International Journal of Humanoid Robotics
Vol. 1, No. 2 (2004) 1–34

World Scientific
www.worldscientific.com

HUMANOID ROBOTS AS COOPERATIVE PARTNERS FOR PEOPLE

CYNTHIA BREAZEAL*, ANDREW BROOKS, JESSE GRAY, GUY HOFFMAN,
CORY KIDD, HANS LEE, JEFF LIEBERMAN, ANDREA LOCKERD
and DAVID MULANDA

MIT Media Lab, Robotic Life Group
77 Massachusetts Ave E15-468, Cambridge, MA 02139, USA
**cynthiab@media.mit.edu*

Received
Accepted

This paper presents an overview of our work towards building socially intelligent, cooperative humanoid robots that can work and learn in partnership with people. People understand each other in social terms, allowing us to engage others in a variety of complex social interactions including communication, social learning, and cooperation. We present our theoretical framework that is a novel combination of Joint Intention Theory and Situated Learning Theory and demonstrate how it can be applied to develop our sociable humanoid robot, Leonardo. We demonstrate the robot's ability to learn quickly and effectively from natural human instruction using gesture and dialog, and then cooperate to perform a learned task jointly with a person. Such issues must be addressed to enable many new and exciting applications for robots that require them to play a long-term role in people's daily lives.

Keywords:

1. Introduction

Many of the earliest motivations for developing humanoids centered on creating robots that can play a role in the daily lives of people. Today, humanoid robots are being developed to provide the elderly with assistance in their homes and to support medical care in hospitals. In other applications, humanoids are being developed to serve as members of human-robot teams. One such example is NASA JSC's Robonaut,[1] a robot envisioned to serve as an astronaut's assistant in space station maintenance operations. In the future, we expect to see more applications for robots that share our environment and tools and participate in joint activities with untrained humans. Robots with a human-like morphology would not require us to re-engineer our environment and tools to accommodate them. Beyond form factor, however, there are critical social issues that concern how robots should interact with us.

1.1. *Working and learning in collaboration*

Rather than viewing robots as semi-autonomous tools that are directed via human supervision, we envision robots that can cooperate with humans as capable partners. For instance, robots should be able to interact with people to learn new tasks and skills in a way that is intuitive and natural for a human teacher. Furthermore, once the robot has learned a new task or skill, the robot should be able to perform it either independently or in partnership with another person, providing him or her with well-timed, relevant assistance.

We argue that robots that can learn from and work with people must exhibit social appropriateness and adeptness. Ideally, teaching a robot a new capability should be as easy and fast as teaching a person. Human learning, however, is quite different from what machine learning is today. Typically, statistical approaches (e.g. supervised or reinforcement learning) require a large number of examples in order to discover the underlying structure of the learning problem. In contrast, human tutelage and other forms of social learning leverage strongly from structure provided through interpersonal interaction. This social guidance enables a learner to acquire new concepts and skills from few examples, and to refine these over time. We propose an approach that models learning as a fundamentally collaborative process that takes place within a tightly coupled interaction between teacher and learner. We suggest learning as *tutelage* rather than through *supervision.*

Likewise, robots need social skills to collaborate with people as teammates. They have to understand our intentions, beliefs, desires, and goals so that they can perform the right actions at the appropriate time. For the human-robot team to succeed, they must also communicate their own set of intents and goals to establish and maintain a set of shared beliefs and to coordinate their actions to execute the shared plan.[2] In addition, each teammate must demonstrate commitment to doing their own part, commitment to the other in doing theirs, and commitment to the success of the overall task.[3,4] We thus propose a view of human-robot *collaboration* instead of mere *interaction.*

Generally speaking, robots today treat us as other objects in the environment (e.g. obstacles to be navigated around), or at best interact with us in a manner characteristic of socially impaired people. They generally do not understand or interact with people as people. They are not aware of our goals and intentions. As a result, they do not know how to appropriately adjust their behavior to help us as our goals and needs change. They often do not display their own goals and intentions in ways easily accessible to us. They do not flexibly draw their attention to what we currently find of interest so that their behavior can be coordinated with ours and information can be focused and shared about the same thing. They do not realize that perceiving a given situation from different perspectives impacts what we know and believe to be true about it. They are not aware of our emotions or attitudes and as a result cannot prioritize what is the most important, urgent, or relevant to us. They do not draw on interaction to learn new skills. Due to all of

Fig. 1. Leonardo is a 65-degree of freedom (DoF) fully embodied humanoid robot that stands approximately 2.5 feet tall. It is designed in collaboration with Stan Winston Studio to be able to express and gesture to people as well as to physically manipulate objects. The robot is equipped with two 6-DoF arms, two 3-DoF hands, an expressive (24-DoF) face capable of near human-level expression, two actively steerable 3-DoF ears, a 4-DoF neck, with the remainder of the DoFs in the shoulders, waist, and hips. The left picture shows the robotic structure, the center picture shows the robot when cosmetically finished, the right shows a simulated version of the robot.

these shortcomings (to name a few) robots cannot cooperate with us as teammates or learn from us efficiently. As a result, human-robot interaction often is reduced to using social cues merely as an interface for supervising the robot's behavior. We argue that robots must be socially intelligent and must understand people in social terms in order to cooperate with us as capable partners.

1.2. *Outline of the paper*

This paper presents our work towards equipping our robot, Leonardo ('Leo') (see Fig. 1), with social skills relevant to learning and collaborating efficiently in human settings. Section 2 outlines what we view as the most pronounced challenges on this road. Section 3 describes the theoretical foundation of our approach. Section 4 applies this framework to model joint reference and object labeling as a collaborative process. Section 5 presents a more advanced example of socially guided learning, where a generalized task representation is acquired from human tutelage. Section 6 describes the robot's ability to perform the learned task jointly with a human collaborator. Finally, Sec. 7 offers a discussion comparing our approach to similar works in the realm of human-robot learning and collaboration.

2. Challenges in Designing a Cooperative Robot

There are a number of core challenges in creating a robot that can learn effectively from human guidance. In previous work we outlined several key issues in teaching a robot a new task, and described how social cues, offered during natural human teaching scenarios, can be used to address them.[5] Many of the important design challenges in robot learning apply to the collaboration domain and also draw on social structure.

2.1. *Knowing what matters*

Faced with an incoming stream of sensory data, a robot must figure out which of its myriad of perceptions are relevant to the task at hand. As its perceptual abilities increase, the search space becomes enormous. If the robot could narrow in on those few relevant perceptions, both the learning and the collaboration problem would become significantly more manageable.

Knowing what matters is fundamentally a problem of determining saliency, which can be guided either externally (by the environment), internally (by the robot's goals, etc.), or socially (by the human). Such cues help the robot to identify the most relevant items to consider and accelerate state-space discovery, where the robot learns new groups of features that have behavioral significance. Alternatively, temporal cues (such as saying "pay attention now") can be used to highlight a distinct environmental context or change that is relevant to the task. To facilitate this process, the state of the robot's attention must be transparent to the human partner so that he or she can easily infer what the robot is attending to, at what time, and what it is about to do.

2.2. *Knowing what action to try*

Once the robot has identified salient aspects of the scene, how does it determine what actions it should take? Knowing what action to try is a precondition to collaboration, and is the ultimate goal of a successful learning process. But there is added value to this skill early on: if the robot had a way of focusing on potentially successful actions, the learning problem itself would be simplified.

Determining which action to try can be addressed in a number of ways. For instance, the robot could experiment on its own by selecting an action based on past experience, as in reinforcement learning. For large state-action spaces this requires a prohibitively large number of trials. Alternatively, in learning by imitation, the robot is given the ability to follow the actions of the demonstrator. The demonstrator can be either a human,[6] or another robot.[7] The learner follows the model, thereby sharing a similar perceptual and motor state, to learn a reactive state-action policy.[8] This mapping often represents a shared inter-personal communication protocol, where the model announces the labels for particular sensory-motor states as they occur and the follower learns their association.[6] This has also been applied to learn a control policy for navigating a maze environment.[9]

A human teacher, however, can play a far more important and flexible role in guiding the learner's selection of the most promising actions in specific contexts. For those actions that the robot already knows how to perform, the human instructor can simply show or tell the robot what to do under particular circumstances. Friedrich and Dillmann make this process explicit where at each step the teacher provides information on which goal he wants achieved and what actions or objects are relevant to achieving that goal.[10] This approach, however, views teaching as explicit programming of the robot. It fails to capture the social and collaborative spirit of the teaching-learning process that characterizes human tutelage.

In contrast, we propose viewing action selection in a collaborative framework rather than as a form of explicit programming. The human intuitively guides the robot through the task, teaching it how to do the portions the robot does not understand yet. The instructor helps the robot determine what action to try either by showing the robot, through demonstration, or telling the robot what to do next, and correcting or encouraging the robot at relevant stages throughout this process.

2.3. *Knowing how to recognize success and correct failures*

Once the robot attempts to perform an action, how can it determine whether it has been successful? How does it assign credit for that success? Further, if the robot has been unsuccessful, how does it determine which parts of its performance were inadequate? It is important that the robot be able to diagnose its errors in order to improve performance. In many situations, this evaluation depends on understanding the goals of the task and intentions of the instructor or collaborator. Fortunately, a human can help the robot do this given that he or she has a good understanding of the task and knows how to evaluate the robot's success and progress.

One way that a human facilitates a learner or teammate's evaluation process is by providing feedback through a number of channels (facial expression, gesture, speech, tone of voice, etc.). This can be utilized in robot design, too. But in order to make feedback most effective, a complex web of social structure must be in place. There must be a shared understanding between human and robot that the instructor wants to teach the robot the goal of each step of the overall task. There must be communication to synchronize the desired outcome for each action, task, or subtask. The human must be able to confirm that the robot properly understood the desired outcome, and the robot in turn must be able to frame the possible feedback. To do so, it should combine verbal and non-verbal cues (internal, external or communicative) in order to properly determine the objects of the feedback act.

2.4. *Knowing how to explore*

In the type of robot cooperation we envision, the human instructor plays an important role in helping the learner explore its state-action space to discover a solution more quickly. We do not view learning and task performance (or collaboration) as separate activities. In a realistic robotic cooperation scenario that is situated in a changing environment, the learning process may never be complete. While the amount of exploration the robot performs reduces as it converges on a satisfactory plan of action, the basic performance-feedback loop remains the same throughout both learning and performance of the task. Thus, the robot's collaboration with a human should also be viewed as an exploration that is used to refine its performance and understanding of the task.

To allow for this process to take place, the human must collaborate closely with the robot even in the early learning stages, guiding its exploration as it quickly and efficiently learns each step of the task. At this stage the robot can leverage from the human's demonstration or verbal instruction to quickly infer the critical preconditions and desired outcome for each step, as well as how these steps relate to one another in the overall task structure. Then the robot should immediately demonstrate its understanding back to the instructor to confirm and move on or to correct if necessary. This prevents misunderstandings or mistakes from persisting for multiple steps — which would make them more awkward to correct later on. Indeed, the robot and human progress fluidly through the task at each step, confirming and correcting along the way. As a result of this collaboration, the robot will be able learn a sophisticated task structure, including learning task components such as actions and subtasks very quickly and efficiently from few examples.

3. Theoretical Background

As we explore applications where people interact with robots as capable partners, it is important to distinguish **human-robot collaboration** from other forms of human-robot interaction. Namely, whereas interaction entails action *on* someone or something else, collaboration is inherently *working with* others.[2,11] Much of the current work in human-robot *interaction* is thus aptly labeled as such, given that the robot (or team of robots) is often viewed as a tool with some limited autonomy that a remote human operator commands to carry out a task.[12–14] This sort of master-slave arrangement where the robot is operated as a tool (although it may be through an advanced interface incorporating speech or gesture input) does not capture the sense of partnership that we mean when we speak of working "jointly with" a robot as in the case of collaboration. Human-robot collaboration is an extremely important yet relatively unexplored topic among the many application and interface scenarios that fall under the domain of human-robot interaction.

In our work on robots capable of learning and working in collaboration with people we draw from Joint Intention and Situated Learning Theory. Both are described briefly in this section.

3.1. *Joint intention theory*

What characteristics must a humanoid robot have to collaborate effectively with its human partner? To answer this question, we look to insights provided by Joint Intention Theory.[3,4] According to Cohen and Levesque's theory, joint action is conceptualized as two or more partners doing something together as a team where the teammates share the same goal and the same plan of execution. In this situation it makes sense to speak of a *joint intention* related to the task at hand.

Several models have been proposed to explain how joint intention relates to individual intention. Searle argues that collective intentions are not reducible to individual intentions of the agents involved, and that the individual acts exist solely in their role as part of the common goal.[15] Bratman's analysis of Shared Cooperative Activity (SCA) introduces the idea of meshing singular sub-plans into a joint activity.[11] Bratman also defines certain prerequisites for an activity to be considered shared and cooperative; he stresses the importance of mutual responsiveness, commitment to the joint activity and commitment to mutual support. Cohen and his collaborators support these guidelines and provide the notion of joint stepwise execution.[16] Their theory also predicts that an efficient and robust collaboration scheme in a changing environment commands an open channel of communication. Sharing information through communication acts is critical given that each teammate often has only partial knowledge relevant to solving the problem, different capabilities, and possibly diverging beliefs about the state of the task. Our work integrates these ideas to model and perform collaborative tasks.

Communication clearly plays an important role in coordinating teammates' roles and actions to accomplish the task. It also serves to establish and maintain a set of mutual beliefs (also called common ground) among the team members. For instance, all teammates need to establish and maintain a set of mutual beliefs regarding the current state of the task, the respective roles and capabilities of each member, and the responsibilities of each teammate.

What happens when things go wrong? Teammates must share a commitment to achieving the shared goal. They cannot abandon their efforts, but must instead continue to coordinate their efforts to try a different, mutually agreed upon plan. Furthermore, each must be committed to complete their agreed-upon part of the execution plan, as well as be committed to others' success in doing theirs.[2,17] Specifically, the actions and goals that each team member adopts to do their part should not interfere with or prevent other team members from carrying out their parts.

Therefore, for cooperative behavior to take place, a mutual understanding for how those internal states that generate observable behavior (e.g. beliefs, intents, commitments, desires, etc.) of the human and the robot must be established so that the two can relate to one another. Furthermore, both human and robot must be able to reason about and communicate these states to each other so that they can be shared and brought in to alignment to support joint activity. Hence, human-style cooperative behavior is an ongoing process of maintaining mutual beliefs, sharing relevant knowledge, coordinating action, and demonstrating commitment to doing one's own part, helping the other to do theirs, and thus completing the shared task.

3.2. *Situated learning theory*

The ability to leverage off of cues that are given in a social interaction, such as the structure of the interaction and the context of the environment, are the bases of the Theory of Situated Learning. An important component of situated learning is Vygotsky's *Zone of Proximal Development*.[18] This theory hypothesizes that a child learns new skills from another person through the process of scaffolding, where the teacher provides structure and assistance in a new activity, such that the learner is able to achieve something they could not do independently.

In a situated learning interaction, the teacher must maintain a model of what the learner has accomplished and the learner's current beliefs about the task. The teacher should know what parts are clear and which are still confusing so that she can provide timely feedback and modify the structure of the task as appropriate. In previous work, we outline a number of social cues that people employ to naturally structure and facilitate the learning process for others,[5] arguing that these could be utilized by a socially savvy robot to learn from natural human instruction in a similar manner (see a detailed discussion in Sec. 7). These include the ability to direct and share attention, participate in turn taking, understand expressive feedback, engage in guided exploration, and more. These social interactions serve to structure, constrain, and guide the learning process of the robot. However, a robot must understand and appropriately respond to these social cues in order to effectively utilize them to constrain and guide its own learning.

The learner assists in this process by communicating their internal state through a variety of expressive acts such as gestures, expressions, and vocalizations. The result is a tightly coupled interaction that allows both the teacher and the learner to guide the process through their corresponding models of the learner's progress. This collaborative teaching process generally takes the form of a dialog that can be usefully considered as the coordination of joint intentions and mutual beliefs.[16,19] Hence, through reciprocal interaction, both the learner and instructor cooperate to (i) help the instructor maintain a good mental model of the learner, and (ii) help the learner leverage from instruction and guidance to build the appropriate task models, representations, and associations.

4. Learning Names of Objects and Following Requests

In this section we describe an initial scenario where the robot follows a human's instruction to associate names with objects in the world and then follow requests to perform actions on the labeled objects (see Fig. 2). This example illustrates how communication between human and robot is modeled as a collaborative dialog where the human uses speech and gesture to communicate to the robot, and the robot communicates with the human using gesture, shifts of gaze, and expressive cues.

The ability to engage in deictic reference in order to learn the names of objects (e.g. "This is the red button.") and then to refer to these labels in the context of a task (e.g. "Turn the red button on.") can be considered a rudimentary form of cooperative learning, as well as an atomic skill for our collaborative scenarios. For instance, in Sec. 5 we build on this ability to teach Leonardo generalized tasks through guided experience where the robot learns the structure of the task and associates goals at various levels in the task representation. This goal-oriented representation then paves the way for successful collaboration between the robot and a human partner on these learned tasks (see Sec. 6).

4.1. *Speech understanding*

Leonardo's speech understanding capability is based on the Nautilus system (under development at the Navy Research Lab by Alan Schultz and his collaborators). It uses ViaVoice as a front end speech recognition system, supports a basic vocabulary, tracks simple contexts, and performs simple dialogs that involve pronoun referents, basic spatial relations (left/right, near/far, front/back, etc.), and shifts in point of view (with respect to my reference frame versus your reference frame, etc.).[13] The vocabulary has been tailored to support what Leonardo perceives during his interactions with objects and people, including: actions (grasping, pressing, looking, etc.), entities (buttons, people, etc.), features (color, button-ON, button-OFF, shape, size, etc.), and gestures (pointing, head nods, etc.).

4.2. *Visual perception*

Leonardo visually perceives the surrounding environment with two camera systems. The first is a wide-angle stereo head that is placed behind the robot to provide peripheral vision information. This system is used to track people and objects in Leonardo's environment. The second is a stereo camera (with a narrower field of view) that is mounted in the ceiling and faces vertically downward to view Leonardo's workspace. This stereo camera is used to track pointing gestures and objects in the workspace in front of Leonardo (e.g. the buttons based on their shape, size, color, and position). This visual information is normalized to real-world coordinates and calibrated to Leonardo's frame of reference. These visual systems allow the robot to detect deictic gestures (discussed below) used by humans to refer to objects and to direct the robot's attention to important aspects of the shared context.

4.3. *Deictic reference and shared attention*

In humans, understanding attention and being able to share attention with others is a critical skill in performing cooperative tasks, communicating with others, and learning from tutelage.[20] Furthermore, it is important to note that the act of "referring to" is fundamentally a collaborative process in human-human communication.[21] The goal is to establish a state of mutual belief among those

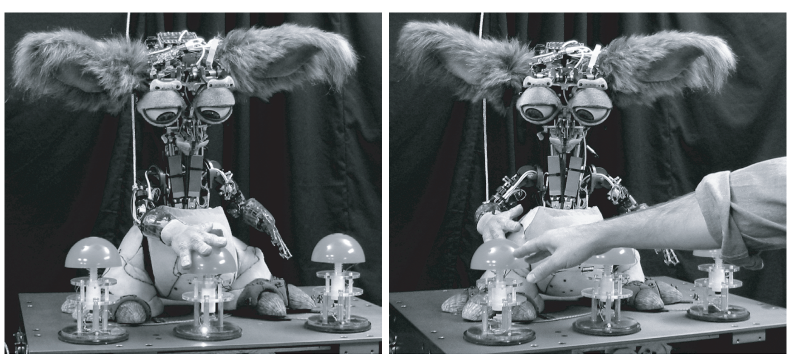

Fig. 2. Leonardo following the human's request to activate the middle button (left). Leonardo learns the labels for each of his buttons by having a person point to a specific button and name it (right). In this picture, Leonardo and the human are both attending to the same button as Leo learns what to call it.

sharing joint attention. In making a reference, the speaker intends that the reference become part of both speaker's and listener's mutual knowledge. Thus, the speaker must adapt his behavior, elaborating or clarifying, whenever he thinks that the listener does not properly understand him. Similarly, the listener must indicate to the speaker that he is being understood (or not). These are important skills for collaborative referential behavior.

Having both human and robot look at the same thing is important part of this process. This is called referential looking and has been demonstrated by previous robots.[22–24] However, these earlier systems did not explicitly represent the attentional state, the referential focus, and the resulting beliefs of the human collaborator. To address these issues and shortcomings, joint attention for Leonardo is modeled as a collaborative process. Accordingly, we have developed an attentional system for Leonardo that determines the robot's focus of attention, monitors the attentional focus of the human, and uses both to keep track of the mutual beliefs and referential focus held by both. Therefore, the robot not only has a model for its own attentional state, but models that of the human as well.

Leonardo's attentional system computes the level of *saliency* (a measure of "interest") for objects and events in the robot's perceivable space. This 3D space around the robot, and the objects and events within this space, are represented by the vision system. The attention system operates on this 3D spatial representation to assign saliency values to the items therein (see Fig. 3). There are three kinds of factors that contribute to the overall saliency of something. These include its *perceptual properties* (its proximity to the robot, its color, whether it is moving, etc.), the *internal state* of the robot (i.e. what the robot is searching for and other goals, etc.),

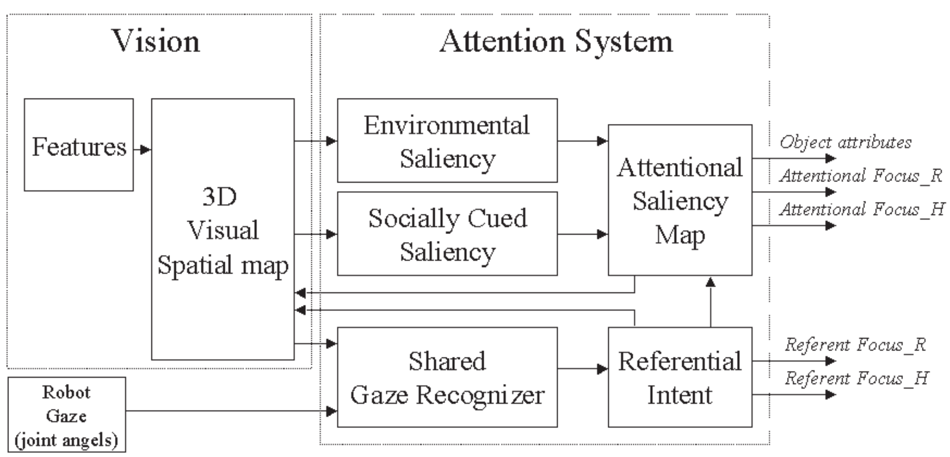

Fig. 3. Schematic of the robot's visual attention system for external factors. It enables the robot to share joint attention with people. The saliency based system deals with environmental aspects of saliency (color, movement, etc.). The socially based system deals with social cues that direct attention (speech, head pose, pointing, etc.). Internal factors come from the robot's cognitive system (not shown).

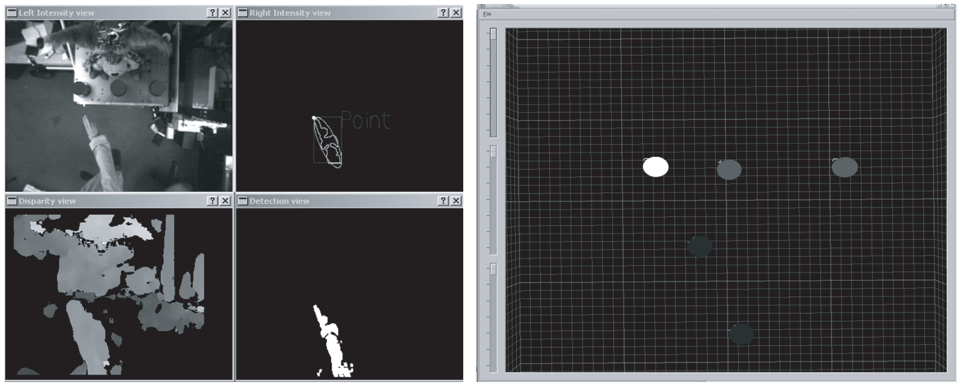

Fig. 4. Computing the deictic reference to an object in the visual scene. Left, an overhead stereo camera identifies the locations of the buttons in the scene and recognizes when a pointing gesture occurs, estimating the location of tip of the finger and the angle of the forearm. This is passed to the spatial reasoning system (right). This overhead viewpoint shows the buttons (red), the location of the tip of the finger and base of the forearm (blue), and the identified object referent (white).

and *socially directed reference* (pointing to (see Fig. 4), looking at, or talking about something to bring something selectively to the robot's attention). For several of these factors, the saliency measure is a time-varying quantity. For instance, socially directed reference assigns a very high saliency upon appearance of the gesture and then gradually decays over time. This strongly biases the robot to participate in referential looking with a human before attention shifts elsewhere.

For each item in the 3D spatial representation, the overall saliency at each time step is the result of the weighted sum for each of these factors. This is done using a similar approach as described in Ref. 25. The left diagram of Fig. 8 shows those factors considered for computing environmental saliency. The right diagram shows those for social cues. The item with the highest saliency becomes the current *attentional focus* of the robot, *AttnFocus_R*. This is where the robot's gaze is directed.[24] The *referential focus, RefFocus_R* is determined as the last object that was the subject of shared attention between robot and human.

Using the same 3D spatial map, the robot also monitors what objects the human looks at, points to, and talks about over time. These items are assigned a tag with value, *Saliency_H*, that indicates which objects have been the human's focus of attention and therefore have been salient (of interest) to him or her. This allows the robot to keep track of items that both human and robot are mutually aware (i.e. the common ground). The human's current attentional focus, *AttnFocus_H* is defined as at what he or she is currently looking. The human's referential focus, *RefFocus_H* is determined by the last object that was the object of shared attention with the robot. See Fig. 5, where robot and human are sharing joint visual attention via head pose and pointing gesture.

These attention following and directing skills can be accompanied by conversational policies along with gestures and shifts of gaze for repair, elaboration, and

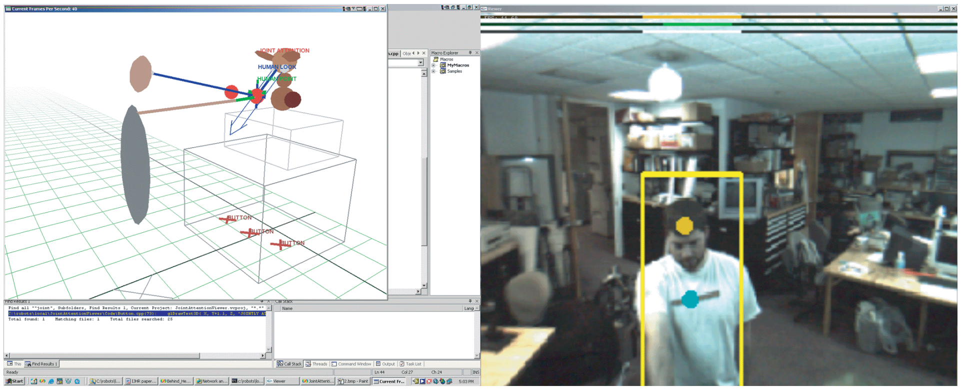

Fig. 5. Visualizer showing the robot and a human sharing joint visual attention on the same object. The right image shows the visual input of a person looking at and pointing to the center button. The left image shows the visualization of the robot's internal model. The human's gaze is shown as the blue vector and his pointing gesture is shown by the brown-green vector. The robot looks at the same button (robot's blue vector) to establish joint attention.

confirmation to confirm a shared referential focus and to maintain mutual beliefs between human and robot. For instance, these skills are of particular importance for situations where an occluding barrier forces a robot and its human teammate to see different aspects of the workspace. In short, human and robot will have to share information, and direct the attention of the other, to establish and maintain a set of mutual beliefs and the same referential focus.

4.4. *Conversational policies*

To enable the robot to participate in simple collaborative dialogs, we have implemented a suite of a collaborative task-oriented conversation and gestural policies for Leonardo. Cohen *et al.* argue that much of task-oriented dialog can be understood in terms of Joint Intention Theory (see Sec. 3.1).[16] Accordingly, each conversant is committed to the shared goal of establishing and maintaining a state of mutual belief with the other. To succeed, the speaker composes a description that is adequate for the purpose of being understood by the listener, and the listener shares the goal of understanding the speaker. These communication acts serve to achieve robust team behavior despite adverse conditions, including breaks in communication and other difficulties in achieving the team goals.

Cohen *et al.* analyzed task dialogs where an expert instructs a novice on how to assemble a physical device.[16] We have implemented conversation policies for those key discourse functions identified by Cohen and his collaborators. These include discourse organizational markers (such as "now," "next," etc.) that are used to synchronize the start of new joint actions, elaborations when the expert does not believe that the apprentice understands what to do next, clarifications when the apprentice does not understand what the expert wants next, confirmations so that both share the mutual belief that the previous step has been attained, and referential elaborations and confirmations of successful identification to communicate the important context features for each step in the task.

It is important to note that expressive cues such as head gestures and facial expressions can be used to serve this purpose as well as speech acts. For instance, Leonardo performs head nods for confirmations (and shakes is head to not confirm), and he shrugs his shoulders with an expression of confusion to request clarification or elaboration from the human instructor. The robot looks to the object that is currently being named by the instructor to confirm successful identification of the target. Leonardo then looks back to the human to confirm that it has finished associating the label with the appropriate object. The robot can demonstrate its knowledge of the object names that it has been taught by pointing to the correct object in response to the human's query "Which is button 1?" This confirms that both human and robot share the same belief regarding which object is called by what name.

4.5. *Turn taking skills*

We have supplemented our models of collaborative dialog and gesture with flexible turn-taking skills modeled after those used by humans.[26] The exchange of speaking turns in human conversation is robust despite interruptions, incomplete utterances, and the like. Well studied by discourse theorists, humans employ a variety of paralinguistic social cues, called envelope displays, to manage who is to talk at which times in an intricate system of turn taking.[26] These paralinguistic social cues (such as raising one's brows and establishing eye contact to relinquish one's speaking turn, or looking aside and positioning one's hands in preparation to gesture in order to hold one's speaking turn when speech is paused) have been implemented with success in embodied conversational agents,[27,28] as well as expressive robots.[29,30]

A number of envelope displays have been implemented on Leonardo to facilitate the exchange of turns between human and robot. To relinquish its turn, Leonardo makes eye contact with the person, raises its brows, and relaxes its arms to a lower position. As the person speaks, the robot continues to look attentively at the speaker and perks his ears so that she knows that the robot is listening to her. When she has finished her utterance, Leonardo lifts its arms to show initiative in taking its turn and breaks eye contact — often looking to the object that the person referred to in her last utterance (e.g. to one of the buttons).

In addition, back-channel signals are given by the listener to let the speaker know that she is being understood. These are important skills for robots that must engage humans in collaborative dialog where communication signals (both verbal and nonverbal) are frequently exchanged to let the conversant know that each properly understands the other — and equally important, when communication breaks down and needs to be repaired. If Leonardo cannot parse the person's utterance, for instance, the robot displays a look of confusion to indicate that it is having problems understanding the speaker. A small, confirming nod is given to indicate when the robot has understood the utterance.

4.6. *A sample dialog*

For the human, teaching Leonardo the names of objects is straightforward — one simply has to point to the desired object and name it for the robot, e.g. "Leonardo, this is the blue button." Given the use of the word "this" in the utterance, the speech understanding system recognizes that the object referent is being communicated to the robot through a deictic gesture. As a result, the speech understanding system passes the linguistic label, e.g. "blue button," to the robot's belief system. We use a whole object bias — the human's point gesture is assumed to refer to a single object in the world (rather than a feature or property of the object). This is a pragmatic simplification as well as a tendency based in psychological findings. Bloom describes children's word learning abilities, including a whole object bias for word reference.[31]

With the information from the vision and speech systems, the cognitive system binds all relevant features into a coherent and persistent belief that an object exists in the world, at a particular 3D location, with the specified name, and is characterized by a number of visual features (color, size, etc.). Furthermore, Leonardo socially cues to the human that he understood the intent of the person's utterance by looking to the object referent as the human points to it and names it, and looking back to the person's face once he has formed the corresponding belief.

If the robot did not understand the utterance, he communicates this to the human as well. Immediately upon hearing a person say something (triggered by sensing a sound threshold on the person's microphone), Leonardo looks to the person's face with an interested expression and perks his ears. This tells the person that Leonardo heard her and is attending to her. If the robot did not understand the utterance, he will display a look of puzzlement that is meant to trigger a repetition from the human (e.g. widen his eyes, lean forward and raise an ear towards the person). This allows the person to correct the misunderstanding and conveys to the human that the robot is committed to trying to understand her by letting her know when she is understood and when she is not. Additionally, if the robot did understand the utterance but was not able to reliably identify an object of attention to relate to the labeling request, he will communicate his confusion by cocking his head to the side and shrugging, thus encouraging a refinement of the pointing gesture and a repetition of the request.

Subsequently, the human can confirm the robot's understanding in a number of ways. For instance, the person can simply ask Leonardo to "Show me the green button." If Leonardo has assigned the "green button" label to an object, the robot responds by first looking to that object, pointing to it, and then looking back at the human to await their response. If the robot has not assigned a "green button" label to an object (perhaps the person did not label anything a "green button" yet, or perhaps the robot did not understand when she did so) the robot communicates that it does not know this by showing an expression of confusion (cocks its head to the side and shrugs). Alternatively, the person can also ask Leonardo, e.g. "Can you show me the green button?" If Leo has assigned a "green button" label to an object, the robot will enthusiastically nod its head "yes" and then point to the corresponding object. Otherwise, Leo will disappointedly shake its head "No" while shrugging to indicate that he cannot complete the request because he does not have enough information.

Therefore, by using natural social cues, the human instructor can quickly assess what the robot has associated and what it has not, and immediately correct any misunderstandings if needed before moving on to the next thing. This leads to an efficient training scenario where errors are quickly remedied, rather than being allowed to persist and thereby potentially lead to more errors and miscommunication in the future. Instead, this natural teaching scenario with immediate communicative feedback, allows both robot and human to establish and maintain a set of mutual beliefs about the situation at hand in a fluid and natural manner.

Once the robot has learned the names of objects, Leonardo can follow human requests to attend to or to manipulate those items. For instance, the human can ask Leonardo to "Look at the green button." To respond, the robot directs its gaze from the human's face to the desired object. After a short period, the robot looks back to the speaker to await her next utterance. The human can also direct the robot to operate on an object, such as "Leonardo, press the green button." The robot responds by applying the desired action to the specified object.

5. Collaborative Task Learning and Execution

Building on the ability to request that Leonardo perform actions on objects in the world, our next endeavor involves the ability to teach the robot more complex tasks made up of these actions. Moreover, we wish to have the robot learn in such a way that it can later perform these tasks either autonomously or in cooperation with a human partner. To allow this, human-robot communication must extend beyond object reference and atomic actions to task representations that include object properties and goals.

5.1. *Representing tasks and goals*

A variety of studies indicate that humans interpret actions based on goals and intentions rather than specific activities or motion trajectories.[32–34] Thus, goals are central to our task representation. In our collaborative task setting, this is particularly important, since goals provide the common ground for communication and interaction. Additionally, viewing tasks in terms of goals allows for alternative paths of action leading towards the same result, simultaneous action of two of more agents as part of the same activity and a dynamic division of labor, throughout the execution of a task.

Action tuples serve as the atomic representational elements of the robot's learned task models and are comprised of set of preconditions, until-conditions and executables that represent an atomic or compound action.[35] We represent tasks and their constituent actions in terms of action tuples with the additional notion of goals. In this representation, goals play a role both in the *precondition* that triggers the execution of a given action tuple, and in the *until-condition* that signals when the action tuple has successfully completed.

Our task representation distinguishes between two types of goals: (a) state-change goals that represent a change in the world, and (b) just-do-it goals that need to be executed regardless of their impact on the world. These two types of goals differ in both their evaluation as preconditions and in their evaluation as until-conditions. As part of a precondition, a state-change goal must be evaluated before doing the action to determine if the action is needed. If the action or task's final state is achieved, either by initial conditions, change in the environment, or a collaborator's actions, the execution of this action tuple should naturally not occur. As an until-condition, the robot shows commitment towards a state-change goal by executing the action, repeatedly if necessary, until the robot succeeds in bringing about the new state. This commitment is an important aspect of intentional behavior.[4,11] Conversely, a just-do-it goal will lead to an action regardless of the world state, as the action itself is the goal of its execution. Also, a just-do-it goal's final state cannot be externally evaluated and will therefore only be performed once.

When executing a task, goals as preconditions and until-conditions of actions or sub-tasks manage the flow of decision-making throughout the execution process. Overall task goals are evaluated separately from their constituent action goals and are independent from the particular recipe the robot has to achieve the task goal. Thus, for example, if the task goal is to screw a set of bolts into place, this does not necessarily translate to a set of sub-goals for each bolt, but can also be satisfied by observing a teammate performing the task successfully (seeing the overall goal state achieved). This top-level evaluation approach is not only more efficient than polling each of the constituent action goals, but a goal-oriented implementation supports a more realistic groundwork for intentional understanding — i.e. performing the task in a way that accomplishes the *overall intent*, rather than just mechanically going through the motions of performing the constituent actions. This approach is also crucial for teamwork and collaboration.

Tasks are thus represented as a hierarchical structure of actions and sub-tasks (the latter defined in the same fashion). Tasks contain a list of actions (with an action goal each and ordering constraints specified among the actions) and a representation of the overall task goal. An action is one of three types: one of Leo's primitive actions, a sub-task, or an abstract action (a result of generalization discussed below). Since tasks, sub-tasks, and actions are derived from the same action tuple data structure, this naturally affords tree-structured tasks. Goals are associated at all levels in the representation: individual actions, sub-tasks, and the overall task.

5.2. *Task manager*

The *task manager module* maintains a collection of known task models and their associated names. Given this set of tasks, the robot listens for speech input that indicates a task-related request from the human partner. These can be in the form of: "Leo, do *task x*" or "Leo, let's do *task x*." These requests can also be made in the form of a question: "Leo, can you do *task x*?" In the case of a question, given Leonardo has no speech generating capabilities yet, the robot will answer by either nodding "yes" or shaking its head "no." If the robot does not recognize the name of the requested task, or if the robot does not know how to perform it, he looks puzzled or shrugs his shoulders, indicating "I do not know."

5.3. *Learning through guided experience*

In the case of encountering an unknown task, the human partner can offer to teach the task ("I can teach you to X . . . "). At this point Leo nods his head, confirming that the learning process has begun. This exchange is important since it initiates the learning process and establishes a mutual belief about the roles of teacher and learner.

The human walks the robot through the components of the task, building a new task from its set of known actions and tasks. While learning, the robot pays attention to what actions it is being asked to perform, encoding inferred goals with these actions. To encode the goal state of a performed action or task, Leo compares the world state before and after execution. If this action or task caused a change of state, this change is taken to be the state-change goal. Otherwise, it is assumed to be a just-do-it goal. This produces a hierarchical task representation, with a goal encoded for each individual part of the task as well as for the overall task.

Currently goals are object-based (where objects are co-located sets of perceived features). The task goal is defined as a set of beliefs about what must be true in the world for this task to be achieved. A goal belief is made for each object in the world that incurred any change during the action or task, and each belief is a set of criteria perceptual features and expectation perceptual features. The criteria features are those that are held constant over the task and the expectation perceptions are those features of the object that changed over the task. For example, if upon finishing Task-X the world state has changed such that a particular toy A remained the same and toys B and C changed from green to red. The goal for Task-X is a state-change goal with two goal beliefs one for each object B and C. Each goal belief has criteria features [type = toy, location = xyz, label = B/C, etc] and an expectation feature [color = red]. This allows for a straightforward evaluation of task goals: for each goal belief, find an object that meets the criteria features and check that the expectation features hold true.

5.4. *Hypothesizing about the task representation*

There are a number of ambiguous situations in a learning interaction. One of them is determining what the overall goal is when learning a new task. When told a task is complete, the robot has to determine which of its actions and effects on the world should be considered the 'overarching goal' of the learned task. Therefore, our task representation is flexible. Rather than associate a single task representation on the first example, each task representation maintains a number of task hypotheses about what the important actions and goals of the task could be. A task hypothesis has a set of actions and goal beliefs and a record of the number of examples that have been consistent with this particular task representation.

Generalization is possible whenever there is a common action denominator to the constituents of a task. Thus, to expand the hypothesis space for a new example of a task, the system first checks for similarity in the actions performed for this task — i.e. all of the actions of the task are of the same primitive type but just have different objects of attention. If this is the case, the primitive action type is noted as the *generalized task action*. Next the system looks at each of the goal beliefs (each of the objects that incurred some change) and collapses these into a single common goal belief (CGB) containing the criteria and expectation features common to all. Then, a number of task hypotheses are made. In each hypothesis, the action is taken to be the generalized task action, and the goal is one of the combinations of goal features from the CGB. For example, if the CGB has four features, one hypothesis will be the generalized task action and a goal belief with all four features (the most specific hypothesis). Another hypothesis will be the generalized task action and a goal belief with three of the four features, and so on. This expansion results in a hypothesis space of task representations that are consistent with the current example of the task.

A single task representation needs to be used when this task is requested next. This is chosen through a Bayesian likelihood method. The likelihood of each of these hypotheses is calculated according to Bayes' rule (Eq. (1)). The probability of a hypothesis, h, given some data, D, is proportional to the probability of the data given the hypothesis scaled by the prior likelihood of the hypothesis:

$$P(h \mid D) \sim P(D \mid h)P(h). \tag{1}$$

In the case of task learning, the data, D, is the set of all examples seen for this task. Thus the probability of D given each particular hypothesis, h, is calculated as the percentage of the examples that contain the characteristics of h (i.e. the state change seen for the example is consistent with the goal representation for h). For prior likelihoods our algorithm prefers specificity to generality (in terms of the number of features in the goal belief). For example, when a task is first learned, every hypothesis is equally represented in the data and the algorithm chooses the most specific representation of this task for the next execution.

5.5. *Task execution*

If Leo is asked to do a task independently that he has already learned, the task manager starts the execution module, expanding the task's actions and sub-tasks onto a *focus stack*.[36] Task execution proceeds by working through the actions on the stack popping them as they are done or, for a sub-task, pushing its constituent actions onto the stack.

The particular task representation that Leo is using to perform a requested task has a confidence associated with it (its relative likelihood compared to the other hypotheses that were available). If the current task representation is not significantly more likely than the others, then Leo expresses his tentativeness to the instructor by frequently looking between the instructor and the object of an action's attention.

Upon finishing the task, Leo waits for feedback from the teacher, leaning forward with his ears perked slightly. The teacher can give positive or negative verbal feedback. With positive feedback, Leo will consider the task complete. However, with negative feedback, Leo will revert to learning mode, expecting the teacher to lead him through the completion of the task. This refinement stage results in a new example, created in a similar fashion to the way in which an original task example was created in learning mode.

Every execution of a task is essentially a new example, thus, the robot is always learning. Upon finishing a task execution — whether it required refinement or not — a task hypothesis is created along with the expanded hypothesis space. This is identical to the expansion process described in the initial learning phase. For each hypothesis, if it does not already exist in the task representation it is added to the hypothesis space, otherwise the coverage is incremented for the existing hypothesis. Finally, the best hypothesis is chosen, for the next execution of this task, through the Bayesian likelihood method described previously.

5.6. *Learning results*

The tutoring of tasks exemplifies our approach to teaching as a collaborative interaction. Joint attention is established both on the object level and on the task structure level. The robot uses subtle expression to indicate to the human tutor when it is ready to learn something new, and its performance of taught actions provides the tutor with immediate feedback about comprehension of the task. Envelope displays such as gaze aversion, eye contact, and subtle nods and are used to segment a complex task learning structure in a natural way to the tutor.

We have demonstrated Leonardo's learning abilities on several tasks comprised of simple dialogs and object manipulation skills. In our experimental scenario, there are three buttons of different colors in front of Leonardo. The buttons can be pressed ON or OFF, which switches an LED on or off. Occasionally, a button does not light up when pressed; this is considered a failed attempt. We designed a task set representing both simple and complex hierarchies and tasks with both state-change and just-do-it goals. In our initial trials, the robot demonstrates its understanding of nested action by recalling tasks that have been learned as sub-tasks of larger activities. Figure 6 shows an example of a learned task hierarchy: Turning the buttons ON, and Turning the buttons OFF, within the larger task of Turning the buttons ON & OFF. The robot correctly associates state-change goals and just-do-it goals in learning new tasks (demonstrated in Leo's understanding of when to perform an action and for how long to persist based on its initial success).

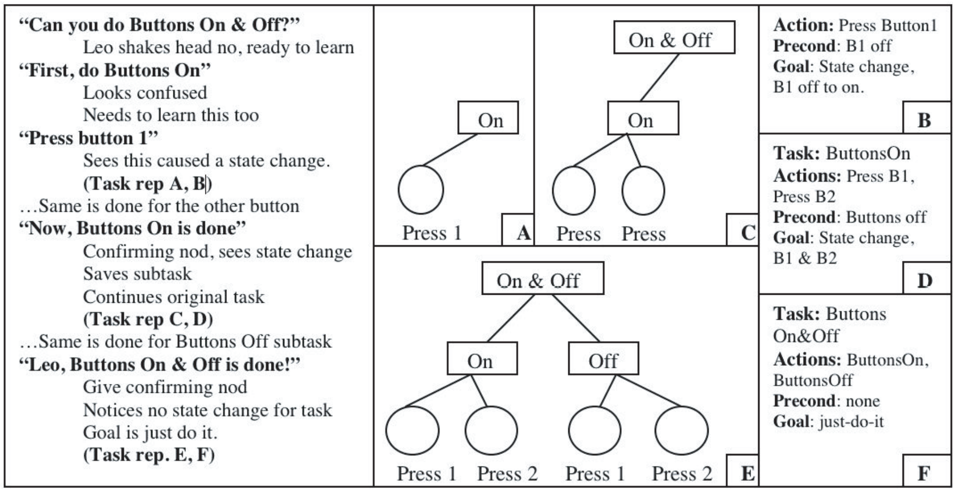

Fig. 6. The diagram on the left is an example hierarchical task representation for the task "Buttons-On-and-then-Off." The right box describes the corresponding task, subtask, and action level goals. The goal associated with the overall task is the completion of subtasks and actions. On the other hand, state changes were seen with each of the subtasks and their respective actions; therefore, these state changes are taken to be the goals.

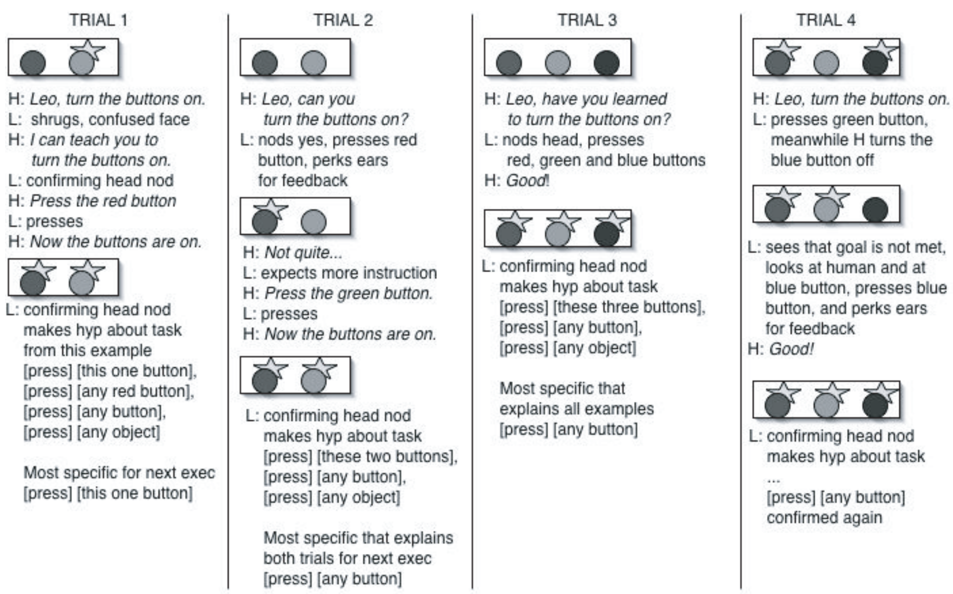

Fig. 7. Four trials of a task learning session, in which a human (H) teaches Leo (L) to 'Turn the buttons ON.' The leftmost button is red, the center button is green, and the rightmost button is blue. Note that in Trial 3 a novel configuration of buttons has been introduced (the blue button has been added) and the robot can generalize the goal to this new configuration and perform it successfully. In Trial 4, the human simultaneously turns the blue button off as Leonardo tries to complete the goal of having all buttons on by turning on the green button. The robot notices this change and demonstrates commitment to the goal by turning the blue button back on.

The gestural cues provide crucial feedback, enabling the tutor to realize when the robot successfully understands a task. Figure 7 shows an actual transcript from a session in which a human teaches Leo to 'Turn the buttons ON.' In the first trial, the initial conditions had the green button already on, so the human only had to ask Leo to press the red button. This led to a variety of hypotheses about the actual task representation, but the most specific is chosen for the next execution of 'Turn the buttons ON.' In the second trial, the human purposely structures the task to resolve an ambiguity from the previous example, and to give Leo another key example of 'Turn the buttons ON.' After this example, the most specific hypothesis that explains both examples is to 'press any button,' which is the correct representation. Trial 3 shows that Leo has learned this generalized goal and can exhibit the correct behavior in a novel configuration. In Trial 4, he demonstrates understanding and commitment to the overall task goal.

While this example is in a low dimensional feature space and little data is needed to resolve the ambiguities, even in this space the advantage of the tutelage paradigm is evident. Through a dynamic turn-taking interaction with a human partner, the robot quickly acquires the representative examples needed to generalize to the correct task representation. The robot's communication of learned abilities through demonstration and gesture helps the human teacher understand what the robot knows and what ambiguities need resolution.

6. Performing a Learned Task in Collaboration with People

Our goal-oriented representation affords task collaboration between the robot and a human partner. We have implemented a turn taking framework in which the human collaborator and Leonardo can work in partnership to achieve a common goal. This is made possible by continually evaluating both the state of the task and the state of the world before trying to execute an action.

A robot performing a complex goal-oriented task requires mechanisms to analyze the world around it, dynamically evaluate goal states and combine these into a useful action selection process. When collaborating with a human partner, however, many new considerations come into play, inspired by the vast body of theory on joint intention. For instance, within a collaborative setting the task can (and should) be divided between the participants; the partner's actions need to be taken into account; mutual support must be provided in cases of one participant's inability to perform an action; a clear channel of communication must be used to establish mutual beliefs and maintain common ground for intentions and actions.

Our implementation begins to address these considerations as the robot engages in a collaborative discourse while progressing towards achieving the joint goal. To do so, and to make the collaboration a natural human interaction, we have implemented a number of mechanisms that people use when they collaborate. In particular, we have focused on communication acts to support joint activity (utilizing gestures and facial expressions), dynamic meshing of sub-plans and turn taking.

6.1. *Dynamic meshing of sub-plans*

Leo's intention system is a joint-intention model that dynamically assigns tasks between the members of the collaboration team. As part of a team, each member must decide on their individual intentions as part of the joint plan. Leo derives his intentions based on a dynamic meshing of sub-plans according to his own actions and abilities, the actions of the human partner, Leo's understanding of the common goal of the team, and his assessment of the current task state.

As stated in Sec. 3, communication is vital for successful cooperation. Leonardo is able to communicate with the human teammate about the commencement and completion of task steps within a turn-taking interaction. Specifically, the robot is able to recognize changes in the task environment, as well as successes and failures on both Leo's and his teammate's side. Moreover, Leonardo is able to communicate to the human teammate his inability to accomplish a task step crucial to the complete joint action.

6.2. *Self assessment and mutual support*

During the collaboration, we employ a turn-taking approach. Before attempting an element of the task, Leo negotiates who should complete it. Leo uses one of two behaviors based on the robot's evaluation of its own capabilities. In the context of the button task, Leonardo can assess whether he can reach each button or not. If he is able to complete the task element (e.g. press a particular button) he will offer to do so. Conversely, whenever he believes that he cannot do the action (e.g. because he cannot reach the button) he will ask the human for help.

Given that Leonardo cannot speak yet, the robot is limited to indicate willingness to perform an action by pointing to itself, and adopting an alert posture and facial expression (Fig. 8(a)). Similarly, when detecting an inability to perform an action assigned to him, Leo's expression displays helplessness, as he gestures toward the human in a request for her to perform the intended action (Fig. 8(b)). Leo also

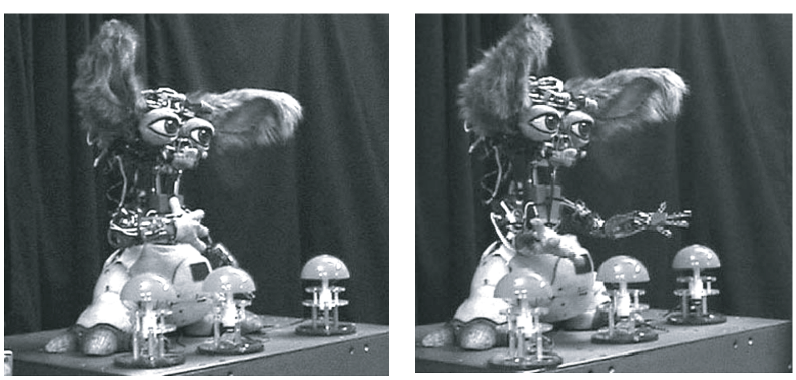

Fig. 8. (a) Leonardo negotiating his turn for an action he is able to perform, and (b) Leonardo asking for help.

shifts gaze between the problematic button and his partner to direct her attention to what it is he needs help with.

6.3. *Communication to support joint activity*

While usually conforming to this turn-taking approach, the robot can also keep track of simultaneous actions, in which the human performs an action while Leo is working on another part of the task. If this is the case, Leonardo will take the human's contribution into account and reevaluate the goal state of the current task focus. He then might decide to no longer keep this part of the task on his list of things to do. However, the robot needs to communicate this knowledge to the human to maintain mutual belief about the overall task state. Another case of simultaneous action handling, in which the human changes the world state in opposition to Leo's perceived task goal. In this case, Leo's commitment to the goal and dynamic evaluation results in the reversal of the human's simultaneous action.

We have implemented a variety of gestures and other social cues to allow the robot to communicate his internal state during collaboration — such as who the robot thinks is doing an action, or whether the robot believes the goal has been met. For instance, when the human partner unexpectedly changes the state of the world, Leo acknowledges this change by glancing briefly towards the area of change before redirecting his gaze to the human. This post-action glance lets the human know that the robot is aware of what she has done, even if it does not advance the task.

If the human's simultaneous action meets a task goal, such as turning the last button ON during the 'Turn the buttons ON' task, Leo will glance at the change and give a small confirming nod to the human. Similarly, Leo uses subtle nods when he thinks he completed a task or sub-task. For instance, Leo will give an acknowledgement nod to the human after completing the 'Turn the buttons ON' sub-task and before starting the 'Turn the buttons OFF' sub-task, in the case of the 'Turn the buttons ON then OFF' task.

We have conducted a few early experiments using the framework described herein, and have found these cues to play a significant role in establishing and maintaining mutual beliefs between the teammates on the progress of the shared plan, and in increasing the efficiency of the human-robot collaboration process. Table 1 shows a sample transcript describing typical task collaboration between Leonardo and a human teammate. We chose to display the following simple tasks for reasons of transcript brevity: *BUTTON-ONE* — Toggle button one, *BUTTON-ONE-AND-TWO* — Turn buttons one and two ON. While these do not illustrate Leonardo's full range of goal-oriented task representation capabilities, they offer a sense of the joint intention and communicative skills fundamental to the collaborative discourse stressed in this section.

Note that Fig. 7, Trial 4, was another case of simultaneous action handling, in which the human changes the world state in opposition to Leo's perceived task goal. In this case, Leo's commitment to the goal and dynamic evaluation results in

Table 1. Sample task collaboration on single-level task. Frames 2–14 present three collaborations on the BUTTON-ONE task (toggling a single button). In the first collaboration, Leo negotiates and completes the task himself. On the second, the human partner completes the task, and Leo's eye gaze and gestures help to communicate mutual beliefs about the task state. The third time, the button is out of reach and Leo sees that he has to ask the human to complete the task. Frames 15–19 present the BUTTON-ONE-AND-TWO task (pressing two buttons ON). This scenario shows Leo's ability to dynamically take his partner's simultaneous actions into account, again using gesture and eye gaze to maintain mutual beliefs about the task state.

#	Human	Leonardo	Notes
1	"Leo, let's do task BUTTONS"	Shrugs "I don't know"	Leo does not know this task
2	"Let's do task BUTTON-ONE"	Looks at the buttons	Leo acknowledges that he understands the task, and visibly establishes mutual belief on the task's initial conditions
3		Points to himself	He can do the first (and only) part of the task, and suggests doing so
4	"OK, you go"	Presses button one, looking at it	Looking away from the partner while operating establishes turn taking boundaries
5		Looks back at his partner	Gaze shift is used to signal end of turn
6		Nods shortly	Communicates the robot's perceived end of task
7	"Leo, let's do task BUTTON-ONE"	Looks at the buttons; points to himself	As in steps 2 and 3
8	"I'll go"	Looks at his partner	
9	Presses button one	Looks at button one	Acknowledges partner's action, creates mutual belief
10		Nods shortly	Communicates perceived end of task
11	Moves button one out of Leo's reach		
12	"Let us do task BUTTON-ONE"	Looks at buttons	Leo acknowledges that he understands the task, and visibly establishes mutual belief on the task's initial conditions
13		Looks at button one, then back at the human partner; extends his arms in "Help me" gesture	Leo assesses his capabilities and consequently requests support
14	Presses button one	Looks at button one; looks back at human; nods shortly	Glance acknowledges partner's action; nod creates mutual belief as to the task's completion
15	"Let us do task BUTTON-ONE-AND-TWO"	Looks at buttons	Leo acknowledges that he understands the task, and visibly establishes mutual belief on the task's initial conditions

Table 1. (*Continued*)

#	Human	Leonardo	Notes
16		Points to himself	He can do the first part of the task, and suggests doing so
17	"OK, you go"	Presses button one, looking at it	
18	At the same time as 17, presses button two		
19		Looks at button two; looks back at the human; nods shortly	Acknowledges partner's simultaneous action; creates mutual belief as to the task's completion

the reversal of the human's simultaneous action. Additional untrained user studies are currently being designed to quantitatively evaluate these perceived performance enhancements by comparing a functionally identical, but socially handicapped version of this system to our current implementation (i.e. the robot performs the task with social skills and cues verses without social skills and cues).

7. Discussion

In viewing human-robot interaction as fundamentally a collaborative process and designing robots that communicate using natural human social skills, we believe that robots will be intuitive for humans to interact with as well as better equipped to take advantage of our socially structured world to learn more effectively. Toward this goal, we have presented our ability to teach a task to a robot through the course of collaborative dialog, and the ability to coordinate joint intentions to perform the learned task collaboratively. In both collaboration and learning we have shown how we incorporate social acts that support collaborative dialog — the robot continually communicates its internal state to the human partner and maintains a mutual belief about the task at hand. This makes learning as well as working together more efficient and transparent.

In this section, we discuss our approach in the context of related work. Viewed in the context of a joint intention framework, our approach is significantly different than other approaches to human-robot interaction. Our goal is broader than interaction; we try to achieve *collaboration* between human and robot partners. Likewise, in the learning problem, we are interested in more than supervision (as traditionally defined in machine learning); we regard the human role in robot learning as that of guided exploration and *tutelage*.

7.1. *Collaboration versus interaction*

As discussed in Sec. 3, human-style cooperative behavior is an ongoing process of maintaining mutual beliefs, sharing relevant knowledge, coordinating action, and demonstrating commitment to doing one's own part, helping the other to do theirs, and completing the shared task. Using Joint Intention Theory as our theoretical framework, we have incorporated the notions of joint intentions and collaborative communication in our implementation. Our goal oriented task representation allows the robot to reason about the task on multiple levels, easily sharing the plan execution with a partner and adjusting to changes in the world state. The robot acts in accordance with joint intentions, and also works to communicate and establish mutual beliefs about the task state as the interaction progresses (e.g. confirming when a particular step is complete, and negotiating who will complete a portion of the task).

In related work, Kimura *et al.* explore human-robot collaboration with vision-based robotic arms.[37] While addressing many of the task representation and labor division aspects necessary for teamwork, it views the collaborative act as a planning problem, devoid of any social aspect. As such, it does not take advantage of the inherent human expertise in generating and understanding social acts. As a result, the interaction requires the human teammate to learn gestures and vocal utterances akin to programming commands.

Fong *et al.* consider a working partnership between human and robot in terms of *collaborative control*, where a human and a robot collaborate in vehicle teleoperation.[14] The robot maintains a model of the user, can take specific commands from the operator, and also has the ability to ask the human questions to resolve issues in the plan or perceptual ambiguities. The role of the human in the partnership is to serve as a reliable remote source of information. A similar approach has been taken by Woern and Laengle.[38] In contrast, our work explores collaboration where the human and robot work together on a collocated task where both the human and the robot can complete steps of the plan. Because the human and robot act upon a shared environment, the robot must therefore notice changes made by the human and dynamically reassess the plan and coordinate actions accordingly.

Some work in the field of human and virtual agent teams also has the notion of shared plans that must be continually maintained and updated according to changes in the world state. For instance, Traum *et al.* have a system in which a human is part of a team of agents that work together in a virtual world.[39] Their system addresses plan reassessment and uses dialog models and speech acts to negotiate a plan as a team. Roles are attached to various steps of the plan, and an authority structure helps in negotiating control. Our work differs in two respects from this virtual teamwork system. First, in our physically embodied scenario, we explore the issues of face-to-face gestures and socially relevant communication acts that facilitate collaboration. Second, we do not utilize an authority structure; instead, the robot and the human negotiate turns in the context of a shared plan.

In sum, on the one hand, previous works have dealt with the scenario of a robot being the tool towards a human's task goal, and on the other, the human being the tool in a robot's task goal. Our perspective is that of a balanced partnership where the human and robot maintain and work together on shared task goals. We have thus proposed a different notion of partnership than has been addressed in prior works:

that of an autonomous robot working with a human as a member of a collocated team to accomplish a shared task. In realizing this goal, we believe that robots must be able to cooperate with humans as capable partners and communicate with them intuitively. Developing robots with social skills and understanding is a critical step towards this goal. To provide a human teammate with the right assistance at the right time, a robot partner must not only recognize what the person is doing (i.e. his observable actions) but also understand the intentions or goals being enacted. This style of human-robot cooperation strongly motivates the development of robots that can infer and reason about the mental states of others, as well as communicate their own internal states clearly within the context of a shared interaction. Our goal-driven joint intention based framework is aimed at this promise.

7.2. *Tutelage versus supervision*

The dominant trend in machine learning has been to eschew built-in structure or *a priori* knowledge of the environment or task at hand, and set out to discover the structure that is in data or the world through exhaustive search and/or sophisticated statistical learning techniques. Two problem domains, pattern recognition and reinforcement learning in particular, have attracted attention and met with some success. Pattern recognition systems typically learn the mapping through a statistical analysis of hundreds or thousands of training examples chosen by a "knowledgeable external supervisor" in which the example contains both the input features and the desired output label. The main approach of reinforcement learning is to probabilistically, and exhaustively, explore states, actions and their outcomes to learn how to act in any given situation where the only supervisory signal is the reward received when it achieves the desired goal. However, as with supervised learning techniques, the actual learning algorithm has no *a priori* knowledge about the structure of the state and action spaces and must discover any structure that exists on its own through its exhaustive exploration of these spaces. As a result, reinforcement learning typically requires hundreds or thousands of examples, in order to learn successfully.

Thus, the progress to date in machine learning has come with some caveats. First, the most powerful techniques rely on the availability of data, and are not appropriate in domains with a small number of examples. Second, they are not appropriate when the environment is changing so quickly that earlier examples are no longer relevant. Third, the underlying representations used in machine learning typically make it difficult for the systems to generalize from learning one particular thing or strategy to another type of thing. Finally, little attention has been paid to the question of how a human can guide the learning process.

Our approach recognizes that for people, learning and teaching form a coupled system in which the learner and the teacher work together. Much of human instruction provides information through social cues and communicative acts that can be used by the learner to infer these constraints. For instance, the teacher often guides the child's search process by providing timely feedback, luring the child to perform desired behaviors, and controlling the environment so the appropriate cues are easy to attend to, thereby allowing the child to learn more effectively, appropriately, and flexibly. The teaching and learning process are intimately coupled, well tuned, and transparent. The teacher and learner read and respond to each other, to more effectively guide the learner's exploration. In this way, the adult becomes a more effective teacher and the child a more effective learner — each simplifies the task for the other.

We have posited throughout this paper that humanoid robots should utilize behaviors that are socially relevant to the humans with which they interact. We believe that the ability to utilize and leverage from human social skills is more than a good human-robot interface, but that it can positively impact the underlying learning mechanism.

7.2.1. *Teaching and learning*

In related work, Lauria *et al.* have introduced Instruction-Based Learning (IBL) where a human can instruct a robot in a navigation task through natural language.[40] In their system, the human uses phrases, which have been grounded to motor primitives in a corpus-based learning approach. The human can additionally teach the robot new phrases built from previously learned phrases. The instruction in their scenario takes place prior to execution, and the human is thus required to mentally represent the whole task prior to the start.

We take a different approach, and believe that the situated and tightly coupled aspect of learning and teaching is a key function of success for both the robot learner and the human teacher. In *learning by doing*, our robot makes its own assessment of the state-action space when building the task representation. Additionally, our guided experience approach simplifies the task of the teacher. By leading the robot through the desired task, adjustments and corrections can be made interactively at the appropriate times.

7.2.2. *Structure and feedback*

With many traditional supervised learning techniques, a human labels examples for the machine. Similarly, our approach has the machine learn through examples from the human. Within our dialog framework, however, the teacher interactively structures the task examples and thereby provides more relevant and seminal examples as learning progresses. In this way, the human works with the robot to actively narrow down the hypothesis space. In addition, our work illustrates how a human instructor can help frame a complex learning task into simpler components that can be learned individually and then combined to define full task specification. Importantly, this is done in a natural and intuitive manner for the human.

In our turn-taking framework, the teacher knows right away what problems or issues remain unclear for the robot, which allows for just-in-time error correction with specific refinement to a failed attempt. In contrast to a typical reinforcement learning algorithm (e.g. Q-Learning),[41] in which it can be ambiguous how far feedback should be propagated, the human in our scenario helps the robot build a correct representation of the failed attempt. Thus, our approach expedites learning by incorporating feedback *and* refinement.

In addition to efficient error-correction, Leonardo's ability to take turns with the human lends significant structure to the interaction that the robot can use to incrementally refine its performance.[29,30] The instructor leads, the learner performs, and if necessary the instructor provides more information, often exaggerating or focusing on aspects of the task that were not performed successfully. Nicolescu and Mataric also interleave instructive demonstrations (where a robot follows the human through a particular sequence of colored landmarks in a navigation course) with supervised practice trials of the robot.[42] Their robot learns a sequentially structured task (e.g. the order of colored landmarks to drive past) from human demonstration. The robot follows the human through the task, and the human can use short verbal commands to point out relevant information and frame the demonstration episodes. When the robot demonstrates the learned task, the human can stop the robot to correct the robot's task model. By saying the word "bad" the human instructor identifies which part of the robot's task model is incorrect, and the subsequent demonstration is used to correct the problem.

Whereas this work uses the guided experience approach, it does not model tutelage as a collaborative process. In their scenario, the human uses dialog to frame the instruction, but the robot does not give feedback, nor does it communicate its understanding until the execution stage, at which point there is still no communication of task certainty. We utilize a more tightly coupled turn-taking framework, in which social cues and gestures form a natural dialog between the teacher and the learner to express the robot's internal state (what is known and what remains unclear) and prompt the teacher for feedback or further instruction. By viewing learning as a collaborative process, communication plays a critical role in our approach. Leonardo maintains a mutual belief with the human teacher about the state of the learning process by continually communicating its internal state through demonstrations, expressive gestures, and eye gaze. Additionally, through gesture and eye gaze, the robot lets the teacher know when the current task representation has a low confidence, naturally prompting the teacher for feedback and further examples of the task.

7.2.3. *Learning for collaborative activity*

Our tutelage paradigm achieves a flexible task representation, which in turn is an important factor in Leonardo's ability to later cooperate on learned tasks. Our task representation supports learning the task on multiple levels: the structure of the actions and effects as well the generalized goal. The robot is able to learn tasks in which the specific actions are what matters as well as those in which a particular world state is the goal. As a result, the robot is more flexible in performing tasks —

it does not have to relearn a new policy when conditions change (the configuration of the task can change, or a partially completed state can be presented and the robot knows how to complete it). Thus, the robot not only learns the specific actions to perform, but also to achieve a desired result and then demonstrates commitment to the goal.

This flexible learning scheme is unlike other work in which a robot learns a new skill from a person. For example, imitative learning has been applied as an efficient way to explore the state-action space, using a human's demonstration to help initialize the robot's own search (see Schaal's work for a review).[43] The robot observes the human's performance, using both object and human movement information to estimate a control policy for the demonstrated motor skill. Providing the robot with *a priori* knowledge of the goal (in the form of an evaluation function) allows the robot to further improve its performance through trial and error, for instance, for a "ball-in-cup" task,[44] or hitting a tennis forehand.[45] In contrast, in our approach Leonardo is able to learn *new goals* instead of optimizing a given goal policy. The ability to flexibly learn new goals from a human partner is an important factor in robot learning.

8. Summary

This paper presents an overview of our work to build humanoid robots that can cooperate with people as capable partners and learn from natural human instruction. We argued that there are many reasons to believe that a social interaction will be the most natural way for ordinary people to work with humanoid robots and to teach them. Aiming at human-robot cooperation rather than mere human-robot interaction, we identified those key issues that must be addressed to build robots that can serve as helpful assistants and effective teammates. Of primary importance is building robots that understand people as people. In contrast to inanimate objects, human behavior arises from a rich network of mental states. Thus, humanoids will have to understand why people do what they do — inferring key mental states such as beliefs, desires, intentions, etc. from people's observable behavior — to be able to provide them with the right kind of help at the appropriate time, or to learn the intended thing at the right time. They must also display their own internal state clearly using channels that are intuitive for humans to read.

Toward this ambitious goal, we have presented how our ideas (informed by Joint Intention Theory and Situated Learning Theory) can be applied to building humanoid robots that communicate with people in human terms, work with people as capable partners, and learn quickly from natural human instruction. We have also highlighted a number of competencies developed for our humanoid robot, which have been applied to teaching the robot how to perform a simple task from natural human instruction. In contrast to standard approaches in machine learning that rely on many examples to infer task structure, our work leverages from social structure provided through natural human instruction to learn hierarchical tasks quickly and efficiently.

We have extended this approach to allow the robot to perform the learned task cooperatively with a human teammate. The robot works with the human to maintain a common ground from which joint intention, attention, and planning are derived. The robot is aware if its own limitations and can work with the human to dynamically divide up the task appropriately, offering to do certain steps or asking the human to perform those steps that it cannot do for itself. If the human proactively completes a portion of the task, the robot can track the overall progress of the overall task (by monitoring the state of the world and following the task model). Leonardo demonstrates this understanding using social cues such as glancing to notice the change in state the human just enacted, or nodding shortly towards the human. As a result they are both in agreement as to what has been accomplished so far and what remains to be completed. We believe that this approach will result in a natural cooperative environment that allows fluent collaboration between human and humanoid teammates, without having to constantly re-examine the robot's internal state.

Based on the work presented in this paper, we argue that building socially intelligent robots has extremely important implications for how we will be able to communicate with, work with, and teach robots in the future. These implications reach far beyond making robots appealing, entertaining, or easy to interact with. This is a critical competence for robots that will play a useful, rewarding, and long-term role in the daily lives of people — robots that will cooperate with as capable partners rather than needing to be operated (either manually or by explicit supervision) as a complicated tool.

Acknowledgments

The work presented in this paper is a result of the ongoing efforts of the graduate and undergraduate students of the MIT Media Lab Robotic Life Group and our collaborators. Leonardo's speech understanding abilities are developed in collaboration with Alan Schultz and his group at the Naval Research Lab. Stan Winston Studio provided the physical Leonardo robot. Leonardo's architecture is built on top of the C5M code base of the Synthetic Character Group at the MIT Media Lab. Geoff Beatty and Ryan Kavanaugh provided the virtual model and animations. This work is funded in part by a DARPA MARS grant and in part by the *Digital Life* and *Things that Think* consortia.

References

1. W. Bluethmann, R. Ambrose, M. Diftler, E. Huber, M. Goza, C. Lovchik and D. Magruder, Robonaut: A robot designed to work with humans in space, *Autonomous Robots* **14**, 179–207 (2003).
2. B. Grosz, Collaborative Systems, 1994 AAAI Presidential Address, *AI Magazine* **2**(17), 67–85 (1996).

3. P. Cohen and H. Levesque, Persistence, intention, and commitment, in *Intentions in Communication*, eds. P. R. Cohen, J. Morgan and M. E. Pollack (MIT Press, Cambridge, Massachusetts, 1990), Chap. 3.
4. P. Cohen and H. Levesque, Teamwork, *Nous* **25**, 487–512 (1991).
5. C. Breazeal, Social interactions in HRI: The robot view, *IEEE Transactions on Man, Cybernetics and Systems: Part C*, forthcoming (2003).
6. A. Billard, K. Dautenhahn and G. Hayes, Experiments on human-robot communication with robots, an imitative learning and communicating doll robot, in *Socially Situated Intelligence Workshop as part of the Fifth Conf. of the Simulation of Adaptive Behavior.* Centre for Policy Modelling technical report series number CPM-98-38 (1998).
7. K. Dautenhahn, Getting to know each other — Artificial social intelligence for autonomous robots, *Robotics and Autonomous Systems* **16**, 333–356 (1995).
8. A. Billard and K. Dautenhahn, Grounding communication in autonomous robots: An experimental study, *Robotics and Autonomous Systems* **24**(1 & 2), 71–81 (1998).
9. G. M. Hayes and J. Demiris. A robot controller using learning by imitation, In *International Symposium on Intelligent Robotic Systems*, pp. 198–204 (1994).
10. H. Friedrich and R. Dillmann, Robot programming based on a single demonstration and user intentions, in *3rd Euro. Workshop on Learning Robots (ECML95)* (1995).
11. M. Bratman, Shared Cooperative Activity, *The Philosophical Review* **101**(2), 327–341 (1992).
12. H. Jones and S. Rock, Dialog-based human-robot interaction for space construction teams, in *IEEE Aerospace Conference* (2002).
13. D. Perzanowski, A. Schultz, W. Adams, E. Marsh and M. Bugajska, Building a multimodal human-robot Interface, in *IEEE Intelligent Systems*, (2001), pp. 16–20.
14. T. Fong, C. Thorpe and C. Baur, Collaboration, Dialogue, and Human-Robot Interaction, in *Int. Symp. of Robotics Research* (2001).
15. Searle, J. Collective Intentions and Actions, in *Intentions in Communication*, eds. P. R. Cohen, J. Morgan and M. E. Pollack (MIT Press, Cambridge, Massachusetts, 1990), Chap. 19.
16. P. Cohen, H. Levesque, J. Nunes and S. Oviatt, Task-oriented dialog as a consequence of joint activity, in *1990 Pacific Rim International Conference on Artificial Intelligence*, Nagoya (Japan, 1990), pp. 203–208.
17. H. Levesque, P. Cohen and J. Nunes, On acting together, in *8th Nat. Conf. on Artificial Intelligence (AAAI-90)* (Boston, MA, 1990), pp. 94–99.
18. L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes* (Harvard University Press, Cambridge, MA, 1978).
19. H. Grice, Logic and Conversation, in *Syntax and Semantics 3: Speech Acts*, eds. P. Cole and J. L. Morgan (New York, Academic Press, 1975).
20. S. Baron-Cohen, Precursors to a theory of mind: Understanding attention in others, in *Natural Theories of Mind*, ed. A. Whiten (Blackwell, 1991), Chap. 16.
21. H. Clark and D. Wilkes-Gibbs, Referring as a collaborative process, *Cognition* **22**, 1–39 (1986).
22. B. Scassellati, Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot, in *Computation for Metaphors, Analogy and Agents*, ed. C. Nehaniv Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence* (Springer-Verlag, Berlin, 1998).
23. B. Scassellati, *Foundations for a Theory of Mind for a Humanoid Robot*, PhD Thesis (MIT Press, Cambridge, MA, 2000).

24. C. Breazeal, A. Edsinger, P. Fitzpatrick and B. Scassellati, Active vision systems for sociable robots, *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, **31**(5), 443–453 (2001).
25. C. Breazeal and B. Scassellati, A context dependent attention system for a social robot, in *Proc. 16th Int. Joint Con. on Artificial Intelligence (IJCAI99)* (Stockholm, Sweden, 1999), pp. 1146–1151.
26. H. Sacks, A. Schegloff and G. Jefferson, A simplest systematics for the organization of turn-taking in conversation, *Language* **50**, 696–735 (1974).
27. J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson and H. Tan, Human conversation as a system framework: Designing embodied conversational agents, in *Embodied Conversational Agents*, eds. J. Cassell, J. Sullivan, S. Prevost and E. Churchill (MIT Press, Cambridge MA, 2000).
28. J. Rickel and W. L. Johnson, Task-oriented collaboration with embodied agents in virtual worlds, in *Embodied Conversational Agents* eds. J. Cassell, J. Sullivan, S. Prevost and E. Churchill (MIT Press, Boston, 2000) pp. 95–122.
29. C. Breazeal, Proto-conversations with an anthropomorphic robot, in *9th IEEE Int. Workshop on Robot and Human Interactive Communication (Ro-Man2000)* (Osaka, Japan, 2000), pp. 328–333.
30. C. Breazeal, *Designing Sociable Robots* (MIT Press, 2002).
31. P. Bloom, Mindreading, communication, and the learning of the names for things, *Mind and Language* **17**, 37–54 (2002).
32. A. Woodward, J. Sommerville and J. Guajardo, How infants make sense of intentional action, in *Intention and Intentionality.* eds. F. Malle, L. Moses and D. Baldwin (MIT Press, Cambridge MA, 2001), pp. 149–169.
33. B. Gleissner, A. N. Meltzoff and H. Bekkering, Children's coding of human action: cognitive factors influencing imitation in 3-year-olds, *Developmental Science* **3**(4), 405–414 (2000).
34. D.A. Baldwin and J.A. Baird, Discerning intentions in dynamic human action, *Trends in Cognitive Sciences* **5**(4), 171–178 (2001).
35. B. Blumberg, *et al.*, Integrated learning for interactive synthetic characters. in *29th Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH '02)* (ACM Press: NewYork, 2002), pp. 417–426.
36. B. J. Grosz and C. L. Sidner, Plans for discourse, in *Intentions in Communication* eds. P. R. Cohen, J. Morgan and M. E. Pollack (MIT Press, Cambridge, MA, 1990), pp. 417–444.
37. H. Kimura, T. Horiuchi and K. Ikeuchi, Task-model based human robot cooperation using vision, in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'99)* (1999), pp. 701–706.
38. H. Woern and T. Laengle, Cooperation between human beings and robot systems in an industrial environment, in *M&R* Vol. 1 (2000), pp. 156–165.
39. D. Traum, J. Rickel, J. Gratch and S. Marsella, Negotiation over tasks in hybrid human-agent teams for simulation-based training, in *2nd Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems* (Melbourne, Australia, 2003), pp. 441–448.
40. S. Lauria, G. Bugmann, T. Kyriacou and E. Klein, Mobile robot programming using natural language, *Robotics and Autonomous Systems* **38**(3–4), 171–181 (2002).
41. C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning, *Machine Learning* **8**, 279–292 (1992).

42. M. Nicolescu and M. Mataric, Natural methods for robot task learning: Instructive demonstrations, generalization and practice, in *AAMAS 2003* (Melbourne, Australia, 2003).
43. S. Schaal, Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences* **3**, 233–242 (1999).
44. H. Miyamoto *et al.*, Learning robot based on bi-directional theory, *Neural Networks* **9**, 1181–1302 (1996).
45. H. Miyamoto and M. Kawato, A tennis serve and upswing learning robot based on bi-directional theory, *Neural Networks* **11**, 1131–1344 (1998).