the robot's own physical embodiment and socio-cognitive architecture, both from the bottom-up through low-level perceptual and motor processes, as well as from the top-down from its intention and deliberation processes. In this way, the robot leverages its own physicality and architectural organization as important resources to make and ground its mental inferences. This enables the robot to make inferences during real-time scenarios about people's likely focus of attention and beliefs to better understand the intention behind their observable behavior. Importantly, these mindreading skills can be applied across different domains and tasks.

2. Related Work

Research at the intersection of human–robot interaction and social robotics strives to endow robots with a variety of human-compatible social skills and socio-cognitive competencies. Several of these computationally modeled skills and abilities have been identified as important precursors to the development of ToM in humans. For instance, some of the earliest

work exploring ToM ideas in robots concerned distinguishing animate from inanimate movement (Scassellati 2001). The ability to share (and learn how to share) attention has been identified as a critical precursor to ToM (Baron-Cohen 1991) and has been modeled on several robots (e.g. Scassellati (2002), Nagai et al. (2002), Fasel et al. (2002), Movellan and Watson (2002), and Deak et al. (2001)). Imitation has also been identified as a precursor Meltzoff (2005) and has been widely explored in robots as a method for learning motor skills and recognizing human actions (e.g. Schaal (1997), Billard et al. (2004), Breazeal et al. (2005), Gray et al. (2005), Demiris and Hayes (2002), Johnson and Demiris (2005), and Jenkins and Matarić (2002)).

Finally, perspective taking abilities have been demonstrated on several notable robotic systems that can take the visual perspective of another agent (often a human) to perform tasks such as playing "hide and seek" (Trafton et al. 2006), disambiguate among multiple possible referents within a cluttered physical space (Trafton et al. 2005), or provide instrumental or informational support during a human–robot teamwork task once beliefs diverge owing to visual occlusions in a workspace (Gray et al. 2005; Breazeal et al. 2006). An architectural challenge for social robotics is to integrate these diverse skills and abilities (along with other social skills) in a principled manner that can be applied across diverse social domains.

To date, related computational work in artificial intelligence has emphasized top-down symbol-based models, such as belief,desire, intentions (BDI) systems (Rao and Murray 1994), adaptive control of thought–rational (ACT-R) models (Emond and Ferres 2001), state, operator and result (SOAR) models (Laird 2001), collaborative discourse systems (Cohen et al. 1990; Pollack 1990; Grosz et al. 1999), and plan recognition (for a review, see Carberry (2001)). In robotics, plan recognition has been approached from the bottom-up by applying probabilistic frameworks on perceptual streams to learn and recognize plans (e.g. Ronnie et al. (2005), Needham et al. (2005), and Intille and Bobick (1999)).

Pollack (1990) identifies several critical shortcomings of many of these plan recognition techniques. First, with respect to top-down techniques, it is problematic to make the common assumption that the library of recipes (i.e. task knowledge) is mutually known to the actor and observing agent. Further, with respect to both top-down and bottom-up techniques, it is too limiting for the recognizing agent to only consider the actor's plan as a recipe for action while ignoring the actor's mental attitudes that resulted in having that plan. As a result, such systems are incapable of inferring and reasoning about misconceptions (i.e. false beliefs) or invalid plans of the actor (Pollack 1990) that frequently arise in complex, dynamic scenarios where each participant only has partial knowledge of the overall situation. (Pollack 1990) addresses this by arguing for an equally important conceptualization of plans as "complex mental attitudes" comprising a principled organization of mental states such as beliefs and intentions that underly the actor's recipe for action. Pollack and successors have applied these insights to develop sophisticated collaborative dialog systems (see Carberry (2001)). A challenge for robotics is to adapt such insights to non-verbal collaborative behavior.

3. Cognitive Architecture Overview

In light of these prior works, we argue that embodied processes for mindreading and their computational counterparts are important to investigate, understand, and assess in their own right. Further, embodied processes have particular relevance to mindreading abilities in robots given the physical coupling of robots to the real world.

Our architecture incorporates simulation-theoretic mechanisms as a foundational and organizational principle to support mindreading skills and abilities. See Figure 2 for a system overview diagram. The two concentric bands denote two different modes of operation. In *generation mode* (the light band) the robot constructs its own mental states to behave intelligently in the world. In *simulation mode* (the dark band) the robot constructs and represents the mental states of its human collaborator based on observing their behavior and taking their mental perspective. By doing so, the mental states of the human and robot are represented in the same terms so that they can be readily compared and related to one another.

For instance, within the perception system, the robot performs a transformation to estimate what the human partner can see from their vantage point. Within the motor system, mirror-neuron inspired mechanisms are used to map and represent perceived body positions of the human into the robot's own joint space to conduct action recognition. Within the belief system, belief-construction is used in conjunction with adopting the visual perspective of the human partner in order to estimate the beliefs the human is likely to hold given what they can visually observe. Finally, within the intention system where goal-directed behaviors are generated, schemas relate preconditions and actions with desired outcomes and are organized to represent hierarchical tasks. Within this system, motor information is used along with perceptual and other contextual clues (i.e. task knowledge) to infer the human's goals and how they might be trying to achieve them (i.e. plan recognition).

In summary, bottom-up processes actively construct likely action, perception, and belief states through an embodied process of simulation. In parallel, high-level task knowledge combined with simulation can be used to deduce likely desires, goals, plans, and beliefs from the top down. These sources of information are integrated to represent the human's mental states.

Our technical discussion proceeds as follows. In Sections 4 and 5 we present our intention system and social learning mechanisms that are the core technical contributions of this paper. These highly integrative systems are presented in significant detail intended to support reimplementation. Before

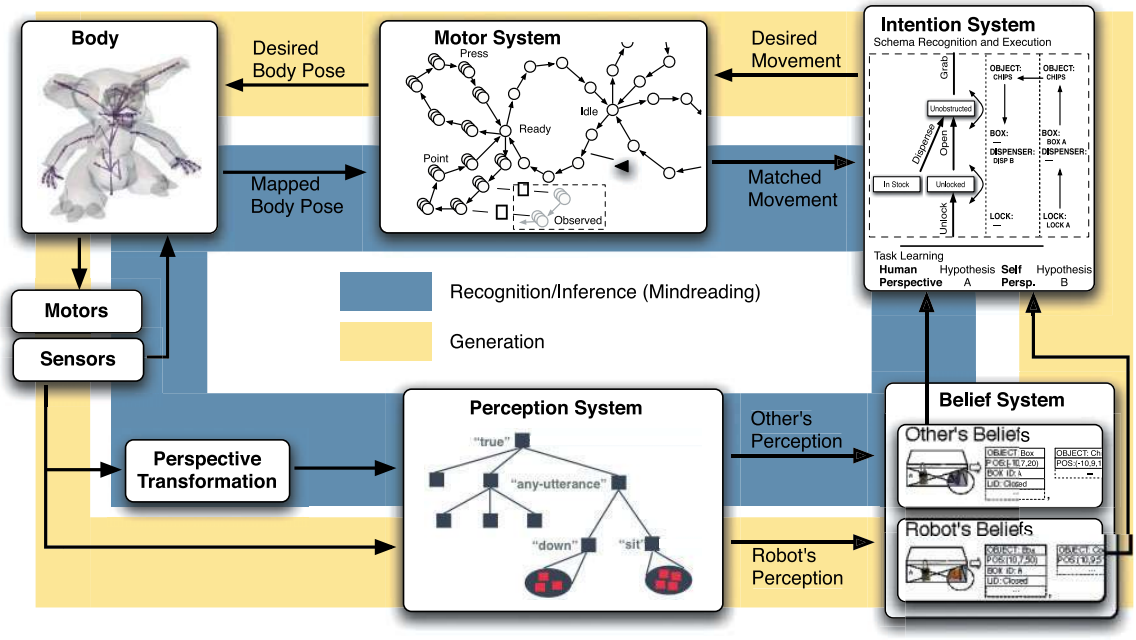

Fig. 2. System architecture overview. See the text for further details.

diving into these systems, we introduce some of the technical details of the perception system, belief system, and motor system components of our architecture, focusing on the technical issues necessary to understand the operation of the intention system and social learning mechanisms. While these subsystems are presented with lighter detail, references are provided to our prior work and to the related work of others for the interested reader.

3.1. The Perception and Belief Systems

In order to convey how the robot interprets the environment from the human'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.

3.1.1. Perception Modeling

On every time step, the robot receives a set of sensory observations $O = \{o_1, o_2, \ldots, o_N\}$ from its various sensory processes. As an example, imagine that the robot receives information about buttons and their locations from an eye-mounted camera, and information about the button indicator lights from an overhead camera. On a particular time step, the robot might receive the observations $O = \{$(red button at position (10, 0, 0)), (green button at (0, 0, 0)), (blue button at (−10, 0, 0)), (light at (10, 0, 0)), (light at (−10, 0, 0))$\}$. Information is extracted from these observations by the perception system. The perception system consists of a set of *percepts* $P = \{p_1, p_2, \ldots, p_K\}$, 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*

$$\begin{aligned} s_i &= \{(p, m, c, d) \mid p \in P, p(o_i) \\ &= (m, c, d), m * c > k\}, \end{aligned} \tag{2}$$

where $k \in [0, 1]$ is a threshold value, typically 0.5. Returning to our example, the robot might have four percepts relevant to the buttons and their states: a location percept which extracts the position information contained in the observations, a color percept, a button shape recognition percept, and a button light recognition percept. The perception system would produce five percept snapshots corresponding to the five sensory observations, containing entries for relevant matching percepts.

3.1.2. Belief Modeling

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: $b = \{(p_x, h_x), (p_y, h_y), \ldots\}$. 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;
- $d = d_1, d_2, \ldots, d_L$ is a vector of belief distance functions, $d_i : B \times B \rightarrow \mathcal{R}$;
- $q = q_1, q_2, \ldots, q_L$ is a vector of indicator functions where each element q_i denotes the applicability of d_i, $q_i : B \times B \rightarrow \{0, 1\}$;
- $w = w_1, w_2, \ldots, w_L$ is a vector of weights, $w_i \in \mathcal{R}$; and
- $c = c_1, c_2, \ldots, c_J$ is a vector of culling functions, $c_j : B \rightarrow \{0, 1\}$.

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:

$$b_i = N(s_i) = \{(p, h) \mid (p, m, c, d) \in s_i, \\ g = G(p), h = g(m, c, d)\}. \tag{6}$$

For the second process, the belief system merges new beliefs into existing beliefs 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:

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

We label this process $merge(B)$. Finally, the belief system culls stale beliefs by removing all beliefs from the current set for which $C(b) = 1$. In summary, then, a complete belief system update cycle proceeds as follows:

1: begin with current belief set B
2: receive percept snapshot set S from the perception system
3: create incoming belief set $B_I = \{N(s_i) \mid s_i \in S\}$
4: merge: $B \leftarrow merge(B \cup B_I)$
5: cull: $B \leftarrow B \setminus \{b \mid b \in B, C(b) = 1\}$

Returning again to the example, the belief system might specify a number of relevant distance metrics, including a measure of Euclidean spatial distance along with a number of metrics based on symbolic feature similarity. For example, a symbolic metric might judge observations that are hand-shaped as distant from observations that are button-shaped, thus separating these observations into distinct beliefs even if they are collocated. For the example, the merge process would produce three beliefs from the original five observations: a red button in the ON state, a green button in the OFF state, and a blue button in the ON state.

The belief system framework supports the implementation of a wide range of object tracking methods, including advanced tracking techniques such as Kalman filters (Kalman 1960) and particle filters (Carpenter et al. 1999; Arulampalam et al. 2002). The ability to specify multiple distance metrics allows sophisticated, general-purpose tracking methods such as these to operate side-by-side with hand-crafted rules which encode prior domain knowledge about object categories, dynamics, and persistence.

3.1.3. Belief Inference and Visual Perspective Simulation

When collaborating on a shared task, it is important for all parties involved to have a consistent representation of the task context. However, in complex and dynamic environments, it is

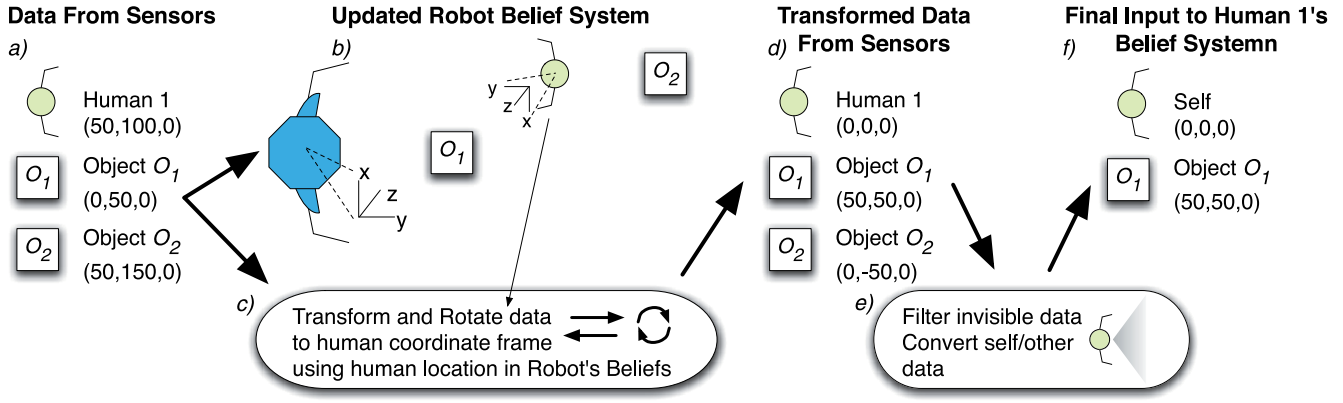

Fig. 3. Perspective transform of sensor data for belief modeling. (a) Data from sensors is used to update the robot's own model of the world (shown in (b)) via the normal belief system update. (b) The real world scenario and corresponding model: the robot (shown as a dark gray hexagon) can see the human (shown as a light gray circle) and two objects. The human can only see object O_1. Coordinate system orientation is shown next to the human and the robot where the origin is centered on each agent. (c) The human's position and orientation from this model is used to transform incoming sensor data to data that is relative to the human's coordinate system. (d) The result of the transformed data. (e) Next, objects that are out of sight of the human (estimated by an "attentional cone") are filtered out, and the data is transformed to a human centric format. (f) This data is now ready to be presented to the belief system that models the human's beliefs.

possible for one collaborator's beliefs about the context surrounding the activity to diverge from those of other collaborators. For example, a visual occlusion could partially block one person's viewpoint of a shared workspace but not that of the other, or an event could occur that one person witnesses, but the other does not. There are many situations where the knowledge that two or more people have of a shared scenario can differ over time. The ability for an agent to estimate what others do and do not know based on their perceptual experience is at the crux of many false belief tasks. In this section we describe our method of modeling the knowledge of nearby humans based on their visual experience by taking their visual perspective.

As described in the previous section, belief maintenance consists of incorporating new sensor data into existing knowledge of the world. The robot's sensors are all in its own 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 reuses the same mechanisms used for its own belief modeling, but first transforms and filters the incoming data stream (see Figure 3). In this way, the beliefs modeled for the human are handled with the same tracking and maintenance systems that the robot uses for its own world model; however, the data is manipulated to simulate first-person experience from the perspective of the human being modeled.

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. If the inputs to the robot's perceptual-belief pipeline are the sensory observations $O = \{o_1, o_2, \ldots, o_N\}$, then the inputs to the secondary pipeline that models the human's beliefs are O', where

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

where

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

and

$$P \;:\; \{\text{robot local observations}\} \rightarrow \{\text{person local observations}\}. \tag{9}$$

Visibility is determined by a cone calculated from the human's position and orientation. The robot also filters out objects whose view is blocked by occlusions (for any occlusions that it can detect).

Maintaining this parallel set of beliefs is different from simply adding "is-visible-to-human" 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, Figure 4 shows an example where this approach keeps track of the human's false beliefs about objects that have changed state while out of the human's view. This method has the advantage of keeping the model of the human's beliefs in the same format as the robot's own, allowing both for direct comparison between the two and operating on these beliefs with the same mechanisms that operate on the

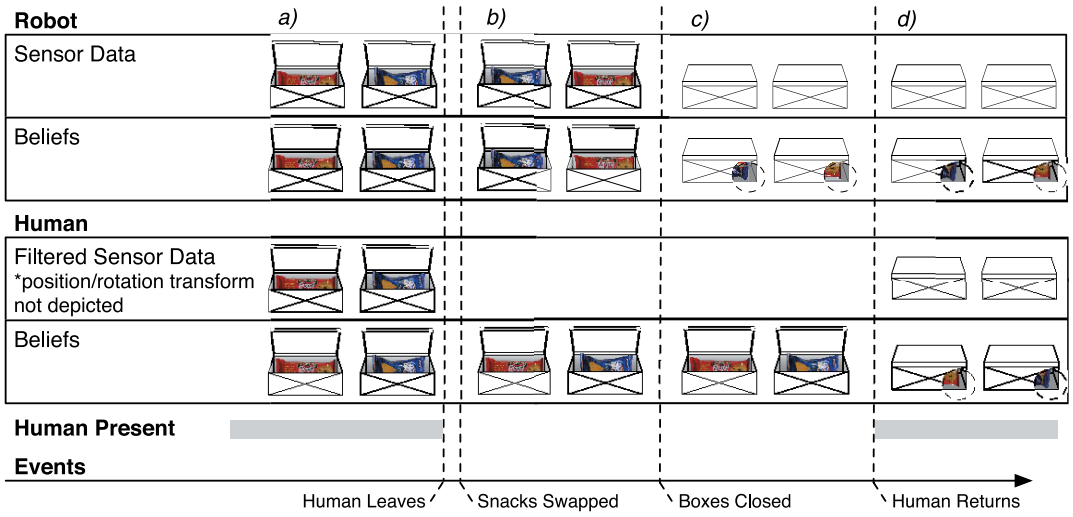

Fig. 4. Timeline showing belief modeling. (a) Initially the robot's model of the human's beliefs agrees with the robot's model of beliefs. (b), (c) When snacks are swapped and boxes are closed, the human is gone and human's model is not updated. (d) The human returns and the model is updated to indicate that the human knows that boxes are closed, however human's model continues to indicate initial (now false) snack positions.

robot's own. This is important for establishing and maintaining mutual beliefs in time-varying situations where beliefs of individuals can diverge over time.

3.2. The Motor System

An important element of the robot's ability to predict and help with goals of people is to be able to make sense of their physical actions. The approach we take is to reuse the physical actions the robot can perform to recognize the actions observed in the human. We do this in a two-stage process. First we transform observed human movements into the same movement space as the robot. Once the observations are in a similar representation to the robot's own motor generation capabilities, we can match the robot-space motions against its own motion repertoire. This dual use of the same motor processes for both production and recognition is inspired by mirror neurons. This gives us a starting point towards understanding the overall activity being performed by the human which also depends on the surrounding context (we discuss this in Section 4).

3.2.1. Body Mapping

In order to compare observed human motions to the robot's motion repertoire, it is important for the human motions to be in the same representation as the robot's own motions. This can be difficult, because human morphology may not be the same as the robot's. Also, whatever sensing technology is used to provide data of human movements is unlikely to provide data in a way that can be related directly to the robot's representation of its own motions.

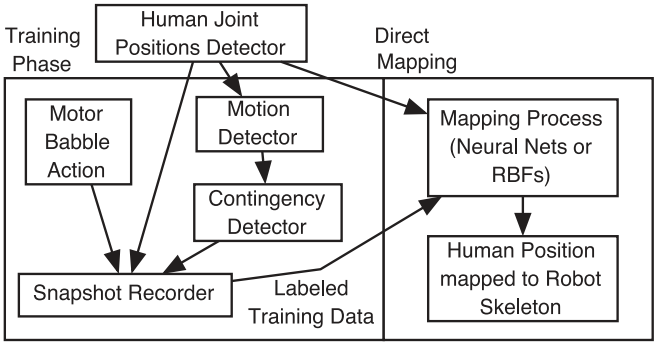

Fig. 5. Mapping perceived human joints onto the robot's skeleton to allow for a comparison between joint configurations of the robot and the human.

We use a mapping technique where the relation between sensed human body positions and the robot's own body positions is learned through an imitative interaction (Breazeal et al. 2005; Gray et al. 2005). This technique allows the joint angle configuration of the human to be mapped efficiently to the geometry of the robot as long as the human has a consistent sense of how to mimic the poses of the robot and is willing to go through the quick, imitation-inspired process to help the robot learn this mapping. Figure 5 presents a schematic of this process.

We have used this technique to learn facial imitation based on facial features tracked using the AxiomFFT system. In this

case, the human imitates facial expressions of the robot until the robot has enough samples to train a neural network that maps between perceived two-dimensional locations of human facial features in image coordinates to the robot's facial joint space (Breazeal et al. 2005). We have also used it to learn a mapping from the arms and torso of an observed human to the corresponding body regions of the robot using a motion capture suit (Gray et al. 2005), and later using an optical motion tracking system (Brooks et al. 2005) as the input observations of the human pose.

3.2.2. Matching Observed Actions to the Motor Repertoire

Once the perceived data is in the joint space of the robot, the motor system represents these observed movement trajectories in the same way that it represents its own movements. The robot's motor repertoire is represented as a directed graph of connected poses, called the *pose graph*. The nodes represent specific body poses, and the arcs represent allowed transitions between them. Families of poses can be represented as a sub-graph of actions (e.g. different kinds of reaching, pointing, waving, etc.) and links between sub-graphs represent allowable transitions between families of actions. In addition, weighted blends of either discrete poses or full trajectories can be generated to enlarge the repertoire of possible movements (Downie 2000). For instance, the robot may have six explicit reaching movements represented in its pose graph (primitives), but can generate a new reaching movement using a weighted blend of reaching primitives to span its entire workspace. The goal of this example-based technique is to satisfy the dual goals of having the robot produce lifelike, expressive motion characteristic of human-made animations while still having the flexibility to behave autonomously. In some cases if we need exact positioning (such as flipping a switch) we start with the blended solution and augment it slightly using inverse kinematics to achieve the end-effector position while attempting to preserve what we can of the animated motion.

This structure is quite a useful way to represent the motions of the robot. In practice, we overlay multiple motor systems for different body regions that can run simultaneously. This allows the robot to perform multiple motions simultaneously, such as pointing at an object, directing its gaze towards it, nodding, and expressing an emotional state such as interest. Individual trajectories specify the joints they require which allows the system to determine which motions are compatible with others (can be run simultaneously).

Once the observed movements are represented within the pose graph, strung together into a trajectory through this space, the next challenge is to determine whether this trajectory is similar to (or can be generated by) any that exist within the robot's motor repertoire. Many interesting techniques exist and others are being developed to determine the match between trajectories based on the relative importance of spacial errors, timing errors, etc. (e.g. Jenkins and Matarić (2002) and Demiris and Hayes (2002)). In the interaction described here we were able to use a simple heuristic to provide a goodness of fit measure: a voting system that chooses trajectories based on a running best-overall-matching-pose measure. However, in interactions with more motions that need to be classified, we have also explored the use of morphable models to provide a more general solution (Brooks et al. 2005).

Representing observed human's movements as one of the robot's own movements is useful for further inference using the intention system. Rather than trying to recognize human behavior purely from a collection of joint angle trajectories, the intention system integrates this motor information with other context provided by tasks schemas (that link environmental conditions with actions to achieve expected outcomes). This is described in the following section.

4. The Intention System

The intention system is responsible for generating the goal-achieving behavior of the robot. Our representation for goal-directed action enables the robot to plan a set of actions under particular circumstances to achieve a desired result. Furthermore, the robot can also introspect over these representations to determine the person's desires, plans, and goals based on what the robot's would be if it were performing the same action in the human's situation. A core feature of this self-as-simulator architecture is that the robot can employ multiple world and agent models to infer introspective states, which it can then apply across multiple task domains. For instance, the robot can use its model of a person's beliefs to help interpret and predict their behavior, and then use its own model of the world to decide how it can best help that person with their goals.

The following sections describe our core task representation and the processes that operate on this representation to generate behavior and to produce intentional inferences. Finally, in Section 4.4, we work through a detailed example of how these processes function, providing additional technical details of our fielded system.

4.1. Task Representation using Schemas

Within the deliberative system of the robot, the atomic-level representation of a goal-directed behavior is a schema that associates its necessary perceptual preconditions with a specific action (optionally performed on a particular object, or with other parameters) to achieve an expected outcome: its goal.

As such, it resembles STRIPS operators within classic planning literature. Schemas can be organized sequentially and/or hierarchically to create larger structures to represent tasks and execute them. When chaining sequential schemas, the

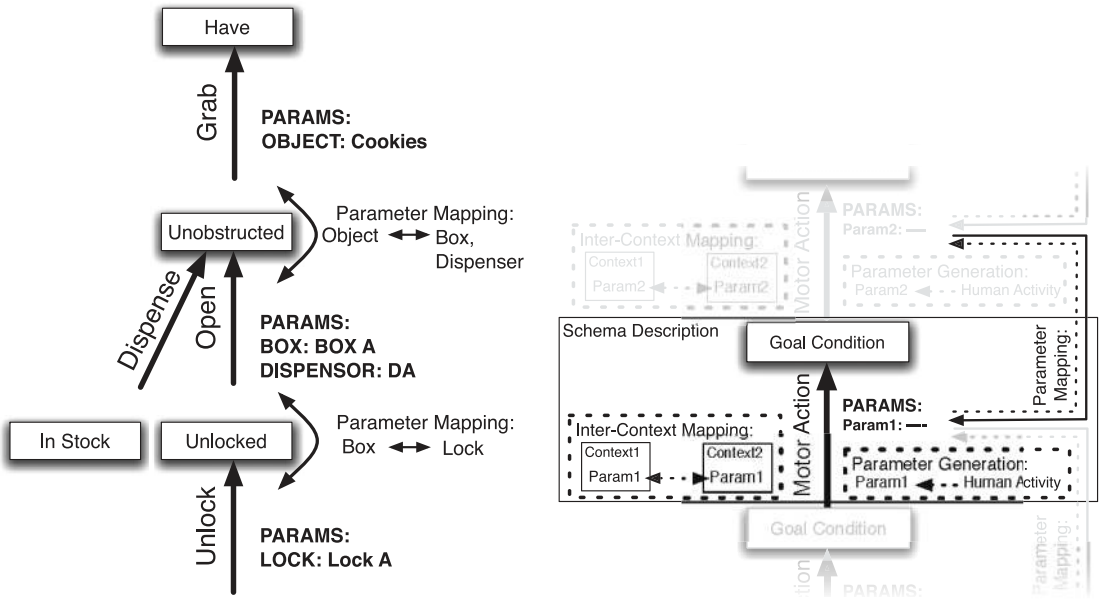

Fig. 6. Example of a task representation. In the simplified task representation to the left, the agent intends to obtain cookies. There are two possible behaviors to obtain cookies: open a locked box that contains cookies, or operate a dispenser to release cookies. In the more detailed figure to the right, the highlighted schema acts as the precondition for the upper schema, while the lowest schema is the precondition for the highlighted schema. In order for the schema to activate a necessary precondition (or to evaluate whether it is necessary using that precondition schema's goal condition) it may need to compute the necessary parameters relevant for that schema based on its own parameters. The downward mapping (solid line) in generation mode is necessary to perform and evaluate precondition schemas based on the parameters of an upper schema. The upward mapping for simulation mode (dashed line) is used to populate later schemas with potential parameters based on known precondition parameters (used during an attempt to predict an ultimate goal for an observed action). Finally the inter-context mapping module is necessary when the robot is trying to compare observed goals with its own world knowledge in order to formulate a helpful plan. It must have a metric to determine how parameters that a human is using (that often relate to their possibly differing beliefs about the world) can be expressed in terms of the robot's own world knowledge.

goal of one schema becomes the precondition of the subsequent schema. Compound tasks are specified as a hierarchy of schemas, where the expected result of multiple schemas are the inputs (i.e. listed in the preconditions) of the subsequent schema. To achieve some desired task goal, only the relevant schema need be activated and all necessary preconditions will be fulfilled. Figure 6 shows an example schema structure.

Each schema has a number of individual components. It has a motor action, which causes the robot to physically perform some sort of movement trajectory by activating the corresponding path within the pose graph (described in Section 3.2.2). It also has an evaluative mechanism to determine the success of the action.

Both the motor action and the evaluative mechanism may depend on additional parameters (for instance, specifying one or more target objects in the world). These parameters may be provided externally from a higher-level deliberative processes if the schema is activated directly to produce the robot's own behavior (see Section 4.2). If the schema is activated in simulation mode in response to an observed action, then the parameters must be discovered based on observation (see Section 4.3). If the schema is activated as a precondition for a sequence of schemas, however, the associated parameters must be determined automatically based on the parameters of the downstream schema.

4.2. Generating Goal-achieving Behavior

In behavior generation mode, schema structures can be traversed top down to achieve goals and automatically satisfy preconditions in the process. In this process, the robot activates the top-level schema which in turn may need to activate other supporting schemas before it can execute fully. When schemas are activated to generate the robot's own behavior, most schemas are parameterized by a set of arguments that

adapts how the motor action operates to suit the situation at hand. For instance, the style of the action may be adjusted to express the robot's affective state, or a particular object could be set as the target for a given action. When this first schema is activated, the parameters (or target) for this schema is based on its goal. It is then up to the schema hierarchy to automatically generate schema parameters for any other required schemas based on the initial parameters provided to this first top-level schema.

At every juncture between schemas in a hierarchy, there exists a parameter mapping module. This module is designed to generate the necessary parameters for precondition schemas based on the existing parameters for the parent schema. For example, in Figure 6 the robot's goal is to obtain cookies. Grasping cookies requires an unobstructed path. In this case, however, cookies are believed to be in a box or in a dispenser that the robot can perceive. The unobstructed schema is activated to reveal a clear path to cookies. One strategy is to open the target box, box A. As it turns out, box A is locked. Accordingly, the robot should attempt the unlocked schema to unlock the correct lock, namely lock A.

Note that for any given situation, many schemas within the robot's task repertoire will not be relevant. In these cases, the parameter mapping module will not be able to assign parameters to instantiate the corresponding schema. This indicates that the current context is not appropriate for that precondition schema to be performed. Hence, this serves an important filtering process whereby the robot only entertains executing schemas that are relevant and performable.

The system currently uses this filtering process to select one goal schema that describes the human's behavior. However, for future tasks in more complex environments, it will be important to revise this system to maintain multiple, probabilistic hypotheses. Another important future addition to the system is in the area of probabilistic actions. The robot's goals are specified relative to perceived world state which gives it some persistence in the face of failed actions, however a more explicit modeling would be required to use that information for planning and replanning in the face of motor failure and other uncertain outcomes.

4.3. Inferring Intent from Observed Behavior

In simulation mode, the robot tries to infer the intention of a person's observed course of action. To do so, the robot traverses schemas in the reverse direction. As schemas are traversed bottom-up, each schema's parameter mapping module is applied to the robot's model of the human's beliefs: mapping parameters relevant to a precondition upwards to parameters necessary for the next higher-level schema. In general, the reverse mapping may be ambiguous (for instance, if someone is opening a box containing multiple items, which one they might want to grab), and it may also be arbitrarily complex. For this reason, the architecture allows for each action to specify its own mapping function which handles both forward and reverse mapping. The actions used in this demonstration employ a mapping function based on object types and spatial relationships, which can operate similarly in either forward or reverse operation.

For instance, if schema S_1 (operating in relation to belief b_1) is a precondition to schema S_2 (operating in relation to b_2), then if either b_1 or b_2 is known the other can be determined according to the following:

$$b_1.isTypeForS_1 \wedge b_2.isTypeForS_2 \wedge r(b_1, b_2), b_1, b_2 \in \text{Beliefs},$$

where r is the relation that must hold between the beliefs.

For the schemas described here, r is a position-based relationship. For example, in the case of a lock and a box:

$$r(b_1, b_2) = |b_1.location - b_2.haspLocation| < 20 \text{ cm}.$$

In simulation mode, some schema parameters must be detected through direct observation, such as the target object of an observed action. In this case, a parameter generation module (associated with each schema) computes the specific arguments necessary to simulate an observed schema in the manner it is being performed by the human. For instance, a person's arm trajectory for a reaching movement has different end purposes depending on what is being reached for: to grasp cookies, to open a lid, to unlock a lock, etc. In this case, the parameter generation module for the reaching schema produces its values based on the robot's models of the beliefs of the person as estimated by the belief system, namely, an object near the person's hand that they can see:

$$target = b \text{ iff } \exists h, b \in B$$

$$\{b.isCorrectType \wedge h.isOwnHand$$

$$\wedge |b.position - h.position| < thresholdDistance\}$$

where B is the subject's Beliefs.

If there exists such a b, then the parameter generation module has determined the relevant target b, and the robot concludes that the attached schema may be relevant to the observed action.

Analogous to the filtering role of the parameter mapping module described previously, these parameter generation modules also serve an important filtering function that narrows the relevant candidate schemas that may describe the human's observed behavior. If the parameter generation module is unable to populate its schema with the appropriate arguments for the current situation, the robot concludes that this schema does not describe the human's current behavior.

To summarize, the intention system runs in simulation mode to enable the robot to observe the human and infer their goal. This is achieved by first determining which schema in the robot's own repertoire matches the human's activity by finding a schema whose motor action matches the observed action of the human and whose parameter generation module indicates that it is a relevant schema in the human's current context. From there, the robot can traverse upwards in the schema hierarchy to try to determine the ultimate goal of the observed behavior. At each step, the robot must attempt to predict the relevant parameters of the higher (temporally later) schema based on the parameters of the lower (preceding) schema using the connecting parameter mapping module. Once it comes to a point where there are no more unique, valid, higher schemas (this can happen because the schema structure has no further schemas, because a parameter mapping module cannot map parameters any further, or because there is more than one valid schema or parameter for the next step), then it has found the farthest goal it can predict into the future without being ambiguous (see Algorithm 1).

Algorithm 1 Finding the human's goal.

abbreviations: PGM = parameter generation module, MA = motor action, PMM = parameter mapping module

ObservedSchema():
- **loop**
  - **for all** $s \in$ SCHEMAS **do**
    - **if** s's PGM succeeds and
      - s's motor action matches the human's **then**
        - **return** s

FindGoal():
- $OS \Leftarrow$ ObservedSchema() {Blocks until Schema is Observed}
- $Params \Leftarrow OS$'s PGM operating on $humansBeliefs$
- **loop** {climb schema tree through unique valid schemas}
  - $matchingSchemas \Leftarrow$ Empty List
  - **for all** $s \in$ SCHEMAS such that OS is precondition of s **do**
    - **if** OS's PMM can map $Params$ to s using $humansBeliefs$ **then**
      - add s to $matchingSchemas$
  - **if** $matchingSchemas$ contains exactly 1 element **then**
    - $OS \Leftarrow$ first element of $matchingSchemas$
    - $Params \Leftarrow OS$'s PMM maps $Params$ to OS using $humanBeliefs$
  - **else**
    - **break** {no higher unique schemas found}
- **return** OS, $Params$ {goal is schema populated with parameters}

4.4. An Example: Goal Assistance

Using all of the parts described above, the robot can infer what the human is intending to do even if their beliefs about the situation are false or incomplete and their resulting course of action will fail to accomplish their goal. How might a robot help a person in this situation? We consider the case where the robot has true beliefs about the situation at hand. The robot can assist the human by first adopting the same goal and then computing a course of action that resolves the errors the human has encountered.

To adopt the human's goal, the robot maps goal information from the context of the human's beliefs into its own set of beliefs. The most common mapping is simply to find a belief in the human's estimated context that preserves a set of properties from the first belief.

Here a belief B_{bs1} from one belief system maps to B_{bs2} in another based on the properties P if

$$\forall p \in P, \quad B_{bs1}\text{'s } p = T(B_{bs2}\text{'s } p)$$

$$T(p) = \begin{cases} \text{perspective transform of } p & p \text{ has location data,} \\ \text{p} & p \text{ has no location data.} \end{cases}$$

This goal can then be used to provide assistance. Consider the case where the robot can use the same schema hierarchy computed according to Section 4.3. Algorithm 2 summarizes this process. For readability the algorithms shown here are simplified to refer to only a single human, however the architecture supports multiple humans.

Algorithm 2 Providing goal-based assistance.

ProvideAssistance():
- $GS, GP \Leftarrow$ FindGoal()
- $GP \Leftarrow GS$ maps goal params from $humansBeliefs$ to $robotsBeliefs$
- **if** $GP =$ NULL **then**
  - **return** NULL {cannot help if robot cannot understand goal}
- $TGS, TGP \Leftarrow$ FindATerminalSchema(GS, GP, $robotsBeliefs$)
- **if** CanPerform(TGS, TGP) **then**
  - perform(TGS, TGP)
- **else if** $(TGS \neq GS)$ or $(TGS = GS \wedge TGP \neq GP)$ **then**
  - pointTo(TGP)

FindATerminalSchema(GS,GP,B):
- **for all** $s \in$ SCHEMAS such that s is precondition of GS **do**
  - $sGP \Leftarrow s$'s PMM mapping GP to s using B
  - **if** $sGP \neq$ NULL
    - and s is not accomplished according to sGP and B **then**
      - **return** FindATerminalSchema(s, sGP, B)
- **return** GS, GP

Figure 7 illustrates this process, following the example in Figure 6, where the human wants a bag of cookies they believe is contained in box A. However, the human's beliefs are false as the cookies were moved to box B while the human was not looking. The robot saw this switch take place and therefore has true beliefs of the situation. The robot uses this knowledge to help the human obtain the object of their desire. This is the general premise for our cooperative behavior experiments in Section 6.

5. Social Learning Mechanisms

The previous section highlighted a cooperative behavior scenario to illustrate the mechanisms within the intention system for behavior generation and simulation for goal inference. In this section we present the task and goal learning mechanisms

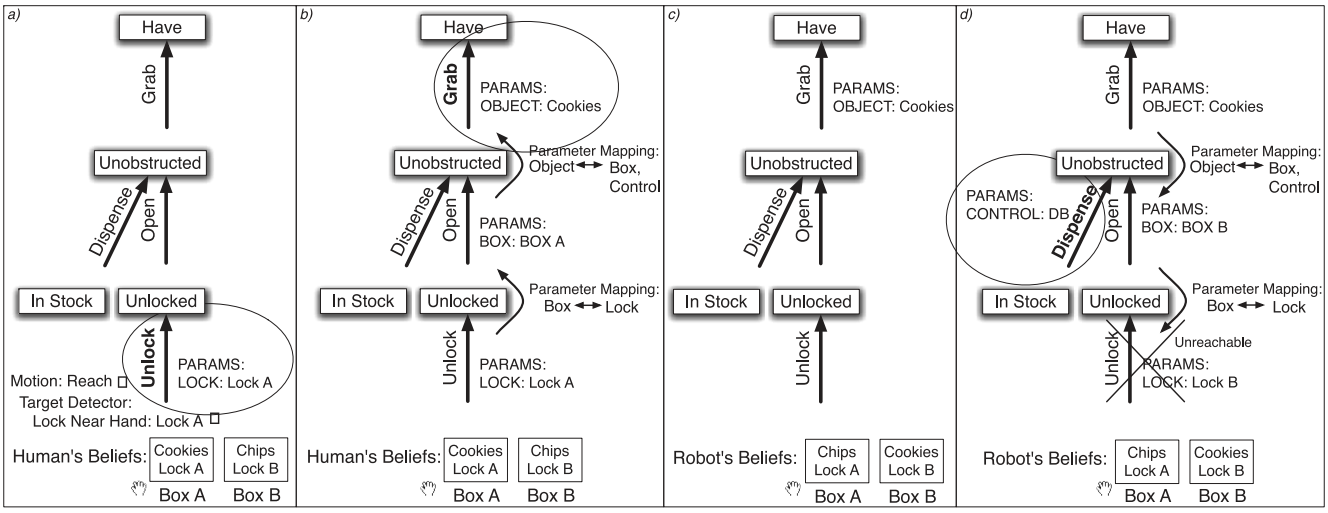

Fig. 7. Example goal inference and helpful behavior. In this example, the human is trying to gain access to a bag of cookies which they believe is locked in box A. The robot saw the cookies moved to box B without the human seeing this event, so the human has false beliefs of the true location of the cookies. The schema hierarchy shown here describes two possible solutions that the robot knows to produce a food item: either unlocking and opening the correct box, or dispensing a matching food item from a dispenser that it can operate. Flow diagrams (a)–(d) represent the corresponding schema hierarchy that is evaluated in the context of a particular set of beliefs (either the human's or the robot's) shown at the bottom. (a) The robot detects a "reach" motion and the relevant context for the "unlock" segment ("own-hand-near-lock" from the human's perspective). This corresponds to the human reaching for the lock on box A. (b) The process traverses up the hierarchy, using a model of the human's beliefs as input to the parameter mapping functions to predict targets for the potential human actions that are likely to follow the current action. In this example, the robot determines that the human's desire is to obtain the cookies. (c) Once a final goal is calculated, the process switches to the robot's own belief context. The robot knows that chips are actually in box A and cookies are in box B. (d) Again, the system uses parameter mapping to determine the targets of relevant actions necessary towards the goal, but this time starting from the goal and working backwards using knowledge from the robot's own beliefs. The robot can then choose an action that helps the human attain his goal: either unlocking box B (the robot realizes the human is looking in the wrong box), or dispensing a bag of chips from the robot's dispenser. For instance, a principle of "least effort" can be applied to decide between the two.

implemented in our cognitive architecture. We are particularly concerned with social learning scenarios where a human teaches a robot through face-to-face interaction. This social context gives rise to interesting issues that do not occur when learning in isolation. Specifically, mindreading skills play an important role in enabling the robot to learn what the human intends to teach.

For instance, when demonstrating a task to be learned, it is important that the context surrounding the demonstration is 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. Consider the situation where a visual occlusion blocks the teacher's viewpoint of a region of a shared workspace but not that of the learner. Consequently this leads to ambiguous demonstrations where the teacher does not realize that the visual information of the scene differs between them.

The ability for the learner to infer the beliefs of the teacher allows the learner to build task models that capture the intent behind the human's demonstrations. To support this, perspective taking processes are interwoven into the learning mechanism to support social learning scenarios. We evaluate this capability in Section 7 as part of our benchmark suite.

Note that our core interest is not on the particulars of the underlying learning mechanism, any number of techniques would suffice. Rather, our focus is how mindreading abilities interface with the underlying learning mechanism to support social learning scenarios.

5.1. 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, owing 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 hold 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:

$$S = \{[(E_1 \dots E_x), G, c]_P, [(E_1 \dots E_x), G, c]_{1\dots n}\},$$

$$E = a|S,$$

$$G = \{B_1 \dots B_y\},$$

$$B = p_{C_1} \dots p_{C_i} \cup p_{E_1} \dots p_{E_j}.$$

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 a 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:

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

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 $Snp(t_{\text{end}})$. 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.

Once the human indicates that the current task is done, S contains the representation of the seen example ($[(E_1 \dots E_x), G, 1]$). 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 Berlin et al. (2006)); 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. Here $P(D|h)$ 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, as a task is learned, the algorithm initially chooses highly specific hypotheses (those with many criteria and expectation features matching the initial demonstration), with more general hypotheses selected as subsequent demonstrations invalidate various specific features.

5.2. Perspective Taking and Task Learning

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. In this paper, we take the perspective of the teacher, and assume that their schema captures the rule they intend to teach. In prior work, we have also explored conflict resolution behaviors where the robot attempts to resolve ambiguities as they arise (Breazeal et al. 2006).

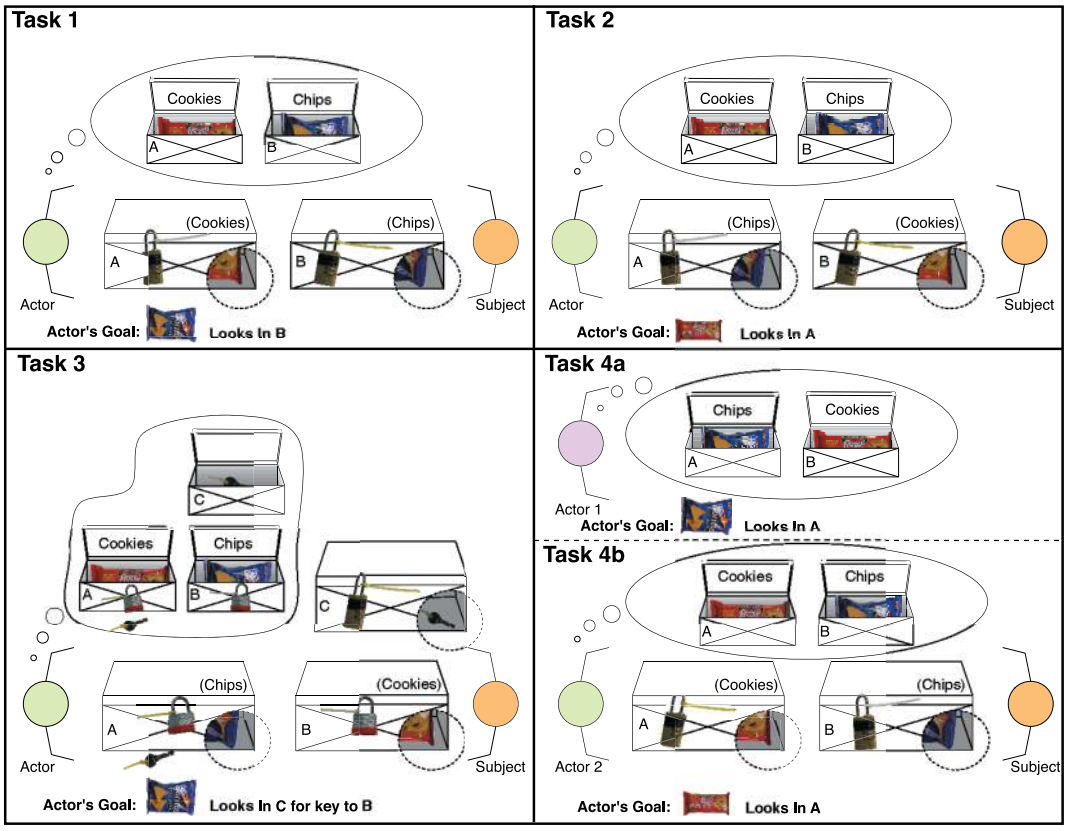

Fig. 8. The four collaborative benchmark tasks: (1) simple goal inference; (2) goal inference with false beliefs; (3) goal inference with false beliefs and indirect, dislocated action; and (4) goal inference with multiple agents and false beliefs. Shown are the actual world state and the actor's "belief" state at the moment when the subject's behavior is classified.

6. Providing Assistance on a Physical Task

In order to evaluate our cognitive architecture, we have developed a novel set of benchmark tasks that examines the use of belief reasoning and goal inference by robots and humans in a collaborative setting. Our benchmark tasks are variants of the classic Sally-Anne false belief task from developmental psychology, but embedded within a live, cooperative setting. Subjects interact face-to-face with a partner (an experimental confederate), and are prompted to assist their partner in any way they see fit. Language is not required to perform these tasks. Instead of probing the participant with an explicit prompt (e.g. "where will your partner look for the cookies?"), we observe their behavior as they attempt to assist their partner. Our objective is to examine the spontaneous use of goal inference and false belief reasoning in collaborative activity.

6.1. Benchmark Tasks

A schematic of four benchmark tasks is shown in Figure 8. In each task, the subject (i.e. a human or robot) interacts with a collaborative partner (actor) who is an experimental confederate. The subject has access to a collection of food objects (cookies in a small red package or chips in a larger blue package) that are identical to hidden target objects locked away in opaque boxes that their partner (actor) may be searching for. It is thus possible for the subject to assist their partner (actor) by giving them the food item that matches the target of their search without requiring the actor to figure out how to unlock the appropriate box. Or the subject can communicate relevant information, such as gesturing to the location of the target item.

In those tasks that call for boxes to be sealed, color-coded combination locks are used. Two of the lock's four numeric dials are covered up and fixed in place by electrical tape, leaving only two dials free for manipulation. This lock mechanism served an important timing function in our study, introducing a delay in the actor's process of opening any sealed box. This gives the subject sufficient time to consider the actor's goal and beliefs and then perform potential helpful actions before the actor unlocks the box.

(1) **Task 1** is a control task examining simple goal inference. The subject and actor both watch as the experi-

menter hides a package of cookies in box A and a bag of chips in box B. The experimenter then seals both boxes. The actor receives instructions written on a notecard to deliver a bag of chips to the experimenter. The actor proceeds to attempt to open box B, and the subject's subsequent behavior is recorded. In order to successfully assist the actor, the subject must infer that because the actor is attempting to open box B, the actor's goal is to acquire the chips contained within the box.

(2) **Task 2** examines goal inference with false beliefs. The setup proceeds as in Task 1, with subject and actor both observing cookies hidden in box A and chips hidden in box B. After the boxes are sealed, the actor is asked to leave the room, at which point the experimenter swaps the contents of the boxes. The actor returns, receives instructions, and attempts to open box A. In order to successfully assist the actor, the subject must infer that the actor's goal is to acquire the cookies, even though box A currently contains the chips.

(3) **Task 3** examines goal inference with false beliefs and indirect, dislocated action. The setup proceeds as in Task 2, however, in this case, the experimenter locks both boxes A and B with color-coded padlocks. The key to box A is left in plain view, but the key to box B is sealed inside of a third box, box C. The actor is then asked to leave the room, at which point the experimenter, using a master key, swaps the contents of boxes A and B, leaving both boxes locked. The actor returns, receives instructions, and attempts to open box C. In order to successfully assist the actor, the subject must infer that the actor's goal is to acquire the chips, even though the immediate target of the actor's actions, box C, contains neither the chips nor even the key to a box containing chips.

(4) **Task 4** examines goal inference with multiple agents and false beliefs. In this task, the subject is introduced to two collaborative partners, actors 1 and 2. All three watch as the experimenter hides cookies in box A and chips in box B, and then seals both boxes. Actor 1 is then asked to leave the room, at which point the experimenter swaps the contents of boxes A and B in view of both the subject and actor 2. Actor 2 is then asked to leave, and actor 1 returns. Actor 1 receives instructions and attempts to open box A. The subject's subsequent behavior is recorded (**Task 4a**). Finally, actor 1 leaves, and actor 2 returns, receives instructions, and also attempts to open box A. The subject's behavior is recorded (**Task 4b**). In order to successfully assist both actors, the subject must keep track of actor 1's false beliefs about the object locations as well as actor 2's correct beliefs about these locations.

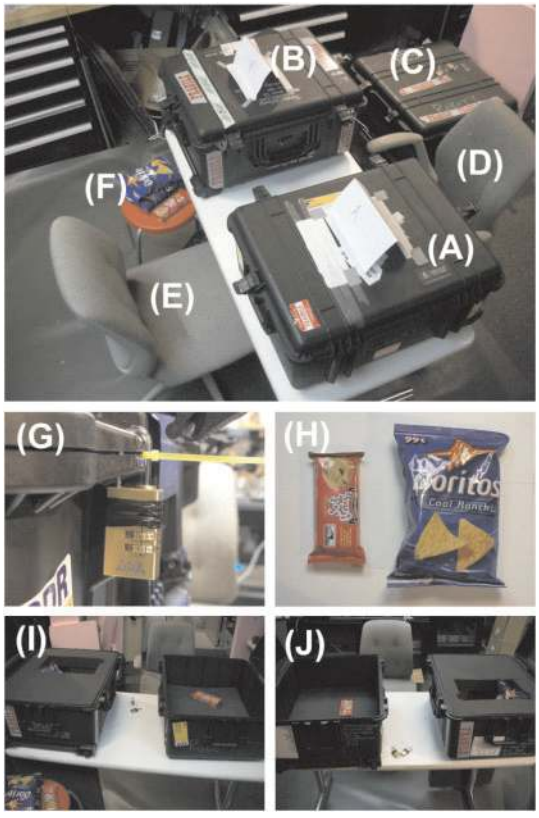

Fig. 9. Setup of the human subjects study. (A)–(C) Boxes in which target objects were hidden. (D) Confederate's chair. (E) Participant's chair. (F) Objects available to participant. (G) Detail of the box with combination lock. (H) Target objects. (I) Participant's viewpoint. (J) Confederate's viewpoint.

6.2. Human Subjects Study

We conducted a human subjects study to gather human performance data on our collaborative benchmark tasks.

Figure 9 shows some of the essential elements of our study setup. Target objects were hidden in three flight cases (A), (B), and (C). Our experimental confederate and the study participant were seated opposite each other at locations (D) and (E), respectively. The participant's stock of food objects was located on a stool, (F), adjacent to their chair and out of the reach and view of the confederate. The target objects, (H), were a bright red package of chocolate-chip cookies and a bright blue bag of corn chips. Also shown are the viewpoint from the participant's location, (I), and the viewpoint from the confederate's location, (J); note that the stock of food objects is not visible from this location.

The detail of our box-sealing mechanism is shown in (G). When attempting to open a sealed box, the actor (experimental confederate) systematically tries two digit combinations in numeric order, starting at zero and tugging at the lock with

Table 1. Behavior Demonstrated by Study Participants on Benchmark Tasks.

Task	Correct object	Guidance gesture	Grounding gesture	Other	No action	Incorrect object
Task 1	16	0	0	1†	1	2
Task 2	14	1	2	0	0	3
Task 3	13⋆	5	2	0	0	0
Task 4a	14	2	1	0	3	0
Task 4b	13	0	1	1‡	1	4

⋆ One participant produced the object only after the key had been retrieved from box C.

† Participant successfully pried open the locked target box.

‡ Participant discovered the combination lock code and revealed it gesturally.

each iteration. The correct code was always 21, so the actor could open the lock within 30 to 45 seconds, giving the subject sufficient time to consider the actor's goal and contemplate potential helpful actions, while keeping the experiment running at a reasonable pace.

We gathered data from 20 participants: 11 females and 9 males, with ages ranging from 18 to 65. Our participants were a mix of undergraduates, graduate students, and staff from the MIT community. Participants were each presented with the four benchmark tasks in randomized order. Participants were instructed not to talk to their partner, but were told that they were otherwise free to perform any action or gesture that might help their partner achieve the goal. Participants were instructed that they might find the objects on the stool next to their chair useful, but that they could only use one of these objects per task.

The results of the study are summarized in Table 1. Participant behavior was partitioned into six categories, from most helpful to least helpful: correct object presented, guidance gesture presented, grounding gesture presented, other, no action, incorrect object presented. Behavior was classified as follows. If the participant presented the correct target object to their partner, they were tallied as "correct", and if they presented the wrong object, they were tallied as "incorrect".

Participants who did not present either object were classified according to the gestures that they displayed. "Guidance" gestures included only direct pointing or manipulation towards the correct target box, lock, or key. "Grounding" gestures included bidirectional pointing gestures indicating that the box contents had been swapped, as well as the use of the matching food objects as a "map" to indicate the correct contents of the various boxes. In the absence of such gestures, behavior was tallied as "no action".

Finally, two unexpected cases were tallied as "other" as described in the table notes. It should be noted that in the case of Task 3, guidance gestures were almost as helpful as producing the correct object, since indicating the correct padlocked box or its readily-available key resulted in the rapid acquisition of the contents of the box.

Fig. 10. Leonardo can operate a remote control box to reveal the contents of two boxes located near the human.

These results indicate that participants were largely successful at inferring the goals of their collaborative partners and engaging in helpful behaviors even in the presence of false beliefs, multiple agents, and indirect goal cues.

It should also be noted, however, that success was not uniform. Several participants found some of the tasks to be challenging and reported difficultly in remembering the locations of the hidden objects and the divergent beliefs of their collaborative partners.

6.3. Robot Experiment

In the robot version of the experiment, the robot, Leonardo, interacts face-to-face with one or more human partners (see Figure 10). The physical robot (and its virtual counterpart) is able to exhibit a large repertoire of non-verbal communication cues such as facial expressions, gestures, and gaze shifts. Leonardo can perform simple manipulation tasks in a small workspace with objects specifically designed for the robot. It can understand simple commands using the *Sphinx4* speech

recognition system, but the robot does not speak. It has a number of camera systems to perceive events, objects, and people in its workspace. In this paper, the robot uses a 10-camera Vicon Motion Capture system to robustly track specific objects and particular human features (tagged with reflective markers) in real-time to millimeter accuracy.

In this set of experiments, the robot's goal is to assist the human (or humans) given the actor's goal of obtaining a desired food item. The robot study followed the same protocol as in the human study. Language was not involved in the robot study either. The same objects and actors were used in both studies, with one exception: as the robot lacks sufficient dexterity to pick up and hand objects to the human, the robot was given a remote control panel that it could use to open either of two small metal boxes (one containing chips and the other cookies) near the actor as shown in Figure 10. The actor can then easily retrieve the target object within.

To participate in these tasks, the robot must track multiple objects (the chips, cookies, box lids, locks, etc.) and multiple aspects of human behavior (each person's head pose and hand trajectory) robustly and in real-time. As can be seen in Figure 11, we use a 10-camera Vicon motion capture system to track the trajectories of reflective markers mounted to people and objects involved in the benchmark tasks. The Vicon system was very useful in this regard instead of using traditional video cameras. Extension 1 demonstrates the robot performing Tasks 4a and 4b in real-time with two human partners.

The actors wore a headband and gloves with a distinct pattern of markers so that the robot could distinguish between the different actors as well as track their behavior. Distinct patterns of markers were also placed on each objects used in the study. We developed customized tracking software to enable the robot to uniquely identify each rigid and near-rigid object (via their pattern of markers) to track their position and orientation. The robot must ascribe meaning to these trajectories, e.g. what food items are in which boxes over time, who is witness to which events, who is performing what actions, etc. The robot's perceptual and belief systems are responsible for constructing the robot's cognitive understanding of the scenario as it unfolds in real-time.

Table 2 displays the robot's behavior generated by our architecture on the various benchmark tasks under two conditions. In the first condition, the robot can offer the human a matching target object by operating its remote control box to reveal the correct item inside. In the second condition, the robot does not have its remote control box, so it cannot provide access to matching items. In this case, the robot can help the person by pointing to the location where the desired object really is. Note that this communicative action manipulates the human's beliefs (rather than actions), helping to lead them to their goal.

On Tasks 1 and 4b, the collaborator attempts to open the correct box, so the robot does not need to generate any helpful behaviors. On Tasks 2 and 4a, the robot uses its knowledge of the human's beliefs to infer which object they are trying to acquire. Using this goal in conjunction with its own true knowledge of the world state allows the robot to direct the human to the correct box via a pointing gesture The robot uses the same inferential mechanism on Task 3 to generate a pointing gesture towards the key lying on the table which opens the correct padlocked box.

Table 2. Under the "Remote Control" Condition the Robot Operates its Remote Control Box Interface to Reveal the Correct Matching Item for All Tasks. Under the "Deictic Gesture" Condition, the Robot can only Help the Actor by Pointing to the Correct Location of the Required Item.

Task	Remote control	Deictic gesture
Task 1	Open chips box (correct)	No action
Task 2	Open cookie box (correct)	Points to target location
Task 3	Open chips box (correct)	Points to key
Task 4a	Open chips box (correct)	Points to target location
Task 4b	Open cookies box (correct)	No action

6.4. Summary

While the robot is not able to generate the full range of gestures and actions observed in our human study participants, the self-as-simulator cognitive architecture nevertheless allows the robot to produce helpful behaviors on a number of sophisticated collaborative tasks requiring goal inference in the presence of potentially divergent beliefs. Further, when given its remote control panel, the robot successfully opens the correct box to reveal the matching target item.

Our original objective of performing human subjects experiments was to gather data on the range of human behavior as they solved each task. We then wanted to compare the robot's performance on these tasks to human performance, fully expecting humans to be better. We were surprised at the number of people who did not perform the tasks correctly, and that people found some of these tasks to be difficult. These tasks are not as simple as one might initially think.

In light of our human performance data, it is interesting that our robot can successfully perform these tasks under both conditions.

7. Learning From Demonstration

Mindreading skills play an important role in many forms of skillful social behavior. In the previous section we examined

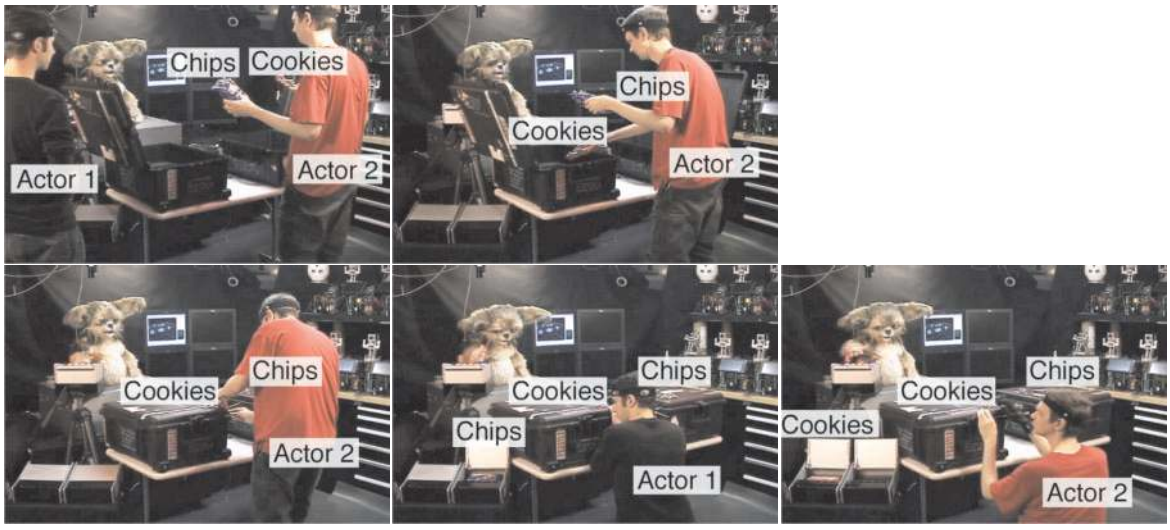

Fig. 11. Setup of the human–robot study for Tasks 4a and 4b. The scenario proceeds from upper left image to bottom right. First, actor 2 places chips in the left box and cookies in the right box for all to see. While actor 1 is absent, actor 2 switches the food items. Actor 1 returns looking for chips, but going to the wrong box. The robot realizes the false belief and invalid plan of actor 1, and gives them the chips they desire. Actor 2 (with true beliefs and a valid plan) returns looking for cookies, and the robot opens the small box revealing matching cookies. See also Extension 1.

the robot's mindreading skills in a cooperative task scenario. As a second domain, we examine the robot's mindreading skills in a social learning context where the robot learns tasks from watching a human teacher. Importantly, the tasks are designed to be intentionally ambiguous, providing the opportunity to investigate how different types of perspective taking might be used to resolve these ambiguities.

For instance, in Section 5 we gave the example of a visual occlusion that blocks the teacher's viewpoint of a region of the shared demonstration area but not the learner's, leading to an ambiguous demonstration where the teacher does not realize the visual information of the scene differs between them. This learning task incorporates a false-belief manipulation in that there are relevant objects in the workspace that the human cannot see but the robot can. In a more subtle situation, both human and robot can see the same workspace, but the teacher focuses their visual attention on a subset of objects in the workspace while ignoring the rest.

To address these kinds of social learning situations, we hypothesize that perspective taking and belief inference integrates with task learning mechanisms, whereby inferring the beliefs of the teacher allows the learner to build task models which capture the intent behind the teacher's demonstrations. In essence, perspective taking acts as a dynamic "social filter" that focuses the hypothesis generation mechanism on the subset of the input space that matters to the human teacher. This enables the learner to successfully learn what the teacher intends to teach despite incompleteness or ambiguity in the observed demonstrations.

To test our hypothesis we devised a benchmark suite of learning tasks where different concepts would be learned depending on whether the learner took the perspective of the human teacher to frame the learning problem, or not. We tested this benchmark suite both on human subjects as well as the robot.

7.1. Benchmark Tasks

Figure 12 illustrates sample demonstrations of each of four tasks. The tasks were designed to investigate how different types of perspective taking might be used to resolve ambiguities in the demonstrations. 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). Figure 13 shows the set of blocks that would be considered to be part of the demonstration consistent with the PT hypothesis or NPT hypothesis for an instance of two of the learning tasks.

Our hypothesis is that human learners engage in perspective taking when the human teacher is present and performing the demonstrations, and would not engage in perspective taking when the human teacher is absent. As a result, they would learn *different* rules depending on whether the *same* learning examples are embedded in a social context or not.

Task 1 focused on visual perspective taking during the demonstration. Participants were shown two demonstrations with blocks with different configurations. The workspace had

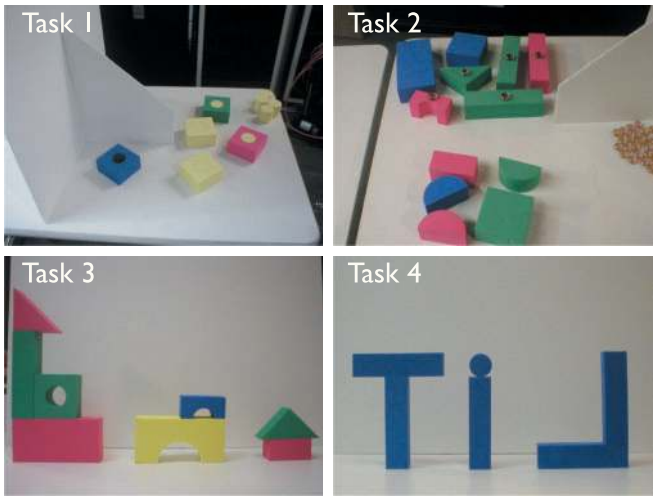

Fig. 12. The four tasks demonstrated to participants in the study (photos taken from the participant's perspective). Tasks 1 and 2 were demonstrated twice with blocks in different configurations. Tasks 3 and 4 were demonstrated only once.

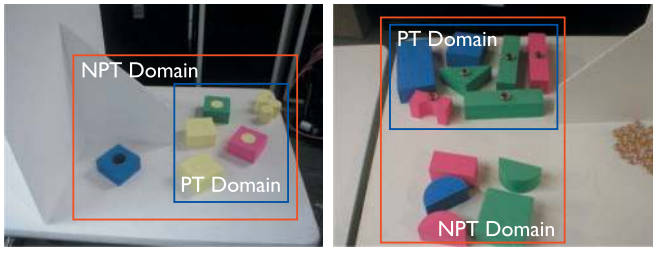

Fig. 13. Input domains consistent with the perspective taking (PT) versus non-perspective taking (NPT) hypotheses. In visual perspective taking (left image), the student's attention is focused on just the blocks that the teacher can see, excluding the occluded block. In resource perspective taking (right image), attention is focused on just the blocks that are considered to be "the teacher's", excluding the other blocks.

square blocks in different colors, each with a circular hole cut out of its center, and a large pile of circular pegs. In the social condition (see Figure 14) for 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 subject but was occluded from the view of the teacher by a barrier. The hole of this blue block was never filled by the teacher. None of the other blocks involved in the demonstrations were blue. In the non-social condition, the subject was shown images on a computer of the same end configuration of blocks from the learner's viewpoint. 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 Figure 13).

Task 2 focused on resource perspective taking and focused visual attention during the demonstration. Again, participants were shown two demonstrations with blocks in different configurations. The blocks had seven different shapes and three different colors: red, green, and blue. In the social condition in both of the demonstrations, the teacher placed marble beads on some of the blocks. 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, and the teacher was careful to only attend to "their" blocks during the demonstration.

In both demonstrations, the teacher placed markers on only "their" red and green blocks, ignoring all of the participant's blocks. Owing to 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 Figure 13).

Tasks 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. Figure 12 shows the examples that were constructed by the teacher. In the social condition for 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 whether people would perform any such perspective flipping in a non-linguistic scenario.

8. Human Subjects Study

We conducted a human subjects study for two purposes. First, to gather human performance data on a set of learning tasks that were well matched to our cognitive architecture's existing perceptual and inferential capabilities. This allows us to compare our system's behavior with human behavior on the same benchmark suite. Second, the study served to investigate the role of perspective taking in human learning. When we began this study, we were not sure what outcome to expect given that

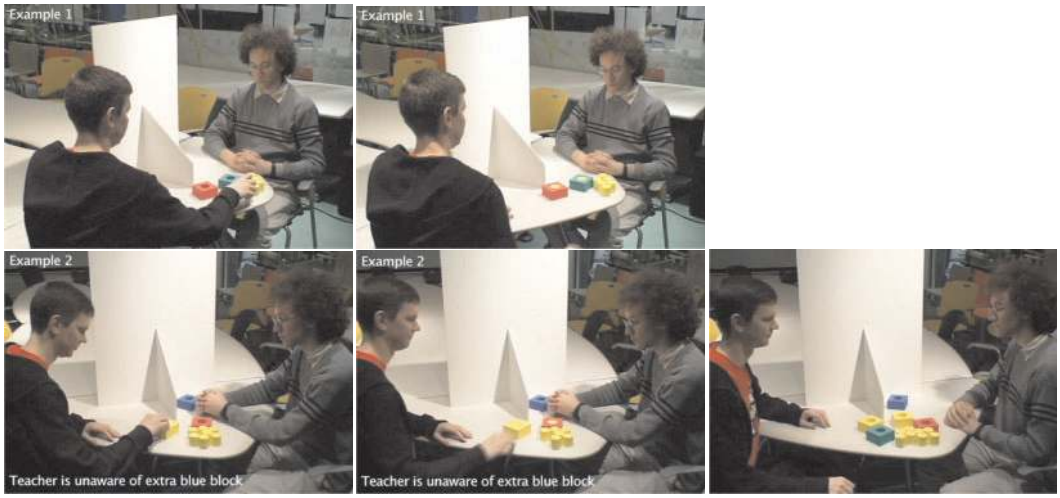

Fig. 14. A learning from demonstration task. Image sequence proceeds upper left to bottom right. The top two photos shows the first example from the teacher's viewpoint. The middle row shows the second example where the subject's viewpoint is also displayed. Note that the student can see a blue block that is occluded from the teacher's viewpoint. The bottom image on the far right shows a novel configuration of blocks for the subject to demonstrate their learned rule.

we could not find a similar experiment in the psychological literature.

Study participants were asked to engage in the four different learning tasks involving foam building blocks. We gathered data from 41 participants: 19 females and 22 males, with ages ranging from 18 to 40. The participants were a mix of undergraduates, graduate students, and staff from the MIT community. The participants were divided into two groups. In the social condition, 20 participants observed demonstrations provided by a human teacher (an experimental confederate) sitting opposite them (see Figure 14). In the non-social condition, 21 participants were shown static images of the same demonstrations on a computer screen with the teacher absent from the scene. Participants were asked to show their understanding of the presented skill either by reperforming the skill on a novel set of blocks (in the social condition) or by selecting the best matching image from a set of possible images (in the non-social condition).

The results of the human subjects study are summarized in Table 3 where participant behavior was recorded and classified according to the exhibited rule. For every task, differences in rule choice between the social and non-social conditions were highly significant (chi-square, $p < 0.005$ or $p < 0.001$ as in Table 3).

Table 4 displays the rules selected by study participants, with the most popular rules for each task highlighted in bold. Note that, while many participants fell into the "Other" category for Task 1, there was very little rule agreement between these participants. These results strongly support the intuition (and our hypothesis) that perspective taking plays an important role in human learning in socially situated contexts.

Table 3. Differential Rule Acquisition for Study Participants in Social versus Non-social Conditions (* $p < 0.001$).**

Task	Condition	PT rule	NPT rule	Other	p
Task 1	Social	**6 (30%)**	1 (5%)	13 (65%)	
	Non-social	1 (5%)	**12 (57%)**	8 (38%)	***
Task 2	Social	**16 (80%)**	0	4 (20%)	
	Non-social	7 (33%)	**12 (57%)**	2 (10%)	***
Task 3	Social	**12 (60%)**	8 (40%)	—	
	Non-social	0	**21 (100%)**	—	***
Task 4	Social	**14 (70%)**	6 (30%)	—	
	Non-social	0	**21 (100%)**	—	***

Table 4. Hypotheses Selected by Human Study Participants. The Most Popular Rule for Each Task is Highlighted in Bold.

Task	Condition	Hypotheses selected
Task 1	Social	***All**; number; spatial arrangement*
	Non-social	***All but blue**, spatial arrangement; All but one*
Task 2	Social	***All red and green**; shape preference; Spatial arrangement*
	Non-social	***Shape preference**; all red and green*
Tasks 3 and 4	Social	***Rotate figure**, mirror figure*
	Non-social	***Mirror figure***

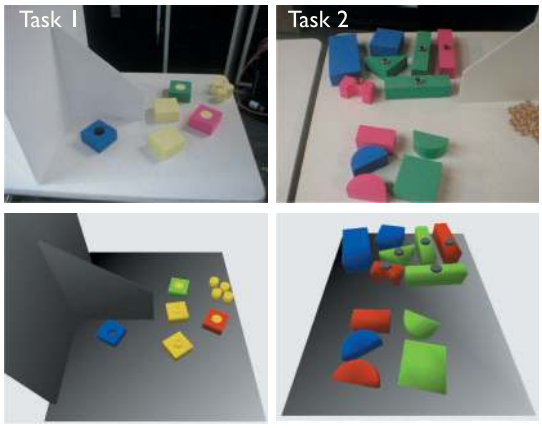

Fig. 15. Virtual Leonardo was presented with the same learning tasks as our human subjects in a simulated environment. The examples for Tasks 1 and 2 are shown. The human demonstrator can use the mouse to grab and place the pegs into the holes of the blocks for Task 1, or to place beads on top of blocks for Task 2. The virtual robot simulates the virtual visual perspective of the teacher.

9. Robot Experiment

The next question is how the behavior of our cognitive architecture compares with our human data. The architecture and learning implementation were developed *prior* to our human subjects experiment. When executed by our system, would its learning behavior predict human behavior? To investigate this question, its learning performance was analyzed under two conditions: with the perspective taking mechanisms intact (to mirror the social condition in the human study), and with them disabled (to mirror the non-social condition).

We used the same tasks and protocols from our human subjects study to test our architecture. As our physical robot lacks the dexterity to perform the object manipulations necessary to demonstrate the learned concepts, we conducted these experiments with our simulated robot in a virtual world running on a desktop computer. The simulated robot runs the same cognitive architecture as its physical counterpart, and interacts with the human demonstrator in real-time where the teacher can move the objects using the computer mouse to perform the same demonstrations. We simulated the tasks and protocols as accurately as possible to preserve the spatial relationships, visual perspectives, and workspace configuration used in the human study (see Figure 15).

Table 5 shows the hypotheses entertained by the robot under the various task conditions at the conclusion of the demonstrations. The hypotheses favored by the learning mechanism are highlighted in bold. As an example, Figure 16 illustrates an interaction run for the social condition.

Table 5. Robot Hypotheses on Benchmark Tasks.

Task	Condition	Hypotheses considered
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***
Tasks 3	With PT	***Rotate figure****, mirror figure*
and 4	Without PT	***Mirror figure***

9.1. Summary

For comparison, Table 4 displays the rules selected by our human study participants, with the most popular rules for each task highlighted in bold. 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 able 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 currently 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 seem to be extremely successful at bringing the agent's focus of attention into alignment with the humans' in the presence of a social teacher.

It was a pleasant surprise that our results with our robot predicted human performance on this set of tasks. We were not expecting to see this amount of agreement between the behavior produced by our cognitive architecture and our human subjects.

10. Discussion

We have tested and evaluated our self-as-simulator cognitive architecture in two different domains: cooperative behavior and learning from demonstration.

As we discussed previously, one of the most important milestones in children's ToM development is gaining the ability to attribute false belief. This may entail understanding how knowledge is formed, that people's beliefs are based on their knowledge, that mental states can differ from reality and from one's own, and that people's behavior and intention can be predicted by their mental states. Our self-as-simulator architecture

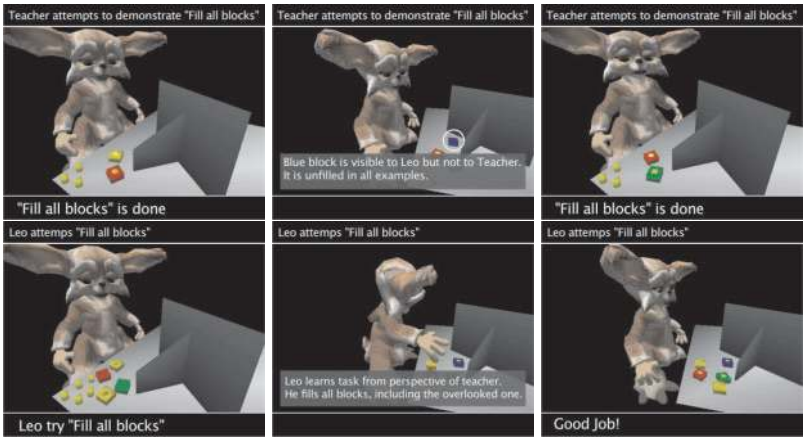

Fig. 16. The teaching scenario for Task 1 with the virtual robot. The top left and top right images show the visual perspective for the human demonstrator. Note how the blue block is occluded from the teacher's view as in the human subjects study. Image sequence proceeds upper left to bottom right.

tackles each of these using a combination of bottom-up and top-down processes.

At the perceptual level, simulation-theoretic mechanisms for visual perspective taking are used to represent what a person can see as a function of body location and head orientation. Hence, visually accessible knowledge is constructed from the bottom-up and sent to the belief system that uses this information to simulate the other's corresponding beliefs (which could be false depending on the perceptually derived knowledge). Meanwhile, mirror-neuron inspired mechanisms in the motor system map the human's observed hand trajectories onto the robot's body to simulate and recognize goal-directed actions, such as reaching movements. These two sources of bottom-up knowledge (perceptual and motor) are sent to the higher-level intention system where task knowledge in the form of hierarchical task models are run in simulation using inferred beliefs and recognized actions to deduce possible desires and goals of the human partner. Meanwhile, in real-time and in parallel, the robot uses the same systems to generate its own perceptions, beliefs, actions, desires and goals, all represented in the same way as those it generates for its human partner.

Depending on the domain and task, the architecture can then reason about and apply these mental states (of self and from others) in different ways. The reasoning process also takes the form of a simulation. In the case of the cooperative tasks, these estimated human beliefs reveal their desired target and impact their intentions to obtain it. Given the goal of helping the human obtain their desired target, the robot can reason about a course of action (simulating within the intention system) given the possible actions it can take. When the robot has access to its control panel, its most direct way of helping the person is to open the box revealing a matching target item. If the robot does not have access to its control panel, it can still help its partner by giving them the knowledge (i.e. pointing to the correct box) that they need to obtain the target for themselves. In the case of learning tasks, the robot uses the beliefs it has inferred from the human teacher to reframe the input demonstrations. By simulating the same learning mechanisms on the reframed demonstrations, the robot infers the rule the human teacher intended to demonstrate.

Thus, in different ways, both domains test the architecture's ability to attribute beliefs (possibly false) to others, to consider the resulting implications on their behavior (e.g. plans, possibly invalid), and to shape its own behavior accordingly. Importantly, these inferences are made from observing a person's real-time, non-verbal behavior as they try to obtain their goal or instruct a task.

Our cross-domain assessment highlights our ultimate goal of endowing social robots with *a way of thinking about and relating to the social world of humans*, inspired and informed by human mindreading skills. We have demonstrated how our architecture can apply the same set of mental state inferences to an assortment of tasks designed to probe different kinds of inferences in two very different domains (task assistance and social learning). Importantly, the robot does not simply infer task-specific "hidden states" (e.g. hidden Markov models) from observing behavior, but rather those specific mental states that are believed to underly human behavior across many different domains (e.g. beliefs, intents, desires, etc.).

This is an important capability for personal robots that need to understand not only what people say but also *what they do*. Many prior works have developed techniques for symbolic or language-based domains (as discussed in Section 2). This work, however, demonstrates how embodied mechanisms (grounded in the robot's own perceptual, motor, and behavior-generation processes) can address the latter. Given the growing scientific evidence that early ToM abilities and critical precursors develop from more embodied processes such as mirror

neurons, imitation, simulation, and perspective-taking, developing computational models of these processes is scientifically significant to understand. Toward this goal, it is exciting to view social robots as an experimental platform to develop and test models and theories to understand human mindreading skills. The ability to test and compare robots and humans on the same tasks following the same protocols is potentially very powerful. As future work, designing new experiments and tasks that allow researchers to compare embodied computational models (i.e. social robots) with prelinguistic children who are developing ToM, or even with other species (e.g. chimpanzees) to explore comparative psychological models, is also exciting.

Furthermore, because we are particularly interested in personal robots that assist and learn from humans, it matters that the robot's socio-cognitive abilities are compatible with those of humans. In short, the kinds of inferences the robot makes and the behaviors it performs need to make intuitive sense to people: the robot needs to be consistent with the theory of mind people ascribe to it. Hence, our human subject studies and "cross-species" assessment are an important contribution toward this goal. In both domains, we found the robot's behavior to be consistent with human behavior. In the cooperative task domain, we were surprised that adults found the tasks to be as challenging as they did. In the learning domain, we were pleasantly surprised to find evidence to support that humans engage in perspective taking to learn from ambiguous demonstrations. We were also pleasantly surprised to find that our architecture was highly predictive of the most popular rules the humans also learn. Hence, our human subject studies are not only an interesting way to assess our system, they also reveal insights into human behavior.

11. Conclusion

Many applications for personal robots require them to be socially intelligent and skillful in their interactions with humans. We argue that personal robots need mindreading skills as a way of being able to think about and relate to the social world of humans. Inspired by human theory of mind, its development and precursors, we have developed a novel integrated architecture, informed by recent scientific findings in embodied cognition and neuroscience for how people are able to take the perspective of others. Accordingly, simulation-theoretic mechanisms serve as the organizational principle for our robot's perspective taking skills over multiple system levels (e.g. perceptual-belief, action–goal, task learning, etc.).

We have evaluated our architecture on a novel benchmark suite and showed that our anthropomorphic robot can apply mindreading skills across two different domains: (1) to assist a human despite their false beliefs and invalid plans in a cooperative setting; or (2) to draw the same conclusions as humans under conditions of high ambiguity in a learning setting. We also performed two novel human subjects studies, that both revealed insights into human behavior as well as providing an important point of comparison for the robot's behavior. This motivates further research into the use of social robots as a flexible experimental testbed for embodying models and theories of human ToM and testing them using the same tasks and protocols used for humans and other species.

Robotic systems that aim to collaborate effectively with humans in dynamic, social environments must be able to respond flexibly to the intentions of their human partners, even when their collaborators' actions are based on false or incomplete beliefs. The architecture enables our robot to infer the task-related beliefs and intentions of its interaction partners based on their observable motor behavior and visual perspective. This is the first work to show how an embodied-cognitive approach enables a real-world robot to produce appropriate behavioral responses in complex collaborative scenarios involving a human partner's divergent, false beliefs and invalid plans. Significantly, the robot is able to make these inferences from watching what people do, rather than from what they say. The non-verbal demands of our tasks highlight the important role that embodied, bottom-up simulation-theoretic mechanisms play in ascribing knowledge, belief, and intents to others.

In addition, 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 and Sidner 1990; Gorniak 2005). Others have looked at the use of visual perspective taking in collaborative settings (Jones and Hinds 2002; Cassimatis et al. 2004). We also draw inspiration from research into learning from humans, which typically focuses on either modeling a human via observation (Lashkari et al. 1994; Horvitz et al. 1998) or on learning in an interactive setting Atkeson and Schaal (1997); Lieberman (2001); Nicolescu and 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.

In our learning experiments, in particular, the behavior of the architecture was surprisingly predictive of human performance. Specifically, we found evidence that people use perspective taking to entertain a different set of hypotheses when demonstrations are presented by another person, versus when they are presented in a non-social context. This data supports that 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. This finding has interesting implications for the role of mindreading skills in machine learning systems that are intended to learn from people in human environments.

Acknowledgements

The work presented in this paper is a result of ongoing efforts of the graduate and undergraduate students of the MIT Media Lab Personal Robots Group. This work is funded by the *Digital Life* and *Things That Think* consortia of the MIT Media Lab, and supported by the Office of Naval Research (YIP grant N000140510623). Jesse Gray would like to thank Samsung Electronics for their support. We would also like to thank Roman Zadov, Amy Shui, Andrew Brooks, Mikey Siegel, Guy Hoffman, Andrea L. Thomaz, Alea Teeters, and Stefanie Tellex for their assistance with our study. IRB approved protocols were used in our human subject studies.

Appendix: Index to Multimedia Extensions

The multimedia extension page is found at http://www.ijrr.org

Table of Multimedia Extensions

Extension	Type	Description
1	Video	Demonstration of Task 4a and 4b.

References

Arulampalam, M. et al. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, **50**(2): 174–188.

Atkeson, C. G. and Schaal, S. (1997). Robot learning from demonstration. *Proceedings of the 14th International Conference on Machine Learning*. Morgan Kaufmann, pp. 12–20.

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. *Natural Theories of Mind*, Whiten, A. (ed.). Oxford, Blackwell, pp. 233–250.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. and Ruppert, J. (2003). Social embodiment. *The Psychology of Learning and Motivation*, Vol. 43. New York, Academic Press.

Berlin, M., Gray, J., Thomaz, A. and Breazeal, C. (2006). Perspective taking: an organizing principle for learning in human–robot interaction. *Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*.

Billard, A., Epars, Y., Calinon, S., Schall, S. and Cheng, G. (2004). Discovering optimal imitation strategies. *Robotics and Autonomous Systems*, **47**: 69–77.

Breazeal, C., Berlin, M., Brooks, A., Gray, J. and Thomaz, A. L. (2006). Using perspective taking to learn from ambiguous demonstrations. *Journal of Robotics and Autonomous Systems*, **54**(5): 385–393.

Breazeal, C., Buchsbaum, D., Gray, J. and Blumberg, B. (2005). Learning from and about others: toward using imitation to bootstrap the social competence of robots. *Artificial Life*, **11**(1–2): 31–62.

Brooks, A. G., Berlin, M., Gray, J. and Breazeal, C. (2005). Untethered robotic play for repetitive physical tasks. *Proceedings ACM International Conference on Advances in Computer Entertainment*.

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

Carpenter, J., Clifford, P. and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *Radar, Sonar and Navigation, IEE Proceedings*, **146**(1): 2–7.

Cassimatis, N. L., Trafton, J. G., Bugajska, M. D. and Schultz, A. C. (2004). Integrating cognition, perception and action through mental simulation in robots. *Journal of Robotics and Autonomous Systems*, **49**(1–2): 13–23.

Cohen, P. R., Levesque, H. J., Nunes, J. H. T. and Oviatt, S. L. (1990). Task-oriented dialogue as a consequence of joint activity. *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, Nagoya, Japan.

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

Deak, G. O., Fasel, I. and Movellan, J. (2001). The emergence of shared attention: using robots to test developmental theories. *Proceedings of the 1st International Workshop on Epigenetic Robotics*.

Demiris, J. and Hayes, G. (2002). Imitation as a dual-route process featuring predictive and learning components: a biologically plausible computational model. *Imitation in Animals and Artifacts*, Dautenhahn, K. and Nehaniv, C. L. (eds). Cambridge, MA, MIT Press, 321–361.

Downie, M. (2000). Behavior, animation, and music: The music and movemnt of synthetic characters. *Master's Thesis*, MIT, Cambridge, MA.

Emond, B. and Ferres, L. (2001). Modeling the false-belief task: an ACT-R implementation of Wimmer & Perner (1983). *Second Bisontine Conference for Conceptual and Linguistic Development in the Child Aged from 1 to 6 Years*.

Fasel, I., Deak, G. O., Triesch, J. and Movellan, J. R. (2002). Combining embodied models and empirical research for understanding the development of shared attention. *Proceedings of the International Conference on Development and Learning*.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, **2**(12): 493–501.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, **1**: 158–171.

Gorniak, P. (2005). The affordance-based concept. *PhD Thesis*, MIT, Cambridge, MA.

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

Grosz, B., Hunsberger, L. and Kraus, S. (1999). Planning and acting together. *The AI Magazine*, **20**(4): 23–34.

Grosz, B. J. and Sidner, C. L. (1990). Plans for discourse. *Intentions in Communication*, Cohen, P. R., Morgan, J. and

Pollack, M. E. (eds). Cambridge, MA, MIT Press, pp. 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. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, Madison, WI, 256–265.

Intille, S. S. and Bobick, A. F. (1999). A framework for recognizing multi-agent action from visual evidence. *AAAI-99*. AAAI Press, pp. 518–525.

Jenkins, O. C. and Matarić, M. J. (2002). Deriving action and behavior primitives from human motion data. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2551–2556.

Johnson, M. and Demiris, Y. (2005). Perceptual perspective taking and action recognition. *International Journal of Advanced Robotic Systems*, **2**(4): 301–308.

Jones, H. and Hinds, P. (2002). Extreme work teams: using swat teams as a model for coordinating distributed robots. *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*. New York, ACM Press, 372–381.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**(1): 35–45.

Laird, J. E. (2001). It knows what you're going to do: adding anticipation to a quakebot. *AGENTS '01: Proceedings of the 5th International Conference on Autonomous Agents*. New York, ACM Press, pp. 385–392.

Lashkari, Y., Metral, M. and Maes, P. (1994). Collaborative interface agents. *Proceedings of the 12th National Conference on Artificial Intelligence*, Vol. 1. Seattle, WA, AAAI Press.

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

Meltzoff, A. N. (2005). Imitation and other minds: the "like me" hypothesis. *Perspectives on Imitaion: From Neuroscience to Social Science*, Vol. 2, Hurley, S. and Chater, N. (eds). Cambridge, MA, MIT Press, pp. 55–77.

Meltzoff, A. N. and Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society: Biological Sciences*, **358**: 491–500.

Movellan, J. and Watson, J. (2002). The development of gaze following as a Bayesian systems identification problem. *Proceedings of the International Conference on Development and Learning*.

Nagai, Y., Asada, M., and Hosoda, K. (2002). Developmental approach accelerates learning of joint attention. *Proceedings of the International Conference on Development and Learning*.

Needham, C. J., Santos, P. E., Magee, D. R., Devin, V., Hogg, D. C. and Cohn, A. G. (2005). Protocols from perceptual observations. *Artificial Intelligence*, **167**(1–2): 103–136.

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

Pollack, M. (1990). Plans as complex mental attitudes. *Intentions in Communication*, Cohen, P. R., Morgan, J. and Pollack, M. (eds). Cambridge, MA, MIT Press, pp. 77–103.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences*, **1**: 515–526.

Rao, A. S. and Murray, G. (1994). Multi-agent mental-state recognition and its application to air-combat modeling. *Proceedings of the 13th International Workshop on Distributed Artificial Intelligence (DAI-94)*, Seattle, WA, pp. 283–304.

Rizzolatti, G., Fadiga, L., Gallese, V. and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, **3**: 131–141.

Ronnie, L., Johansson, M. and Suzic, R. (2005). Particle filter-based information acquisition for robust plan recognition. *Eighth International Conference on Information Fusion*.

Scassellati, B. (2001). Distinguishing animate from inanimate visual stimuli. *Proceedings of the International Joint Conference on Artificial Intelligence*.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, **12**: 13–24.

Schaal, S. (1997). Learning from demonstration. *Advances in Neural Information Processing Systems*, Vol. 9, Mozer, M., Jordan, M. and Petsche, T. (eds). Cambridge, MA, MIT Press, pp. 1040–1046.

Tapus, A., Mataric, M. and Scasselatti, B. (2007). The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine*, **4**: 35–42.

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.

Trafton, J. G., Schultz, A. C., Perzanowski, D., Bugajska, M. D., Adams, W., Cassimatis, N. L., and Brock, D. P. (2006). Children and robots learning to play hide and seek. *HRI '06: Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction*. New York, ACM Press, pp. 102–109.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: representation and constraining function on wrong beliefs in young children's understanding of deception. *Cognition*, **13**: 103–128.