Int J of Soc Robotics (2014) 6:315–327
DOI 10.1007/s12369-014-0234-2

Manipulating Mental States Through Physical Action

A Self-as-Simulator Approach to Choosing Physical Actions Based on Mental State Outcomes

Jesse Gray · Cynthia Breazeal

Accepted: 1 March 2014 / Published online: 13 April 2014

Abstract We present our implementation of a self-as-simulator architecture for mental state manipulation through physical action. The robot attempts to model how a human's mental states are updated through their visual perception of the world around them. This modeling, combined with geometrically detailed, perspective correct simulations of the immediate future, allows the robot to choose actions which influence the human's mental states through their visual perception. The system is demonstrated in a competitive game scenario, where the robot attempts to manipulate the mental states of an individual in order to win. We evaluate participants' reaction to the system, focusing on their perception of a robot with mental state manipulation capabilities.

Keywords Human robot interaction · Mental state manipulation · Perspective taking

1 Introduction

This paper focuses on a demonstration of mental state manipulation in a competitive game scenario and an evaluation of human perceptions of this behavior. The motivation for this work is to explore the connection between (hidden) mental states of an embodied agent and the (observable and modifiable) world in which that agent exists: in particular, how can an agent modify inaccessible mental states of another agent by manipulating the accessible world the two agents share.

"Mental state manipulation through physical action" may sound unusual, however consider: every external action an embodied agent takes is a physical action; if an agent does cause a mental state change in another agent, it will be through these actions. Agents can only act on the physical world (pointing, speaking, waving), but the goal of these communicative actions is often not simply the physical action of moving the arm or producing speech. Rather, those physical actions are mechanisms used to attempt to change a mental state in the viewer/listener. A physical action alone will not directly change another's mental state; mental states are only changed by the agent itself, based on its observations of the world and its own internal mental processes. To attempt a mental state change, an agent determines how to alter the world so that an observer's perceptual and mental processes will bring about that change.

The implementation described here focuses on manipulating mental states in this way by choosing the correct physical actions. Every action has potential mental state consequences for surrounding agents. These changes are never direct; instead, the actions can change the world, and the perception of the changed world may cause agents to update their mental states. To intentionally alter mental states at this level of detail, our strategy is to have the agent model both of these mappings:

1. *Action Simulation* (mental state → world) once an agent has chosen to perform an action, how will that action alter the world.

2. *Mental State Simulation* (world → mental state) how will the changes to the world alter the mental states of other agents.

Electronic supplementary material The online version of this article (doi:10.1007/s12369-014-0234-2) contains supplementary material, which is available to authorized users.

J. Gray (✉) · C. Breazeal
Cambridge, MA, USA
e-mail: jg@media.mit.edu

C. Breazeal
e-mail: cynthiab@media.mit.edu

Using these mechanisms, the agent can perform *Mental State Manipulation* by choosing and performing a set of actions which bring about its mental state goals.

While speaking is a very useful physical action for conveying mental states, mental state manipulation does not require the use of language; the demonstration described here focuses on visual perception and manipulation of simple "world-modeling" mental states.

The mental states used in this implementation are mental states the agents hold to internally model the physical world around them such as object locations, properties, and relationships (e.g., "Red cup is being held by Agent2 at x, y, z"). Thus the *Mental State Simulation* presented here focuses on how the visual perception systems of other agents will update their "world-modeling" mental states given the state of the world and their visual perspective and occlusions (e.g., an agent is modeled as knowing "Red cup at x, y, z" if and only if that agent appears to have seen the cup there). The *Action Simulation* focuses on how the agents' actions alter the visible features of these objects, as well as altering the perceptual context such as perspectives, occlusions, etc. (e.g., during action performance an object may be visibly moved, or the agent's body might temporarily occlude an object from the perspective of an observer). By combining *Action Simulation* with *Mental State Simulation*, potential future actions can be evaluated according to their predicted effects on the mental states of other agents (e.g., what will Agent2 see if I perform X). This implementation's *Mental State Manipulation* mechanism explores the space of possible future action sequences to find a sequence of actions that achieves a desired set of mental state goals

The following section (Sect. 2) motivates this implementation with research on human behavior and situates it within the space of robotic systems in the areas of mental perspective taking and manipulation. We present a demonstration scenario in Sect. 3 which illustrates the mental state manipulation capabilities of our system through a competitive game played with a human. Section 4 describes our self-as-simulator mental state manipulation implementation. In Sect. 5 we present the results of a human-subjects study, which found that participants were more willing to team with and attribute mental states to a robot that can perform mental state manipulation through physical actions.

2 Background

2.1 Human Perspective Taking

The ability of humans to perceive hidden mental states of others is well studied. Researchers have shown that humans can determine the goals behind observed actions [18], and that similar brain responses occur upon performing one's own actions as well as observing the actions of others [20]. People are also able to both infer certain mental states of others based on geometrically correct perception models and maintain that model even when it differs from one's own [26,27]. Furthermore, inferring another's mental state can help with communication: researchers have shown that people resolve the meaning behind ambiguous communications by attributing a communicative goal to the speaker (alternative meanings that could have supported more specific communications can be eliminated) [7]. The use of visual and mental perspective taking facilitate many human-human interactions and communications, and we believe that endowing a robot with these skills is crucial for allowing a robot to communicate and interact naturally with humans.

2.2 Perspective Taking and Manipulation Systems

One strategy to implement perspective taking is through re-use of one's own systems; e.g., by re-using perceptual or mental processing mechanisms to conduct a simulation of the other agent. Cassimatis et al. even show how certain probabilistic and logical inference strategies could be implemented as perceptual simulations [5]. While powerful, certain problems which appear to rely on these mechanisms can be solved without simulation or even perspective taking: Trafton et al. demonstrate a robot that learns "hide and seek" without relying on perspective taking [22,23]. Nevertheless, our strategy is to enable the robot to perform perspective taking through simulations which re-use the robot's own systems to simulate possible mental states in others.

Researchers have shown the usefulness of detailed perceptual modeling and visual perspective taking in a number of domains, such as to improve the accuracy of activity recognition [11], to resolve ambiguities in an operator's command [24], to approach a target while hiding from sight [14], and to recognize a human's action by comparing it to the robot's own library of first-hand experience [12].

Others have shown the value of using perspective taking to simulate not only the sensory perspective but also the decision making of another agent to predict their next action in competitive [16] and cooperative [13] scenarios.

Building on mental perspective taking used for inference, researchers have also created systems which incorporate mental state manipulation. In a simulated school setting, researchers have shown agents that can plan actions based on complex mental state consequences including affecting nested beliefs (beliefs about beliefs) [17]. However, simulated systems tend to abstract away the connection between mental states and physical action, instead providing the agents with actions that have well understood mental state outcomes. Our work focuses on modeling the mental state

consequences of physical, situated action performance, and using these actions to achieve the desired results.

Moving from the virtual to physical world, others have demonstrated robotic systems that determine the actions necessary to modify the action choice or performance of another agent in a desired way. In a cooperative example, a robot chooses how to hold out its hand when receiving an object based on a model of how this action will modify the "give" action of the participating human [19]. In a deceptive example [25], a hiding robot attempts to mislead the seeker by first experimentally determining the visible evidence that would cause the seeker to check a particular location (e.g., visible tracks on ground), then producing that evidence while hiding in a different location. In both these systems, the robot's goal is to change the action the other agent will perform. To do this, the robot takes action to alter the world, which then causes the other agent to perform the desired action. We focus on a sub-part of this problem: given a desired mental state change, how can the robot act to cause that mental state change in the observer. In contrast to these systems, we do not have an a priori perceptual effect associated with each action; instead, we simulate the physical performance of the action in the current spatial context to capture the intended and unintended perceptual consequences of the action for the observer and the resulting mental state changes ("place object on table" may not convey "object on table" to the observer if the robot's hand blocks the view). It is our belief that these techniques are complimentary; a fielded system could benefit from using the above mechanisms for cases where the perceptual effects of certain actions are pre-determined, avoiding the extra time cost of performing the geometric simulations presented here.

2.3 Perceptions of Adversarial Robots

Other researchers have also studied human reactions to a competitive robot that defeats them in an unexpected manner; researchers have shown that a robot that openly cheats to win in a game of "rock paper scissors" elicits greater attributions of mental states from the participants [21]. Rather than cheating, we focus on reactions to a robot capable of mental state manipulation in a deceptive scenario. We embrace the definition of deception as "the process by which actions are chosen to manipulate beliefs so as to take advantage of the erroneous inferences" [6]. We focus on the subproblem of how a robot chooses its actions so as to achieve a specific belief manipulation, given that such an interaction is mediated by the physical world. A more detailed definition could include "the process by which actions are chosen to [alter the physical world such that a perceiver's belief is updated in a targeted manner]." We demonstrate these abilities using a deceptive scenario, however the mechanisms for this subproblem can apply equally well to a cooperative scenario.

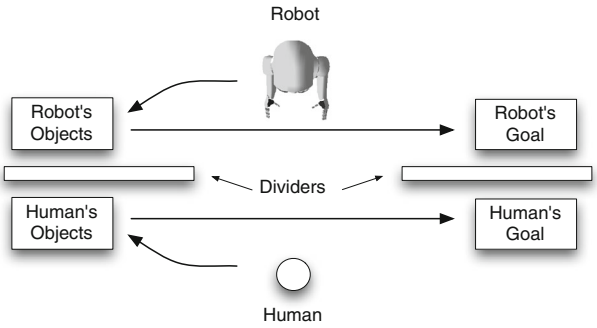

Fig. 1 *Top down* view of the demonstration scenario, a competitive game between the human and the robot. The robot stays on the *upper part* of the diagram pictured, and the human on the *lower part*. Each player has access to a matching set of objects on the *left side*, and each has their own goal area on the *right side*. The game ends when each player has placed an object into their goal—the robot wins if the two players placed different objects, the human wins if the objects are the same. Occlusions block the view of each player from the opposing player's object and goal areas, however they can see each other as they travel between the object repository and goal

3 Demonstration Scenario

The ability to form and act on mental state goals is a key ingredient to a robot's ability to robustly communicate and to maintain common ground. However, in a competitive scenario this same ability may be called deception. In both cases the robot intentionally modifies the world (taking into account the perceptual capabilities of the target agent) to produce desired mental state effects in another agent, however in a competitive scenario those mental states may be divergent from the ground truth or otherwise to the disadvantage of the observer. A competitive scenario was chosen for this demonstration because it fully exercises the system in a relatively simple scenario (see Sect. 6 for discussion).

The architecture for mental state manipulation presented here makes heavy reuse of the motor actions and perceptual processing that the robot uses for its own behavior. As such, it can apply to varied contexts, as long as the robot is configured to perceive and operate there. The scenario chosen for this demonstration revolves around a simple competitive game played between the human and the robot. The game is illustrated in Fig. 1.

The rules of this game create a situation where the player who goes second has the advantage of potentially seeing the item played by their opponent. If the human goes second and sees the item played by the robot, it is straightforward for them to win by playing the object they saw the robot play.

For this demonstration, the robot takes its turn first. It is thus to the robot's advantage to manage the information that can be observed from its behavior. If it proceeds in a straightforward manner, the human will be able to watch and observe the object the robot plays, then play the same object

Condition	Robot's Goals	Robot's Behavior
1	•Cylinder in goal •Human doesn't see me carry cylinder •Human sees me carry football	•Transports cylinder behind back •Carries decoy football
2	•Cylinder in goal •Human doesn't see me carry cylinder	•Transports cylinder behind back
3	•Cylinder in goal	•Transports cylinder openly

Fig. 2 Set of robot's goals and resulting behavior for each of three demonstration conditions

and win. To win, the robot must instead hide this information from the human.

For the demonstration, the game was played three different times, each time with a different set of mental state goals for the robot (see Fig. 2). These different mental state goals change the behavior of the robot as it plays the game. In each case, the robot has the same overall task goal—transport the cylinder to the goal location. However, the way it accomplishes this task varies in the three conditions based on the mental state goals.

In condition one, the robot attempts to cause the human to believe that the robot is transporting only the football, while actually additionally transporting (and playing) the cylinder. The robot finds that carrying the cylinder hidden behind its back, with the football carried out in the open, satisfies these conditions. In this way it may fool the human into thinking that the robot is playing the football, causing the human to lose by playing the football in response. In condition two, the robot's goal is to keep the cylinder (which it is transporting) hidden from the human. The chosen action sequence results in carrying the cylinder with its left hand, hidden behind its back from the human. The human can't see what the robot played, so is likely to choose arbitrarily and win half the time (there are two possible object choices). In the final case, the robot has no mental state goals, and therefore its only goal is to transport the cylinder. It simply carries the cylinder over to the goal (likely causing the robot to lose in this case). The robot's performance of these three conditions is shown in Fig. 3. Section 4 describes our system used to find the action sequences which correctly manipulate the mental states.

4 Implementation

The implementation described here builds on the existing R1D1 system, originally designed for interactive graphical characters [1,4], then later adapted for robots [9,10]. The system employs self-as-simulator techniques for theory of mind tasks, and previous publications have described work in mental state modeling, perspective taking, and goal inference using this system (please refer to [2,3,8] for more details).

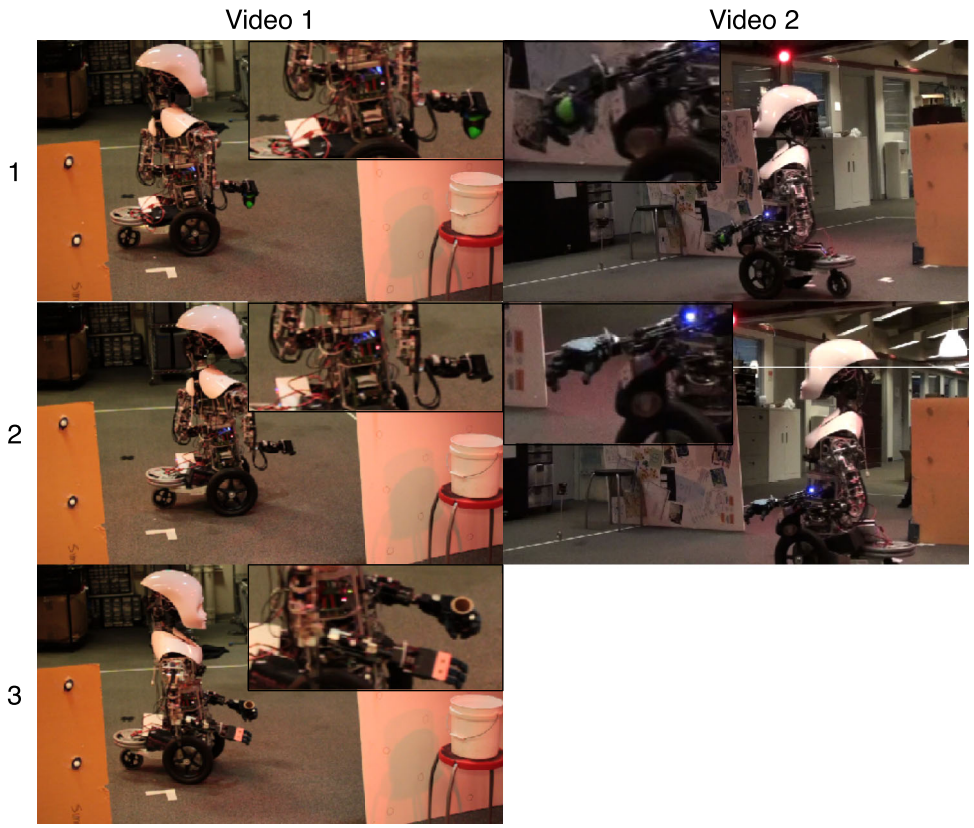

Fig. 3 Photos taken during the robotic performance of the demonstration scenario. The *left column* shows a video still from the perspective of the human opponent during the three conditions (Fig. 2). The *right column* shows the same scene from a different angle, revealing what the robot is hiding from the human. In all *rows* the robot must transport the cylinder to the goal. In *row 1*, the robot's mental state goals are to reveal that it carries the football, and hide that it carries the cylinder. It accomplishes this by carrying the football in the *right hand*, and carrying the cylinder in the *left hand* which it hides behind its body. In *row 2*, the robot's mental state goal is to hide that it carries the cylinder, but with no goal about carrying a decoy object. It does this by carrying the cylinder in its *left hand*, again hidden behind the body. In *row 3*, the robot has no mental state goals. In this case it simply carries the cylinder in its *left hand* openly

These previous implementations and demonstrations focused on modeling human mental states by monitoring the human's physical actions and visual perspective. The robot then re-uses parts of its own behavioral mechanisms in three main ways: (1) reusing its own world modeling capabilities to connect the human's visual perspective to possible human mental state formation; (2) reusing its own action performance mechanisms to connect the human's observed physical motions to possible higher level actions; (3) reusing its own goal directed action system to infer goals based on inferred mental states and actions. Inspired by work in human psychology, the self-as-simulator architecture provides the advantage of a common vocabulary between the robot's own behavioral mechanisms and the properties inferred in an observed human; since the purpose of mental state inference is to inform the actions of the robot, it is critical that inferred mental states be mapped into the space of its behavior generation systems.

Using these systems, the robot is constantly modeling the mental states of nearby agents. Whenever the robot discovers a new agent, along with updating its own model of the world state to reflect the presence of this agent, it also spawns a new copy of its own modeling systems. This new copy will maintain a world state model from the perspective of the new agent. These copies are provided with sensory data that is *re-imagined* by the robot from its own world state model, then transformed and filtered to best match what that agent should be experiencing.

Since these copies have the same capabilities as the robot's own systems, they too spawn copies when they sense another agent (including the robot), allowing for recursive mental state modeling. We currently cap this recursion at two levels, to allow for second level mental state goals such as *Robot Demonstrates To Human That Robot Knows X.*

In this previous work, the robot takes advantage of its own embodiment by using its behavior generation mechanisms as a common language between its behavior and the human's. Human actions and mental states are understood through their relation to the robot's own actions and mental states. This common language allows the robot to leverage its own structures for better inferences (e.g., inferring a goal from a physical action). However, the robot was passive observer of mental states, watching the human as an isolated actor and taking independent action only once it had completed a particular inference.

This implementation describes improvements made to the above techniques allowing the robot to move out of the role of an observer and instead become an active participant in theory of mind activities. Firstly, the robot's presence in an interaction with a human cannot be ignored. The robot must take into account the effect its own presence and actions are having on the human's mental states (including recursive mental states, i.e., those that the human has about the robot's own mental states). Second, it is important to move beyond the present. While monitoring the mental states of a human "right now" is useful, it is also critical to make short term predictions. Finally, the robot needs to be able to take action to modify the future mental states of the human. This can be thought of as a very low level type of communication—deciding how the robot should perform to cause the human to form a desired mental state.

The goal of this implementation is to add mental state manipulation capabilities to a physical robot. Robotic actions cannot be modeled solely by a set of pre- and post-conditions: they take time to perform, occur across physical space, and are observed subject to the perspective of an observer. We embrace these realities and include motion and geometry in our computations for mental state manipulation. For example, the robot may need to turn a certain way before performing an action in order for a critical part of a motion or effector to be visible (or hidden) from an observer.

The implementation sections are somewhat abstract, describing how we re-use elements without fully defining the underlying systems. More details about the underlying systems are available in the works referenced above, but to give better context to the following sections:

Mental States used by the robots described here consist of knowledge about the world around them; this takes the form of a collection of known objects and the properties of those objects. For example, the robot may have seen an orange football at location *(x,y,z)*. For animate objects, such as humans, aside from basic properties the robot is also modeling their mental states as described above. In addition to these external objects, the robot also knows its own location and body configuration.

Actions performed by the robot take the form of physical motions with associated goals and expectations. Some may be an open loop motion, such as "wave," however many take one or more parameters (in the form of known objects) and may have dynamic elements. For example, "grab" operates on an object with a known location, includes an expectation (object ends up in hand), and has a dynamic controller instead of a fixed motion sequence (closed loop feedback gets the hand to the surround the target object).

Actions can be selected based on the current set of mental states, as well as higher level goals (not discussed here). For example, "grab" is only possible if a graspable object is nearby.

The implementation is broken up into two main challenges: how can the robot simulate future actions along with the mental state results of those actions, and, given this ability, how can it choose a course of action to bring about its mental state goals.

Algorithm 1 Implementation Outline

```
Find Action Sequence:
  Clear List of Failed Action Sequences
  while Viable Sequences Remain do
    Init Future Simulation From Current State
    time = 0
    Begin Simulation
    while Simulation Running AND time < MAX_TIME do
      time++
      if Action In Progress then
        Keep Performing Action
      else
        Select and Begin relevant Unexplored Action
      if Mental State Goals Succeeded then
        return Recent-Action-Sequence
      else if Mental State Goals Failed then
        Save Recent Action Sequence to Failed List
        End Simulation
  return Failed-To-Find-Sequence
```

4.1 Simulating the Future

In the previous section we described how we have used mental state inference to model the current values of the human's hidden mental states to help resolve ambiguities and better assist the human with their goals. In order to proactively manipulate mental states, we use these mechanisms within the context of a simulator which simulates the immediate future, allowing the robot to evaluate and choose between multiple actions based on mental state outcomes (see basic outline in Algorithm 1).

The robot's simulation of possible futures consists of a copy of its own behavioral mechanisms, identical except that it is disconnected from the real world inputs (sensors) and outputs (motors). This "hypothetical" robot includes a copy of the virtual model used for motor planning, so it still has access to a body for performing its motor actions, however the final stage of synchronizing that model to the motors of the physical robot is not performed. This allows the robot to maintain a detailed representation of the hypothetical actions being performed, down to specific positioning of parts of its own body.

By copying the mechanisms of the main robot's systems to make this "hypothetical" robot, the hypothetical robot inherits the same mental state modeling capabilities as the original. However, using these capabilities to make future predictions creates challenges not faced by the real-time robot. Though the hypothetical robot can simulate motions with the virtual body provided to it, it cannot rely on real-world physics to close the sensory-motor loop and change the world state as a result of its motions. Additionally, in order to predict future mental states, rather than current, it must have a way to advance its modeling to look ahead in time.

4.1.1 Mental States as Simple Physics

The real robot performs real actions, which alter the state of the world, which changes the sensory input it receives

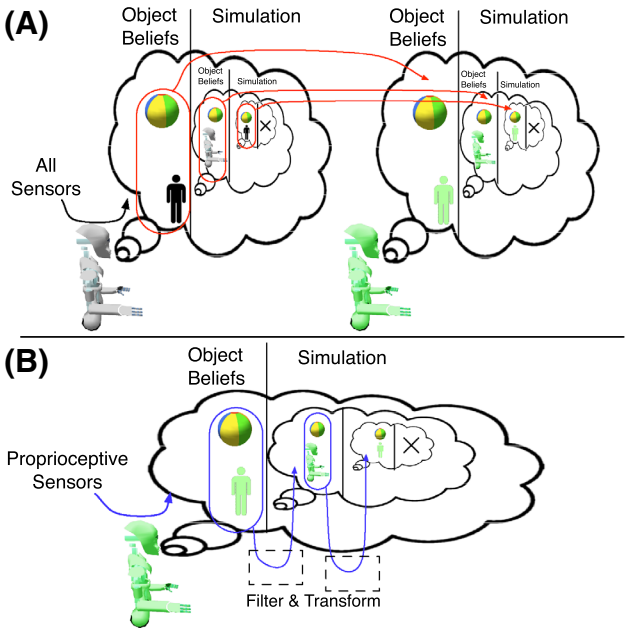

Fig. 4 Dataflow during reset and operation of hypothetical robot. **a** shows the real robot on the *left*, and hypothetical robot on the *right*. The *red arrows* represent object beliefs being copied during the reset of the hypothetical robot at the beginning of a future simulation (the hypothetical robot, including its models of human mental states, should start from realtime robot's current estimate). **b** shows the operation of the hypothetical robot during a future simulation. *Blue arrows* show how the data about objects propagates in this configuration (except for the cutoff from the sensors, this is identical propagation as in the real robot). (Color figure online)

from the world, which in turn results in updates to its mental states (and the mental states of the agents it is modeling). Our hypothetical robot cannot rely on this sensory-motor loop, since it is not interacting with the physical world. To overcome this absence, we reuse mechanisms designed for robust world modeling in the face of sensory lapses and noise.

The first such mechanism is mental state maintenance. For many reasons, such as occlusion, distance, or viewing angle, the sensory stream of data about a particular object may be interrupted. In these cases, as long as no conflicting data is received, hidden objects are assumed to remain as they were last seen. The real-time robot is constantly maintaining this information about the world, and also updating the models it is maintaining about nearby humans, which in turn maintain this information in the same way. The hypothetical robot, cut off from the sensors, will have no new data coming in; however, if we perform a full copy of the models the realtime robot has created, the hypothetical robot will start with an accurate model, and the belief maintenance mechanisms will retain the initial data over time. This copy must be performed recursively, copying the mental states over for each of the agents the robot is currently modeling the mental states for, so they too start from the correct initial state (Fig. 4).

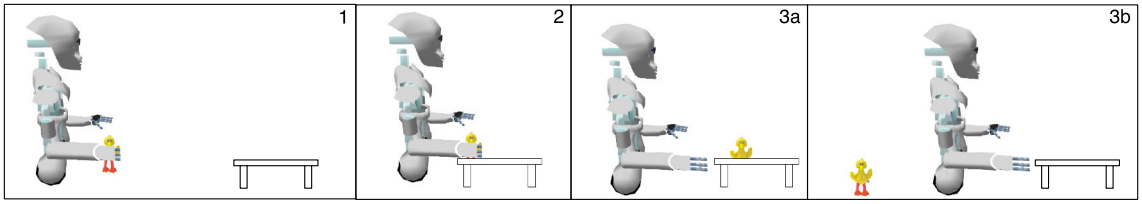

Fig. 5 When the physical robot successfully carries an object, the result is frame *3a*, and the robot can model this result because its sensors will detect the toy on the table. When the hypothetical robot carries a virtual object, the physical sensor feed will not be affected. In both the case of the physical and hypothetical robots, the belief maintenance mechanisms assume that an object, while grasped, stays in the hand, and thus they update the toy's location during the carry action. This allows both real and hypothetical robots to experience frame 2, even though the hypothetical robot is not moving a real object, and the physical robot is unlikely to be able to track the toy visually during this process. It is important to cut off the physical sensor feed from reaching the hypothetical robot so the object belief of the toy will stay put once released (frame *3a*) and not snap back to its real-world location (frame *3b*)

The next mechanism is an expectation mechanism which operates in conjunction with physical actions. Many physical actions modify properties of objects in the world. For example, when carrying an object, it is expected that the object move along with the motion of the robot's hand. The expectation mechanism updates the robot's model of the world state to account for these expected property changes. During an action like "carry", it is difficult for the robot to actually see the object in its hand, but by using the expectation mechanism the robot can continue to update its world model until sensory data is re-acquired (conflicting data can override this expectation). This helps the real-time robot keep a more accurate model of the world state in the face of incomplete sensory information by updating its models based on the expected results of its actions (See Fig. 5). For the hypothetical robot, which has no access to actual sensory information, this process serves the critical role of allowing it to update its world model to reflect expected outcomes as actions are performed. The hypothetical robot's world model is what it will use to re-imagine the visual input provided to the perceptual systems of the humans it is modeling, therefore these changes will also update the inferred mental states of the humans it is simulating (if it judges the change to be within their visual perspective).

The two mechanisms above allow the hypothetical robot to maintain and update a simple world state model without resorting to a heavy-weight physics simulation. Instead, the representation is entirely within its own mental states, no additional simulated sensors or physics modeling is required. This does have the limitation that all simulated actions always occur as expected (producing the expected results).

The proprioceptive sensing of the hypothetical robot's kinematics and locomotion can function almost as normal as they are tied to the virtual model. This means that as it performs actions, it motions and moves around the hypothetical world appropriately, and those motions can be constantly fed not only into its own world model, but can be used to calculate accurate occlusions and sight-lines while updating

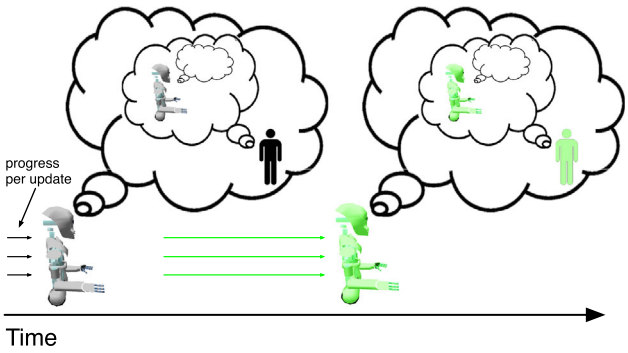

Fig. 6 A *hypothetical* copy of the robot is used for mental state predictions. The robot has the capability to model mental states of agents around it (*left*). To make short term predictions, a copy of the robot (*green*, *right*) starts from the robot's current state and performs (in a virtual space) the actions the robot is about to perform, but performs them much faster. While doing this, it maintains the mental states of the surrounding agents as they participate in this accelerated timeline. This gives the robot the ability to predict the mental states of surrounding agents in the short term future. (Color figure online)

the mental models of the humans. Thus the modeling of the human's mental states can take into account a detailed, 3D representation of future situations and how that layout will affect their perception of the events as they unfold.

4.1.2 Time

As described above, the hypothetical robot is as close as possible to a direct copy of the mechanisms that run the actual robot. The hypothetical robot, however, is not limited by the constraints placed on the physical robot and its motors; we can thus send it forward in time by running it much faster than the physical robot (in simulation, Fig. 6).

To allow this, the hypothetical robot's progress through motor actions is increased: joints move faster to complete motor actions more quickly. In addition, instead of updating the robot's behavior, motor, and perceptual systems at the constant rate of 30 hz, as in the real robot, the hypothetical

robot is allowed to update as fast as the CPU allows with a virtual clock keeping pace such that 1/30th of a second appears to have elapsed between each update.

This allows the robot to send a probe into the near term future, by copying its current state into the hypothetical robot, having the robot quickly perform the next actions planned, in simulation, and then examining the predicted mental states produced in that simulation after an action or series of actions.

4.2 Finding Correct Action Sequence

In the last sections we described how to use the self as simulator system to model another agent's mental state as well as to simulate hypothetical futures. In this section we use these two capabilities together to search for an action sequence that achieves our particular mental state modification goals. This section details the critical ingredients for this search. First, mental state goals must be well defined—it is not sufficient to pair a mental state with an agent, the path through the recursive modeling tree is important to the meaning of the goal. Second, taking this goal into account, the robot must search through its space of possible action and parameter sequences, determining action relevance as it proceeds.

4.2.1 Mental State Goals

Mental state models exist in a recursive hierarchy, with each agent modeling the agents around them, and those models in turn modeling the agents known to that model. This process allows us to specify complicated mental state goals (Fig. 7). We traverse this structure using *Agent Specifiers*, which are a mechanism to specify a particular model, or models, in the recursive model graph. For example, we might want all humans to think that the robot knows *X*. This specifier would then create several paths through the graph to pinpoint the appropriate models, and when paired with a particular mental state goal (*X*), together they specify the overall desired goal state.

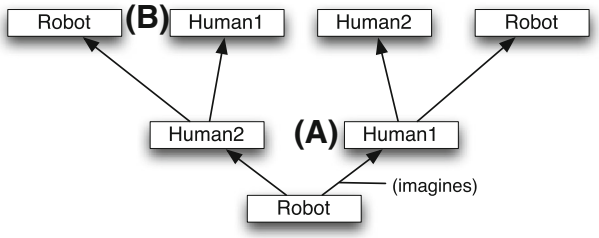

Fig. 7 A two level deep example of the recursive structure of the mental models maintained by the robot. The robot is maintaining a model of the mental states of two humans, and each of those mental models, in turn, is maintaining a model of the other agents. Mental state goals, then, must not just indicate a desired mental state and an agent which should have that state, but also a path to that agent. It is different to try to get *Human1* to believe *X* (model *(A)*) than to get *Human2* to believe that *Human1* believes *X* (model *(B)*)

In Fig. 8, arrows show the robot tracking these goals during a simulation. Arrows visually show the path through the agent models in the mental state graph to a particular model's belief, in this case going from human to robot to object for the goal "human knows that robot knows it is carrying *X*" (first arrow, originating at the root node "Robot" is always omitted).

4.2.2 Action Sequences

Having specified mental state goals, and a hypothetical robot which looks forward while tracking mental state effects caused by its actions, we can now search through the action space for sequences that achieve the desired result. Actions are often parameterized, and each action has a mechanism to determine current valid parameters, as well as whether the action can even be performed in the current situation. For example, a *Grab* action will be able to produce a list of target objects, which are nearby objects that can be grabbed; it can also report that the action is inappropriate, in this case if the robot's hands are full, or no objects are in range.

Because the set of appropriate actions and parameters change as the robot acts and alters the world, it does not build an exhaustive tree initially. Instead, the tree is filled out as it searches (Fig. 9). Through this process, the robot can find the path though its parameter and action space that most achieves its mental state goals. Once a successful sequence is found, the search is terminated. The parameters associated with the sequence (e.g., which object to grab) are composed of mental states held by the hypothetical robot, so to be performed by the real robot they must be mapped back to the mental states of the real robot, which may be different (object properties may change during the simulation, for example). We have found that simple heuristics suffice for these mappings, such as relying on similarity of key object properties like location and identifying information.

5 Study

In order to evaluate reactions of people toward a robot teammate with this mental state manipulation ability, a video based human-subjects study was performed. Along with testing if the robot's manipulative actions provided any advantage to the robot in the game, the study measured if these behaviors had any effect on the subjects' perception of the robot's competencies and their evaluation of the robot as a potential partner.

Subjects participated in the study online, by accessing a website. The subjects were broken into three different groups. All subjects were instructed that they would be playing a simulated game with the robot. After the game and rules were

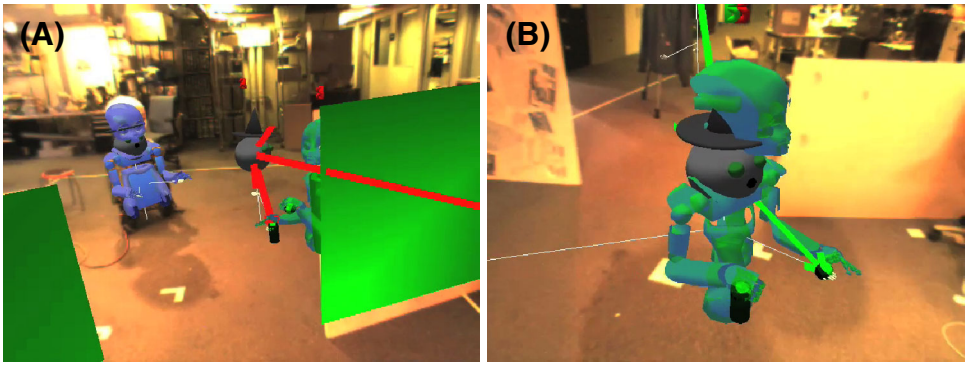

Fig. 8 This augmented reality visualizer demonstrates the robot's planning system. The hypothetical robot (*green*) simulates action sequences based on the current goals and the most recent sensory data. **a** Human player's perspective of a failed trial: the hypothetical robot has just revealed that it is carrying the cylinder, failing a mental state goal (visualized by the *red arrow*). **b** Opposite perspective of another action sequence: the robot has achieved a mental state goal (show human that it is carrying the football, *green arrow*) and not yet failed its goal of keeping the cylinder hidden (*no red arrow*)—a possible successful action sequence in progress. (Color figure online)

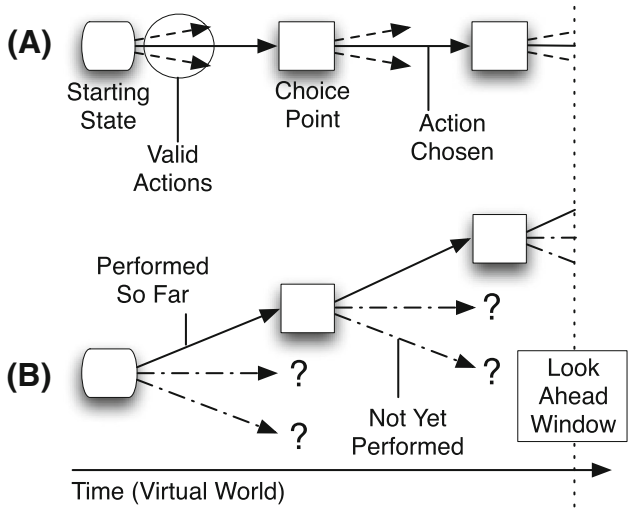

Fig. 9 **a** diagrams the process of simulating a single possible future (Sect. 4.1). **b** shows of search through action/parameter space, with lazy discovery of possible subsequent actions (to account for each action altering the world state, and thus changing which actions and parameters are available). Robot maintains mental state models as it searches so as to monitor mental state goals

described, they were shown a video of the robot performing its turn. This video was recorded from the perspective of the human player, with the robot programmed to treat the camera as if it were the opposing player (thus any actions which would hide an object from the competing human would hide that object from the camera).

Each of the three groups corresponded to one of the conditions in Fig. 2 and saw videos of the robot motivated by the goal in that condition. After watching this video, the subjects were instructed to fill out their answers to several questions. The first question asked them to indicate which item they would place in their goal area in response to the robot's actions. Next they were asked if, in future games, they would prefer to team with the robot or play against the robot. Finally a set of questions asked them to rank the robot on several criteria.

In the conditions with attempted concealment (conditions one and two), after filling out the information above the subjects were shown the same interaction from a second video angle allowing them to see any originally occluded objects. After seeing this second video, the subjects are then asked the same questions again to evaluate how their answers change in response to this new information.

5.1 Study Results and Discussion

The answers provided by the subjects were analyzed to address the following hypotheses:

- **Hypothesis 1** The mental state manipulation is successful, as measured by the subjects' choice of object. If the robot is successful, people will be fooled by the robot's decoy object in condition one, they will be unsure what to play in condition two, and they will correctly see the robot's actions in condition three and thus be able to win. After seeing the second video, revealing the robot's hidden hand, people will choose the object the robot was hiding.
- **Hypothesis 2** Subjects will choose the robot as a teammate more frequently when they observe its mental state manipulation capabilities. People will be more willing to team with the robot that hides objects behind its back than the robot that openly carries objects, and will change their mind about teaming with the condition one robot once they realize it had been manipulating mental states.
- **Hypothesis 3** People are more willing to attribute mental states to the robot once they see that it is pursuing a strategy of mental state manipulation, rather than sim-

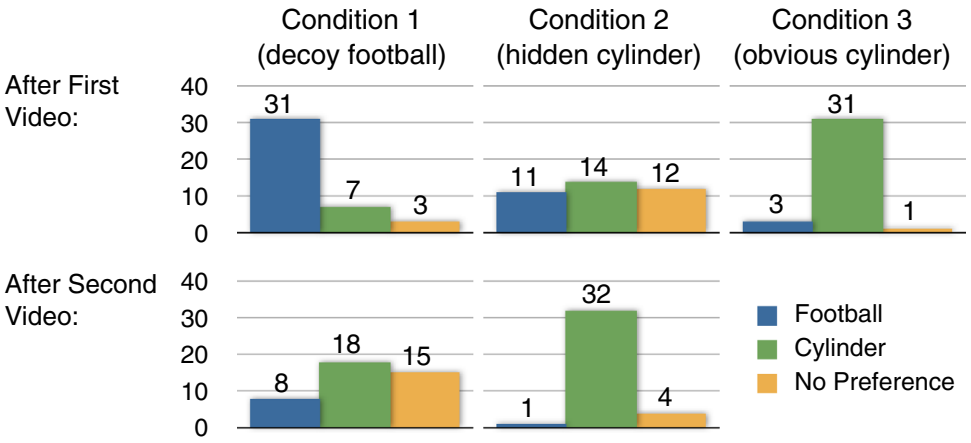

Fig. 10 Data showing object choice by human players across each condition, before and after having seen the second video. The participant is instructed to choose the winning object, which is defined to be the same object the robot placed in its goal. Subject choice differs significantly by condition (across *first row*) ($p < 0.01$) and changes significantly after seeing second video (*columns*) ($p < 0.01$)

ply transporting an object to the goal. This hypothesis is evaluated by the subjects' change in rating of several statements after the robot's deception is revealed.

Across the three conditions, 113 subjects completed the entire questionnaire. 41 subjects were in the condition one group, 37 in condition two, and 35 in condition three.

5.1.1 Hypothesis 1: Success of Mental State Manipulation

Participants' choices of object to play indicated that the robot successfully occluded its chosen object as described in hypothesis one (Fig. 10). In condition two (no object visible) the participants showed no strong preference for either object; in the other conditions the participants chose the same object as the robot was openly carrying: the football in condition one (the deception is successful, and the human loses) and the cylinder in condition three (the human is correct, and wins).

In condition one and two, many subjects change their choice of object after seeing the second video (revealing both of the robot's hands). In condition two this change happens as expected; after the first video the subjects have little preference, but then after seeing the second video they switch their answer to the newly revealed cylinder.

In condition one, when the deception is revealed many participants switch from their initial choice of football to the now revealed cylinder. While technically the robot could play either item (it has both in its hands), cylinder is chosen most frequently as expected by the hypothesis. This choice is consistent with applying a deceptive motive to the robot: it was hiding the cylinder on purpose, and therefore means to play it. In written responses, 11 of the 18 who chose the cylinder (the choice predicted by hypothesis one) used language that had some relation to mental state manipulation—that they chose the cylinder because the robot was "hiding" it from the subject.

Despite cylinder being the most frequent choice, many participants were still undecided or chose the football. An informal analysis of the written comments sheds some light on the wider distribution of choices in condition 1 part 2 as compared to condition 2 part 2. Six of the no-preference respondents give a mechanistic description of the robot's behavior, without mentioning any motive (without attributing a motive for hiding the object, it is not clear which of the two objects the robot would play). Three of them reacted in the opposite direction—having seen that the robot tricked them, they attribute such high levels of deceptive capability that they were unsure which object to now choose, imagining further trickery. The remaining six indicate a level of confusion with the task, or seem to have missed some parts of the video (e.g., some did not seem to see the cylinder in the robot's previously hidden hand when shown the second video).

5.1.2 Hypothesis 2: Willingness of Subjects to Team with Robot

After each video, subjects were asked whether in future games they would prefer to have the robot on their team or on the opposing team (Fig. 11).

Hypothesis two predicts that subjects will be more willing to team with a robot that is able to perform mental state manipulations. From the analysis of hypothesis one, we know that subjects largely were fooled by the robot's deception in condition one, choosing the football. Consistent with this expectation, after watching only the first video, subjects in conditions one and three were less likely to want to team with the robot as compared with condition two, where they witness the robot hiding an object. Additionally, the subjects are more willing to team with the robot in condition one after the second video reveals the robot's manipulation. These differences indicate that when people are aware of mental state manipulation capabilities, they are more willing to team with the robot.

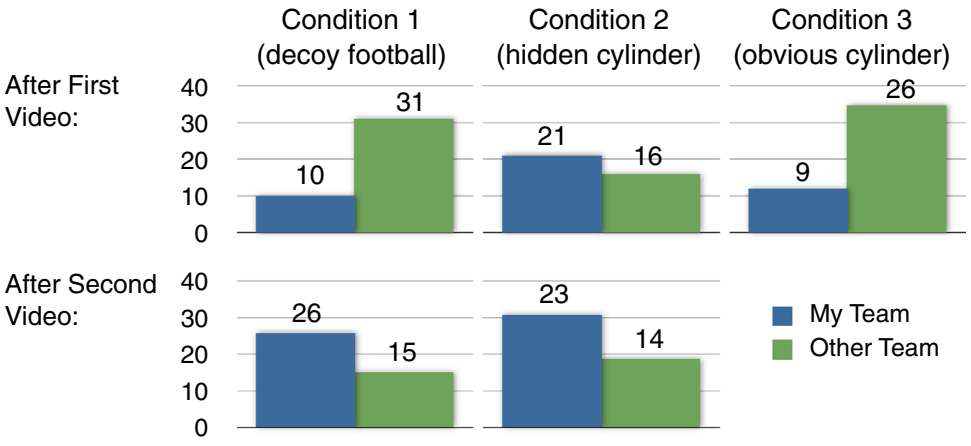

Fig. 11 Human is asked whether, if they were to play another game, they would choose to have the robot on their team or the other team. After the first video, participants want to team with the robot significantly more in condition two than in one or three ($p < 0.01$). In condition one subjects change their answer in favor of teaming with the robot after the second video ($p < 0.001$)

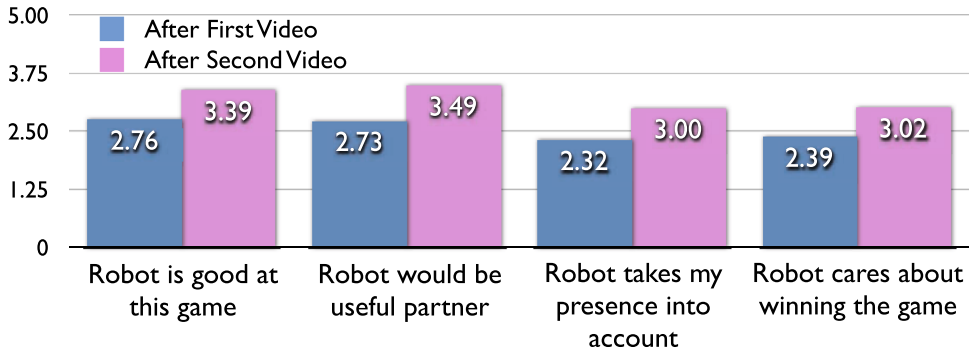

Fig. 12 Data showing subjects' rating of the robot on four questions (using a five point scale) in condition one. Subjects in condition one were asked these questions once after watching the first video of the robot's turn. They are asked to rate the robot again after the robot's hidden behavior is revealed through the second video. For each question, the responses change significantly after watching the second video ($p < 0.01$)

In condition two, it is expected that the subjects were largely aware of the robot's strategy of occluding the item. The second video reveals the item, allowing the participant to refine their item choice, but we don't expect their desire to team with the robot to change, as no new capabilities are revealed. The teaming question responses are consistent with this interpretation—there is very little change in their response to the teaming question after watching the second video.

5.1.3 Hypothesis 3: Attribution of Mental States to Robot

In addition to the above questions, subjects were asked to rate their agreement with four statements about the robot's performance and internal mental functions on a five point scale. By asking these questions before and after the deception is revealed in condition one, we can examine how that revelation changes the participants' evaluation of the robot and to what extent it affects their attribution of mental states. Figure 12 shows how the subjects' opinions changed in support of hypothesis three. Subjects with lowered expectations have been shown to be more affected by the positive performances of a robot in some cases, which may amplify this effect [15].

5.1.4 Results Summary

Through the subjects' object choices in the three conditions, the study showed that the mental state manipulation performed by the robot was successful. The mental state manipulation goals the robot pursued did indeed change the behavior of the subjects.

The study also showed that these behaviors were readable to the subjects. After watching the manipulation behavior from a second angle, subjects were able to better predict the robot's actions based on a correct understanding of its deceptive motivation for hiding its actions.

Finally, these capabilities had a positive effect on subjects' willingness to work with the robot, and raised their rating of the robot's capabilities. Subjects' discovery of the mental state manipulation changed both their mechanistic description of the robot's behavior, as well as their description of its behavior in terms of intentions.

This multi-perspective video strategy allowed us to efficiently collect data from many participants, while preserving certain important aspects of the game experience. Other or stronger effects might be observed when subjects can physically interact with the robot, however we were able to obtain significant results on the effect of mental state manipulation capabilities on teaming preference and mental state attribution using this method.

6 Discussion and Future Work

The focus of this work has been to leverage how embodiment connects the observable and alterable world with the hidden mental states of other agents which cannot be directly observed or operated on. Humans and robots, while vastly different, share a common problem of being embodied agents with sensory motor loops based on affecting and observing the physical world around them. By modeling a human's connection between mental states and the world as similar to its own, and reusing those mechanisms to help evaluate mental state consequences, the robot can add altering mental states in others to its repertoire of possible goals.

This gives the robot a primitive type of communication that operates without language, using only the robot's own actions and perceptions as its vocabulary.

The demonstration shown here focuses on a competitive scenario, and the mental state manipulation performed could be classified as "deceptive." It is not the specific intention of the authors to create deceptive robots; rather, it is to explore the base level skill of taking action to modify mental states. It is interesting to note that this same skill could be described as part of "communication" or "maintenance of common ground" when goals are aligned, but becomes part of "deception" when they are not. In the scenario presented, the robot's goal is to modify the human's mental states to differ from the actual world; this scenario was chosen because it allows a simple demonstration to exercise the full capabilities of the system.

An equivalently interesting demonstration of the system in a cooperative scenario could be to have the robot modify a human's incorrect mental state back to ground truth, for example in a task where a human is overloaded and forms incorrect mental states as a result. In order to show off the robustness of the belief manipulation, this task should be chosen such that basic heuristics ("speak information near partner", "point to object partner hasn't seen") would be insufficient. While this could be done (e.g. by causing interference, distractions, etc.), it was deemed more difficult to set up a suitably complex task repeatably in the laboratory.

One limitation to this self as simulator approach is that the robot cannot model mental states that are not part of its own mental repertoire. While this limits the possible space of observations and manipulations, this limitation also means that all observations that the robot can make are in the vocabulary of its own mental states, readily applicable to its own behavior; observing or inferring mental states outside this vocabulary would not be directly useful by the robot's systems, as they would not have any meaning to the robot: meaning is derived by the role of that mental state in the robot's own behavioral mechanisms.

Due to the detailed nature of the mental state modeling and simulations of the future, it would be computationally expensive to create long term plans with these mechanisms. However, a long term plan at this level of detail is not necessarily productive—likely it is not worth considering the exact hand motion I'll need for a very specific situation occurring tomorrow. Instead, this level of detail is useful in the very short term, for determining how to perform the next actions appropriately. Interesting future work is to integrate these techniques with a longer term, more abstract mechanism, allowing longer plans with mixed levels of detail. It could also be helpful to include a set of heuristics that define several simple communication strategies and the context to which they apply, to allow for simple communication in simple situations ("speak information near partner"). When one of these fails or is not applicable, then the computationally intensive modeling/planning described here could be used to determine a course of action.

The major contributions of this paper are: an implementation which proactively manipulates human mental states at the level of perception and physical action and an evaluation of how this ability is perceived by humans.

References

1. Blumberg B, Downie M, Ivanov Y, Berlin M, Johnson MP, Tomlinson B (2002) Integrated learning for interactive synthetic characters. ACM Transactions on Graphics, vol. 21, no. 3. In: Proceedings of ACM SIGGRAPH 2002
2. Breazeal C, Buchsbaum D, Gray J, Blumberg B (2005) Learning from and about others: toward using imitation to bootstrap the social competence of robots. Artif Life 11:31–62
3. Breazeal C, Gray J, Berlin M (2009) An embodied cognition approach to mindreading skills for socially intelligent robots. Int J Robot Res (IJHR-09) 28(5):656
4. Burke R, Isla D, Downie M, Ivanov Y, Blumberg B (2001) CreatureSmarts: the art and architecture of a virtual brain. In: Proceedings of the game developers conference, San Jose, pp 147–166
5. Cassimatis NL, Murugesan A, Bignoli PG (2009) Reasoning as simulation. Cognit Process 10(4):343–353
6. Ettinger D, Jehiel P (2010) A theory of deception. Am Econ J 2(1):1–20
7. Goodman ND, Stuhlmueller A (2013) Knowledge and implicature: modeling language understanding as social cognition. Top Cognit Sci 5:173–184
8. Gray J, Berlin M, Cynthia B (2007) Intention recognition with divergent beliefs for collaborative robots. In: Society for the study of artificial intelligence and simulation of behaviour (AISB-07)

9. Gray J, Breazeal C, Berlin M, Brooks A, Lieberman J (2005) Action parsing and goal inference using self as simulator. In: 14th IEEE international workshop on robot and human interactive communication (ROMAN), Nashville, TN
10. Gray J, Hoffman G, Adalgeirsson SO, Berlin M, Breazeal C (2010) Expressive, interactive robots: tools, techniques, and insights based on collaborations. In: HRI 2010 workshop: what do collaborations with the arts have to say about HRI?
11. Johnson M, Demiris Y (2005) Perceptual perspective taking and action recognition. Int J Adv Robot Syst 2(4):301–308
12. Kelley R, Tavakkoli A, King C, Nicolescu M, Nicolescu M, Bebis G (2008) Understanding human intentions via hidden markov models in autonomous mobile robots. In: Proceedings of the 3rd international conference on human robot interaction, pp 367–374
13. Kennedy W, Bugajska M, Harrison A, Trafton J (2009) "Like-me" simulation as an effective and cognitively plausible basis for social robotics. Int J Soc Robot 1(2):181–194
14. Kennedy WG, Bugajska MD, Marge M, Adams W, Fransen BR, Perzanowski D, Schultz AC, Trafton JG (2007) Spatial representation and reasoning for human–robot collaboration, vol 7. In: AAAI, pp 1554–1559
15. Komatsu T, Kurosawa R, Yamada S (2012) How does the difference between users' expectations and perceptions about a robotic agent affect their behavior? Int J Soc Robot 4(2):109–116
16. Laird JE (2001) It knows what you're going to do: adding anticipation to a quakebot. In: AGENTS '01: proceedings of the fifth international conference on autonomous agents. ACM Press, New York, pp 385–392
17. Marsella SC, and Pynadath DV (2005) Modeling influence and theory of mind. In: Artificial intelligence and the simulation of behavior
18. Meltzoff AN (1995) Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. Dev Psychol 31:838–850
19. Pandey AK, Ali M, Alami R (2013) Towards a task-aware proactive sociable robot based on multi-state perspective-taking. Int J Soc Robot 5(2):215–236
20. Rizzolatti G, Fadiga L, Gallese V, Fogassi L (1996) Premotor cortex and the recognition of motor actions. Cognit Brain Res 3:131–141
21. Short E, Hart J, Vu M, Scassellati B (2010) No fair!! an interaction with a cheating robot. In: Human–robot interaction (HRI), 2010 5th ACM/IEEE international conference on, pp 219–226
22. Trafton G, Hiatt L, Harrison A, Tamborello F, Khemlani S, Schultz A (2013) Act-r/e: an embodied cognitive architecture for human–robot interaction. J Hum Robot Interact 2(1):30–55
23. Trafton J, Schultz A, Perznowski D, Bugajska M, Adams W, Cassimatis N, Brock D (2006) Children and robots learning to play hide and seek. In: Proceedings of the 1st ACM SIGCHI/SIGART conference on human–robot interaction. ACM, Salt Lake City, pp 242–249
24. Trafton JG, Cassimatis NL, Bugajska MD, Brock DP, Mintz FE, Schultz AC (2005) Enabling effective human-robot interaction using perspective-taking in robots. IEEE Trans Syst Man Cybern 35(4):460–470
25. Wagner AR, Arkin RC (2011) Acting deceptively: providing robots with the capacity for deception. Int J Soc Robot 3(1):5–26
26. Wellman H, Cross D, Watson J (2001) Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev 72(3):655–684
27. Wimmer H, Perner J (1983) Beliefs about beliefs: representation and constraining function on wrong beliefs in young children's understanding of deception. Cognition 13:103–128

Jesse Gray is a cofounder of IF Robots LLC, a robotics consultancy in Cambridge, MA. After receiving a BA from Tufts University, he received his Ph.D. and S.M. from the MIT Media Lab's Personal Robots Group directed by Dr. Cynthia Breazeal. At MIT Jesse worked on developing autonomous behavior and expressive motion for the robots Leonardo and Nexi. His research focused on self-as-simulator techniques to enable embodied agents to make sense of others' mental states.

Cynthia Breazeal is an associate professor at the MIT Media Lab where she founded and directs the Personal Robots Group. She is a pioneer of social robotics and human robot interaction. She is author of Designing Sociable Robots and has published more than 100 peer-reviewed articles on these topics. Her research program develops personal robots with interpersonal skills that enable them to work and learn collaboratively with people. Recent work focuses on socially assistive robots targeting applications in education and health that require long-term interaction. She received her Sc.D. in electrical engineering and computer science from the Massachusetts Institute of Technology in 2000.