Submitted for review to International Journal of Humanoid Robots
December 15, 2003

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 NE18-5FL, Cambridge, MA 02139
Email: cynthiab@media.mit.edu

Abstract

We want to build humanoid robots capable of rich social interactions with humans, including natural communication, cooperation, and tutelage. Humans display a remarkably flexible and rich array of social competencies, demonstrating the ability to interpret, predict, and respond appropriately to the behavior of others. 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. Developing robots with these social abilities is a critical step towards enabling them to be intelligent and capable in their interactions with humans, able to cooperate with people as capable partners, able to learn quickly and effectively from natural human instruction, are intuitive to communicate with, and are engaging for humans to interact with. 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. This paper presents an overview of our work towards building socially intelligent, cooperative humanoid robots that can work and learn in partnership with people.

1 Introduction

Many of the earliest motivations for developing humanoids centered on creating robots that can play a role in the daily lives of ordinary people. Our living spaces and the artifacts we use are engineered to support our human morphology. Accordingly, developing robots with a humanoid morphology would not require that we re-engineer our environment and artifacts to accommodate them. In addition, humanoid robots support human social cues and therefore would not require that people be specially trained to communicate with them. These arguments center on the issue of designing interfaces for robots that are well matched to the human environment and to people who are already experts in social interaction.

Beyond interface issues, however, there are other critical issues that concern how humanoid robots interact with us. For instance, their interactions with us must exhibit social appropriateness and adeptness. They must be able to *quickly* learn new skills and learn how to perform new tasks from human instruction and demonstration. Ideally,

helping a robot acquire new capabilities would be as easy and fast as teaching a person. Finally, to cooperate with humans as capable partners, robots need to understand our intentions, beliefs, desires, and goals so that they can provide us with well-timed, relevant assistance.

Yet robots today interact with us either as other objects in the environment, or at best 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 don't know how to appropriately adjust their behavior to help us as our goals and needs change. They do not flexibly draw their attention to what we currently find of interest so that their behavior can be coordinated and information can be focused 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. Consequently, they do not bring important information to our attention that is not easily accessible to us when we need it. They are not aware of our emotions, feelings, or attitudes. As a result they cannot prioritize what is the most important to do for us according to what pleases us or to what we find to be most urgent, relevant, or significant. Because of all these shortcomings (to name a few) they cannot cooperate with us as teammates or help us as assistants. At best, we can only use them as sophisticated tools rather than work with them in partnership.

1.1 Robots in the Daily Lives of People

This is a serious problem given that many of the most exciting and new applications for humanoid robots require them to cooperate with people as a capable, cooperative, and socially savvy partners. For instance, humanoid robots are being developed to provide the elderly with assistance in their home. Such a robot should be persuasive in ways that are sensitive to the person, such as helping to remind them when to take medication, without being annoying or upsetting. It must understand what the person's changing needs are and the urgency for satisfying them so that it can set appropriate priorities. It needs to understand when the person is distressed or in trouble so that it can get help. Ideally, the robot would play a useful and pleasing part in a person's life. People should enjoy having the robot in their life because they are useful and pleasant to have around.

In other applications, humanoid robots are being developed to serve as members of human-robot teams. NASA JSC's *Robonaut* is an excellent example (Bluethmann *et. al.* 2003). This robot is envisioned to serve as an astronaut's assistant to help its human counterparts maintain the space station using the same tools. To provide a human teammate with the right kinds of 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 within the context of the interaction they share.

For instance, consider the following collaborative task where a human and a humanoid robot work together shoulder-to-shoulder. The shared goal of the human and the robot is

to assemble a physical structure. Both have different capabilities---the human being more dexterous. The task requires different tools (e.g., screwdrivers, wrenches) and different equipment (e.g., beams, nuts, bolts). The workspace is partially occluded so that neither teammate can see the entire workspace. Due to the occluding barrier, each can see different parts of the workspace and therefore hold different beliefs regarding the placement of tools and equipment.

Given these constraints, the human's responsibility is to operate the tools necessary to assemble the structure. The robot's responsibility is to be a helpful assistant, providing the human with the appropriate tools at the right time, sharing relevant knowledge, and helping to maneuver the awkward pieces of the assembly into place so that they may be fastened together by the human. To be an effective assistant, the robot must be able to take multiple points of view to reconcile what the human sees and believes to be true with what it is able to see. For example, "Please pass me the wrench" may be ambiguous if the human can see only one wrench while the robotic assistant can see several. The correct response of the robot in this case would be to hand the person the wrench that is visible from the human's point of view. If the robot notices that an occluded part of the structure has not been fastened yet, the robot should realize that the human is not aware of this problem and consequently alert the person to the uncompleted part of the task.

Both must be in agreement as to the sequence of actions that will be required to assemble the structure so that the robot can manage the tools appropriately. If the robot must use some of the same tools to assemble parts of the structure in tandem with the human, it must carry out its task while being careful not to act in conflict with what the human teammate is trying to do (e.g., hoarding tools, assembling parts of the structure in a sequence that makes the human's task more difficult, etc.).

Hence, for the human-robot team to succeed, both must communicate to establish and maintain a set of shared beliefs and to coordinate their actions to execute the shared plan (Grosz 1996). In addition, each must demonstrate commitment in doing their own part, commitment to the other in doing theirs, and commitment to the success of the overall task (Cohen & Levesque, 1991; 1990).

2 Designing Robots for People

Does it make sense for robots to serve as cooperative partners for people, or should we be content to use them as tools? We argue that for many humanoid robot applications, people will naturally try to interact with robots in anthropomorphic, social terms. This is a natural fit for interacting with a robot as a partner.

When humans are faced with non-living things of sufficient complexity (i.e., when the observable behavior is not easily understood in terms of physics or its underlying mechanisms), we often apply a social model to explain, understand, and predict their behavior. We automatically attribute mental states (i.e., intents, beliefs, feelings, desires, etc.) to understand and explain its behavior (Reeves & Nass, 1996). People rely on social models (also called "folk psychology" by philosophers or "Theory of Mind" by

psychologists) to make complex behavior more familiar, understandable, predictable, and intuitive (Gordon, 1986; Premack & Woodruff, 1978). We do this because it is enjoyable for us, and it is often surprisingly quite useful (Dennett, 1987).

Therefore, from a design perspective, it makes sense to design humanoids to adhere to people's social model for them. By doing so, it will be possible for untrained people to naturally and intuitively explain and predict what the robot is about to do, its reasons for doing it, and how to elicit a desired behavior from it. As argued eloquently in Reeves & Nass (1996), humans are already experts at social interaction and it is a universally understood interface. According to Norman (2001), personality can be used as a powerful design tool for helping people form a conceptual model that channels beliefs, behavior, and intensions in a cohesive and consistent set of behaviors. Thus, designing robots with a life-like personality may help provide people with a good mental model for them. The parameters of the personality must fall within recognizable human (or animal) norms and be understandable and predictable to people, however, otherwise the robot may appear mentally ill or completely alien. Natural behavior is a very useful guide in this respect.

This raises an important question: to what extent does a robot's design support the social model? Simply stated, does applying a social mental model to understand and interact with the robot actually work? Furthermore, to what extent must the robot have its own social model to understand people? Many early examples of "social robots" (i.e. robot toys or entertainment robots) only project the surface appearance of possessing social and emotional intelligence. This may be acceptable for a sufficiently structured scenario (such as theme park entertainment, etc.) where the environment and the audience's interaction with the robot are highly constrained.

However, as the complexity of the environment and the interaction scenario increases, the social sophistication of the robot will clearly have to scale accordingly. Once the robot's behavior fails to support the social model a person has for it, the usefulness of the model breaks down. Ideally, the robot's observable behavior will continue to adhere to a person's social expectations during natural interactions in the full complexity of the human environment. Similarly, as the complexity of the robot's interactions with people increases, the robot will need better ways of understanding human behavior so that it can continue to behave appropriately. We argue that it will not be possible to achieve this degree of scalability without tackling the (hard) problem of developing "deep" architectures for socially and emotionally intelligent robots.

The remainder of this paper presents an overview of our approach for tackling the challenging research issues associated with endowing robots with social skills and social commonsense. This is a critical competence for robots to cooperatively work with and learn from people in partnership. We present specific examples of social skills and social understanding that we have developed for our highly expressive humanoid robot, Leonardo (shown in Figure 1), in the context of learning a task from natural human instruction.

3 Toward Robots as Cooperative Partners

For applications where robots interact with people as 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 (Bratman, 1992; Grosz, 1996). Much of the current work in human-robot interaction is thus aptly labeled given that the robot (or teams of robots) is often viewed as a tool capable of some autonomy that a remote human operator commands to carry out a task (Jones & Rock, 2002; Perzanowski *et. al.,* 2001; Fong *et. al*., 2001). This sort of master-slave arrangement where the robot is operated as a tool (albeit, perhaps using speech or gesture as a natural interface) does not capture the sense of partnership that we mean when we speak of working "jointly with" others as in the case of collaboration. Human-robot collaboration is an extremely important yet relatively unexplored scenario of human-robot teamwork.

Figure 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 with tactile sensation, an expressive (24 DoF) face capable of near human-level expression, an active binocular vision system (4 DoF), two actively steerable 3 DoF ears, a 4 DoF neck, with the remainder of the DoFs in the shoulders, waist, and hips. A speech recognition and understanding system (developed by and used in collaboration with the Naval Research Lab) allows the robot to engage in task-oriented dialogs with a human. 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.

3.1 Joint Intention Theory

What characteristics must a humanoid robot have to collaborate effectively with its human collaborator? To answer this, we look to insights provided by Joint Intention Theory (Cohen & Levesque, 1990; 1991). According to this theory, joint action is conceptualized as doing something together as a team where the teammates share the same goal and the same plan of execution. 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.

Communication plays an important role in coordinating their 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, the responsibilities of each teammate, etc.

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 hold up their end, as well as be committed to others' success in doing theirs (Grosz, 1996; Levesque *et. al*., 1994). Specifically, the actions and goals that each team member adopts to do their part should not interfere or prevent the others in carrying out theirs.

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 to 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 completing the shared task.

3.2 Toward Robots that Understand Other Minds

For robots to cooperate with people in a human-like way, they will need to be able to infer these kinds of mental states from people's observable behavior (e.g., gesture and social cues, language, actions, and more), the surrounding context, and internal models that they use to represent others. In humans, this competence is frequently referred to as a *theory of mind (ToM)* (Premack and Woodruff, 1978), *folk psychology* (Gordon, 1986), *mindreading* (Whiten & Byrne, 1997), or *social commonsense* (Meltzoff & Moore, 1997). It entails that each participant has a set of mechanisms and representations for predicting and interpreting other's actions, emotions, beliefs, desires, and other mental states; it plays a critical role in adult human-style communication and social understanding. It gives us the ability to take the perspective of others (Gopnick & Moore, 1994), share joint attention (Baron-Cohen, 1991), appreciate how perception relates to beliefs (Povinelli, 1996), experience empathy (Baldwin & Moses, 1994), engage in social referencing (Hornick *et. al.,* 1987), understand that others can have beliefs and desires that differ from our own (Wellman, 1991), reason about how these mental states lead to observable behavior (Astington & Gopnick, 1991), infer intentions and goals that underlie action (Baldwin & Baird, 2001; Woodward *et. al,* 2001), and more.

In a human-style social exchange, these capabilities are accomplished in part by each participant treating the other as a conspecific --- viewing the other as being "like me" (Meltzoff, 1996). Perceiving similarities between self and other is an important part of the ability to take the role or perspective of another, allowing people to relate to and to

empathize with their social partners (Heal, 2003). This sort of perspective shift may help us to predict and explain other's emotions, behaviors and mental states, and to formulate appropriate responses based on this understanding (Goldman, 2001). It enables each participant to infer the intent or goal enacted by the behavior of the other --- an important skill for enabling richly cooperative behavior.

The idea of having a robot adopt an experiential or "empathetic" understanding of others (rather than using a symbolic reasoning approach, for instance) has been explored in the context of having a robot learn a communication protocol from another robot (Dautenhahn, 1995; Billard & Dautenhahn, 1998) or a human model (Billard *et. al.,* 1998). In this style of work, the learner follows or mimics the model, where the model already knows the communication protocol and uses it to describe its own interactions as they unfold. As the learner follows or mimics the model, it learns to associate the communicated symbols with its own perceptual experience. In this way, the learner acquires the protocol with respect to its own embodiment through a process of sharing a similar experience with the model and assimilating it as its own.

It is very important for human-style cooperative behavior, however, that the robot *not* simply adopt the perspective or mental states of the other as its own. Although the robot can use its own cognitive and affective mechanisms as a "simulator" for inferring the other's internal states, it is critical that they be represented as distinct from the robot's own states. The reason being that the robot must be able to determine what is held in common, what is not, and therefore what must be communicated and agreed upon so that coordinated joint activity can be established and maintained. Hence, capturing this representational aspect of Theory of Mind of the robot's own states and the states of others is very important for our purposes.

4 A Collaborative Approach to Robot Learning

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 "knowledgable 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 a great deal of data. Thus, they are often not appropriate in domains when the number of examples may be very small. 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. Fourth, little attention has been paid to the question of how a human can guide the learning process. Fifth, and not insignificantly, few would argue that current approaches to machine learning, however successful, have much to tell us about how learning occurs in animals and humans.

By contrast, any survey of animal learning will quickly convince one that learning in nature is characterized by fast and robust, albeit, constrained learning. Indeed, an important way that internal structures simplify the learning task is by acting so as to bias the learner to take maximal advantage of external environmental and social interactions that serve to structure and constrain the learning task. Hence, learning is the result of a complex interplay of structures and processes, both internal and external to the learner.

For example, our approach recognizes that for people in the world, learning and teaching form a coupled system in which the learner and the teacher work together. For instance, humans are born with innate cognitive, social, behavioral, and affective machinery into a rich and benevolent learning environment, surrounded by adults and older children who are motivated and natural teachers. Much of human instruction and scaffolding 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. At its best, the teaching and learning process are intimately coupled, well tuned, and transparent. The teacher and learner read and respond to each other, tuning their behaviors 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. Ultimately, the teacher helps the learner to not only achieve certain learning outcomes, but also to become a better learner.

5 Overview: Architecture for Learning and Working in Partnership with People

The remainder of this paper outlines progress in a pedagogical scenario where our robot, Leonardo, learns how to perform a task from natural human instruction as part of a collaborative process as outlined in the previous section. Here, we present an overview of the perceptual, cognitive, social, and behavioral systems of the robot's computational architecture that allows the robot to work and learn in partnership with people. We focus on the perceptual and speech understanding system in this section. Sections 6 through 10 present the remaining systems in detail.

Similar to the cooperative assembly task example presented in section 1.1, Leonardo has to cooperate with the human to learn how to perform this task. This requires the robot to engage in several ToM competencies, including the ability to establish joint attention about external objects, the ability to represent belief states of self and other, the ability to take the visual perspective of others, and the ability to align these mental states between the human and the robot through communicative acts. Section 6 presents progress towards developing the subsystems that implement these ToM competencies in more detail.

Figure 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.

In Leonardo's "Button Task," there are a number of buttons in front of the robot (see Figure 2). The human instructor stands facing Leonardo, across from the buttons, to teach the robot to how to perform the button task using natural social cues (speech, gesture, head pose, etc). All of the buttons start in the OFF position and the task is to first turn them all ON and then turn them all OFF (see section 9). In route to learning a high-level representation for the overall task, the human has to teach Leonardo the labels for each of the buttons (see section 8), as well as the motor skill for how to press a button (see section 7). After the robot has learned the task, it is able to perform it cooperatively with a person (see section 10).

5.1 Overview of the Computational Architecture

A schematic of Leonardo's cognitive architecture is shown in Figure 3. This section presents a brief summary of the speech understanding system (section 5.2) and the vision system (section 5.3). Section 6 presents those systems (i.e., the *attention system*, the *spatial reasoning system*, the *belief system*, and parts of the *action system*) that implement a variety of communication skills (e.g., turn-taking, gestures, deictic reference) and other abilities associated with Theory of Mind competencies (e.g., joint attention, modeling beliefs, and perspective taking). Section 7 describes the *motor system* and its role in learning motor skills (such as how to press the buttons) from internal demonstration via a telemetry suit worn by the demonstrator. Section 8 presents an overview of how these systems interact to allow the robot to learn names of objects and follow requests. Section 9 describes how the *learning system* interacts with other aspects of the cognitive

architecture to enable the robot to acquire a high-level task model for the Button Task from natural human instruction. Section 9 describes how this task model is used to allow the robot to perform the learned task cooperatively with a human.

5.2 The Speech Understanding System

We have been working in collaboration with Alan Schultz and his group at the Navy Research Lab to extend their natural language understanding system (Nautilus) to support collaborative task-oriented dialogs and accompanying communicative and expressive gestures with humanoid robots. The current *Nautilus* speech understanding 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.) --- see (Perzanowski *et. al.,* 2001). The vocabulary has been tailored to support the kinds of actions (grasping, pressing, look-at, etc.), entities (buttons, people, etc.), features (color, button-ON, button-OFF, shape, size, etc.), and gestures (pointing, head nods, etc.) that Leonardo perceives during his interactions with objects and people. As described in sections 6 through 10, we have been developing perceptual, cognitive, and learning systems to support these dialogs.

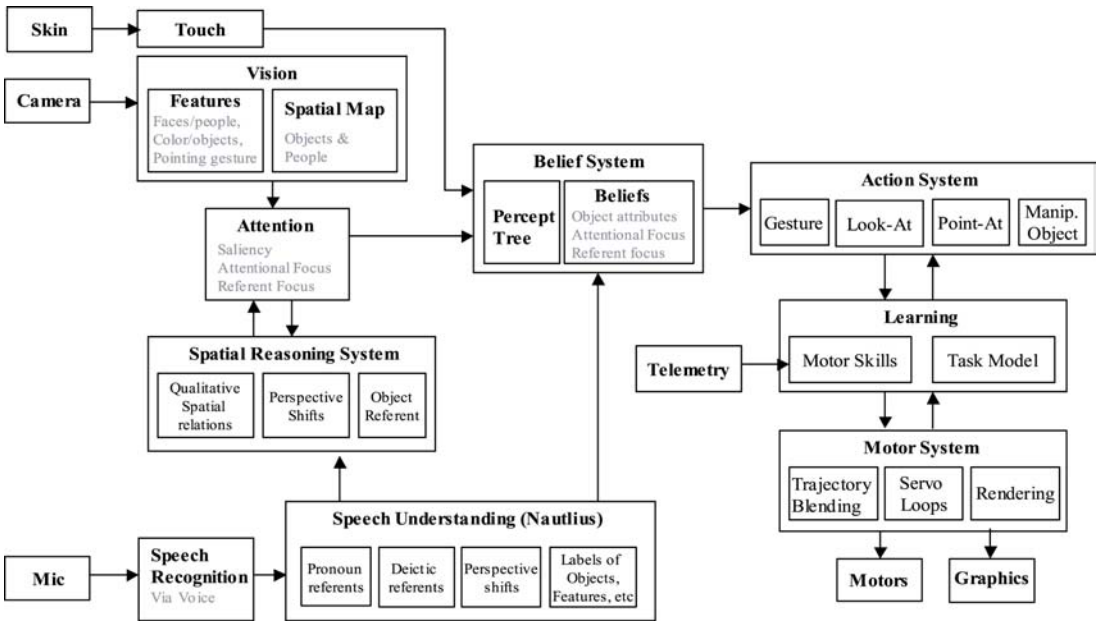


Figure 3: Leonardo's cognitive architecture for learning and performing tasks and motor skills (see text). It is built on top of the *c5* codebase; a behavior engine originally designed for computer animated interactive characters (Burke et. al., 2001). We use the graphic abilities as a simulator for the robot.

5.3 The Vision System

Leonardo perceives the surrounding environment with three camera systems. Two of the stereo camera systems are presented in Figure 4. 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 (e.g., the buttons based on their shape, size, color, and position) in the workspace in front of Leonardo. This visual information is normalized to real-world coordinates and calibrated to Leonardo's frame of reference. The workspace defines the position and angle of the robot base with respect to the camera system.

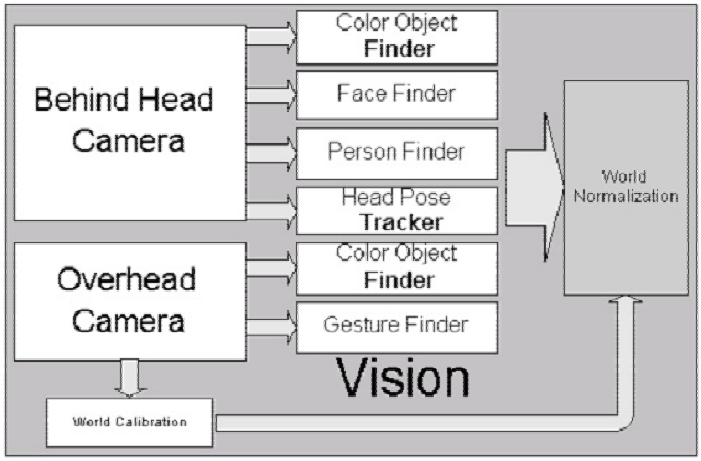


Figure 4: A schematic of Leonardo's vision system. Two stereo camera systems are shown in this diagram with extracted features (calibrated to world coordinates) that allow Leonardo to visually characterize people and objects.

Each button is detected and tracked via saturated color matching on the intensity data. The RGB video data is first converted to hue, saturation, and value (HSV) space, and the saturation values are then thresholded. Pixels that pass the threshold are windowed by hue to isolate red (indicative of buttons) and green (indicative of special calibration markers that define the distal corners of the robot's workspace). The calibration markers need only be present upon initiation of the system and can then be removed, as neither the camera nor the robot's base are expected to move during an interaction.
Once the appropriately saturated pixels have been extracted, a pixel-labeling algorithm is run over the results. Connected and near-connected pixels are given the same label, and the extents of a particular label are computed and defined as a unique object (a button). For button objects, the center pixels of the object are examined to determine whether they match the exterior pixels. There is a colored LED at the center of each button whose lighting is toggled when the button is pressed. This enables the robot to have visual feedback of when a button has been turned ON. The locations of calibration objects are not required after the initial workspace calibration, but the number, positions and pressed status of the buttons are sent out over the robot's computational network on every frame.

Leonardo also has two narrow field of view cameras, one in each eye, used for post-attentive processing such as face recognition and tracking of facial features (using the Axiom ffT SDK from Nevengineering, Inc.).

6 Social Skills for Collaborative Interaction and Learning

As discussed in section 4, human-style tutelage is fundamentally a social and a collaborative process. Breazeal (2003) outlines a number of social cues that people employ to naturally structure and facilitate the learning process for others, 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 section 11). 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.

To be a good instructor, the human must maintain an accurate mental model of the learner's state (e.g., what is understood so far, what remains confusing or unknown, etc.). This allows him or her to appropriately structure the learning task and to provide timely feedback and guidance. In short, the instructor relies upon this mental model to tailor his or her instruction to the needs of the learner. The learner (robot or otherwise) can help the instructor keep an accurate mental model by communicating its state to the instructor via communicative acts (e.g., expressions, gestures, or vocalizations that reveal the learner's understanding, confusion, attention, and so forth).

Hence, through reciprocal interaction, both the learner and instructor cooperate to 1) help the instructor maintain a good mental model of the learner, and 2) help the learner leverage from instruction and guidance to build the appropriate task models, representations, associations, etc. The remainder of this paper describes the social capabilities that allow Leonardo to participate in this cooperative and social process to learn the Button task from natural human instruction.

6.1 Communication Skills

6.1.1 Conversational Policies

We have implemented a suite of a collaborative task-oriented conversation and gestural policies for Leonardo. Dialog is fundamentally a cooperative (Grice, 1975). Cohen *et. al.* (1990) argue that much of task-oriented dialog can be understood in terms of Joint Intention Theory (see section 3.1). 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.,* (1990) has analyzed task dialogs where an expert instructs a novice on how to assemble a physical device. 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 it shrugs his shoulders with an expression of confusion to request clarification or elaboration from the human instructor. The robot looks to the button 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 button. The robot can demonstrate its knowledge of the button names that it has been taught by pointing to the correct button in response to the human's query "Which is button 1?" This confirms that both human and robot share the same belief regarding which the button is called by what name.

6.1.2 Turn Taking Skills

We have supplemented our models of collaborative dialog and gesture with flexible turn-taking skills modeled after those used by humans (Sacks *et. al.,* 1974). 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 para-linguistic social cues, called *envelope displays*, to manage who is to talk at which times in an intricate system of turn taking (Sacks *et. al.*, 1974). 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 (Cassell *et. al,.* 2000; Rickel & Johnson, 2000) as well as expressive robots (Breazeal, 2000; 2002).

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 non-verbal) are frequently exchanged to let the conversants know that each is being properly understood by 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.

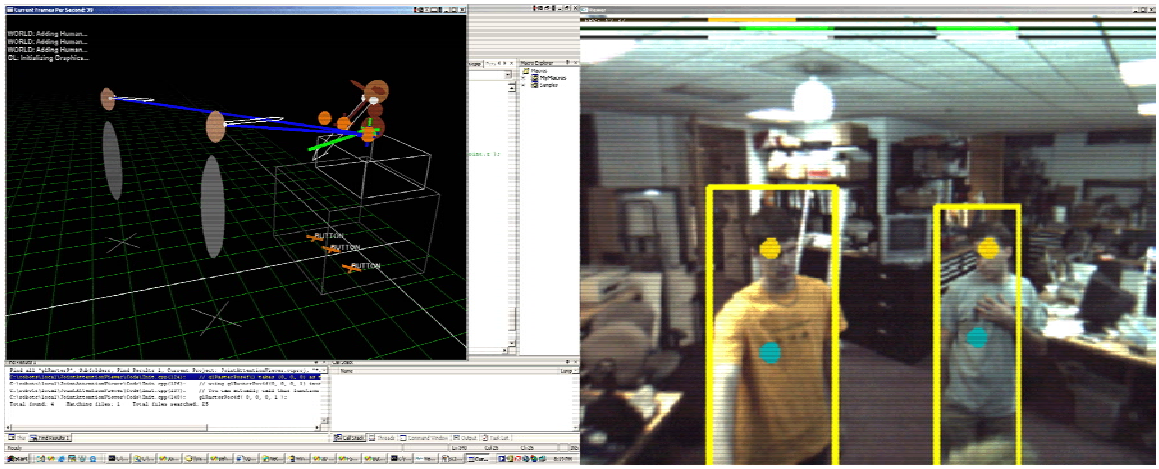


Figure 5: The vision system can detect head pose and pointing gestures. In this figure, two people are looking at the button to the left of Leonardo (blue vectors). The leftmost person is also pointing to that button (green vector). The robot happens to be looking at the center button.

6.2 Deictic Reference

We have implemented a suite of deictic gestures used by humans to refer to objects and to direct the attention of others to important aspects of the shared context. For instance, pointing gestures and following a person's direction of gaze (which we estimate using head pose) allow people to establish joint attention with others (see Figure 5). We have implemented visual routines for recognizing pointing gestures (see Figure 6). In addition, we have developed a suite spatial reasoning routines that allow the robot to geometrically follow a human's head pose (section 6.2.1) or pointing gesture (section 6.2.2) to the indicated object referent (section 6.2.3).

6.2.1 Following Head Pose

To track a person's head pose we are using the head tracking system developed by the CSAIL Vision Interface Group (Morency et. al., 2002). This software system uses the wall mounted stereo camera to provide 3D head pose estimation (pan, tilt, and rotation) in

real-time. This provides the attention system with a vector that characterizes the direction of the person's gaze.

To calculate which object a person is focusing on, each of the observed objects in the visual scene (from the 3D spatial map) is projected onto this vector. If the object with the least error from the gaze line is within a cone angle of 16 degrees (the cone defines the visual field of focus of the person) then that object is taken to be the focus of attention for the human provided that the person maintains fixation on that object for a minimum length of time (a couple of seconds).

The focus of attention cone is defined by:

If (C(E/D) <= 1.0) then Within focus*
If (C(E/D) > 1.0) then Outside visual focus*

Where:
C is the angle constant for 16 degree wide cone of focus.
E is the error in the projection of the object location onto the gaze vector.
D is the distance along the vector of the projected object.

6.2.2 Detecting Pointing Gestures

We use the overhead stereo camera to detect when the person is performing a pointing gesture. Detection of the arm is accomplished by background subtraction in two visual domains: intensity and a stereo range map.

Intensity background subtraction is a common technique for distinguishing the salient features of an environment based on the fact that items move, but may not be continuously in motion. Our technique uses a fairly standard unimodal Gaussian model where the mean value and standard deviation of the intensity of each pixel is computed over a number of video frames. Once this training has completed, every pixel on each frame is compared with its mean value, and the difference is thresholded to produce a foreground image. The threshold value of 3σ is typical. More sophisticated statistical models, such as bimodal distributions (Haritaoglu *et. al.,* 2000), can be used (for example, to account for image variation of a periodic nature) but we have not found this necessary, as a result of our use of stereo processing (Eveland *et. al.,* 1998; Ivanov *et. al.,* 2000) to account for illumination effects such as shadows and false positives present in the intensity map. Our background models are continuously updated with a two-element IIR lowpass filter.

To compute the foreground depth map, first a real-time depth map is produced using off-the-shelf hardware and correlation software for fixed-baseline parallel cameras (Konolige, 1997; Faugeras *et. al.,* 1993). The stereo depth map is then filtered to account for correlation noise, depth discontinuities, and insufficient texture (Darrell *et. al.*, 2001). To compute the foreground depth map, background subtraction is performed on this depth image using the same algorithm as in the intensity domain.

The master detection image is computed by performing a logical AND operation on the intensity foreground and depth foreground maps. The foreground depth image is more robust to illumination effects, but whereas the intensity foreground tends to suffer from false positives, the stereo foreground more commonly suffers from undefined areas due to correlation noise at depth discontinuities and patches of insufficient texture for stereo matching. We therefore perform complementary morphological cleaning operations on each before combining them. Multiple pre-detection techniques of this type are increasingly popular in human segmentation for gesture extraction as processing power becomes available (e.g. motion and skin color (Habili *et. al,* 2001)).

To extract the extended arm from the master image, separate regions in the master detection image are extracted via a multi-pass labeling and extents collection algorithm. Only the first pass operates at the pixel level, so on a computational cost basis it is comparable to a single-pass approach. The result regions are then sorted in decreasing size order, and compared against a region history for one final accumulation stage, to combat any breakup of the segmented body part. The largest candidate regions are then fit with a bounding ellipse from image moments within each region, and evaluated for likelihood of correspondence to an arm based on orientation and aspect ratio. The best candidate passing these tests is designated to be the pointing arm and used to compute the gross arm orientation.

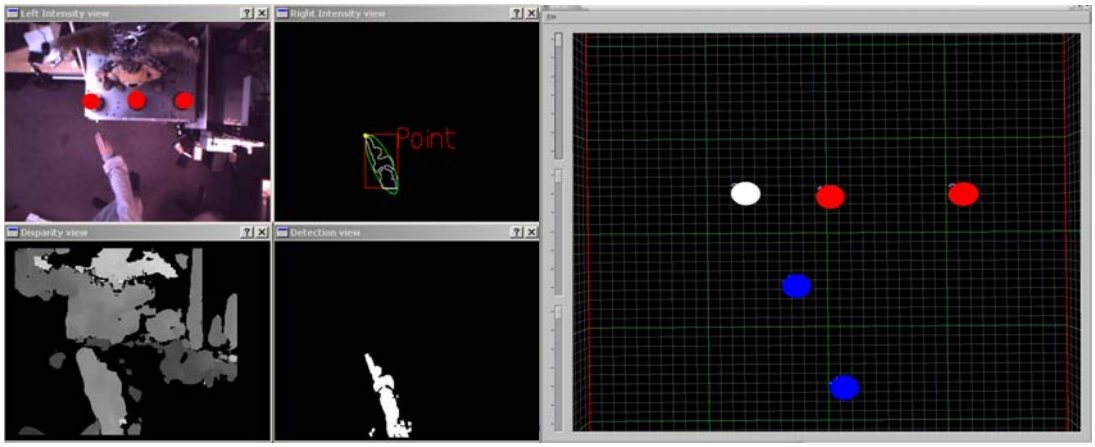


Figure 6: Computing the deictic reference to an object in the visual scene. Left, an overhead stereo camera identifies the locations of the buttons in the scene and recognizes when a pointing gesture occurs, estimating the location of tip of the finger and the angle of the forearm. This is passed to the spatial reasoning system (right). This overhead viewpoint shows the buttons (red), the location of the tip of the finger and base of the forarm (blue), and the identified object referent (white).

Once the arm has been extracted, we recognize whether the hand is configured in a pointing gesture or not (see Figure 6). We accomplish this by estimating the kurtosis, or "pointiness", of the hand. The distance from each arm result pixel to the arm centroid is computed, and the most distal point identified in the direction of the gross orientation. This point is assumed to be the hand tip. The outline of the hand is then traced by performing a morphological erosion on the detection image and then reapplying this

result to the master detection image with a logical XOR operation.
The hand outline is treated as a 1D bidirectional array of 2D features to be traversed in both directions, with the deviation between each pair of points summed. This result is compared against empirical results for various hand configurations, and the 2D general traverse direction is compared with the gross arm orientation. Since the video frame rate is fast enough that a small amount of additional latency is acceptable, and it is not necessary to be able to reliably detect unusual pointing gestures that last less than a fraction of a second, several adjacent video frames vote on whether or not a pointing gesture has been detected.

6.2.3 Determining the Object Referent from Pointing

Humans are able to fluidly switch between multiple points of view to perform spatial reasoning from different perspectives (Franklin *et. al*., 1992). We have developed a basic 3D spatial reasoning system to handle issues regarding spatial relations and reasoning about objects from a different person's point of view (i.e., from the robot's visual perspective or from an adjacent person's viewing position). To take the visual perspective from a different point of view, the system can perform rotations of the robot's 3D spatial model of its environment and the people within it. This allows the system to compute spatial relationships from different vantage points. In addition, the spatial reasoning system supports the spatial relations capabilities of the Nautilus speech understanding system. Hence, it can compute basic spatial relations such as front/back, right/left, and above/below with reference to a particular location from a specific visual vantage point.

Given this ability to compute perspective shifts and spatial relations, the *spatial reasoning system* is responsible for computing the correct object referent in visual the scene from the button locations, gross arm orientation, and hand configuration (with a confidence measure). To determine what object the human is pointing to, the spatial reasoning system first performs a perspective shift to place the point of view on the tip of the person's finger. It then computes the FRONT-OF spatial relationship between the fingertip and objects that are in front of it within a bounded range (see the rightmost image of Figure 6).

Ultimately, the ability to take the visual perspective of another person is an important capability when a workspace is partially occluded so that the robot human each see different partial views of the scene (as discussed in section 1.1). To be able to communicate relevant information to each other, each must appreciate what the other can see and therefore what each knows and believes about the situation (e.g., I know that there are three objects in the scene because I can see them. However, my teammate must believe that there are only two since that's all they can see from their vantage point). This competence requires integrating the spatial reasoning system with the systems that handle joint attention (section 6.3) and mutual beliefs (section 6.4).

6.3 Joint Attention

Leonardo's Button Task requires that the robot understand and participate in shared attention with a human in order to learn the names of objects and actions, follow requests, learn tasks from instruction, and perform tasks cooperatively with people. Therefore, the robot must be able to interpret and respond appropriately to human attentional cues (such as pointing and referential looking). Leonardo must also be able to use pointing gestures and other deictic cues to direct the attention of its human partner.

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 (Baron-Cohen, 1991). Therefore it is a critical capability for Leonardo. According to Baron-Cohen, understanding *seeing* verses understanding *attending* hinges on the latter requiring that the representation of the object encode something perceived as being "interesting" to the viewer. Normal human children first begin to understand attention (i.e., that vision can be directed selectively to objects or events of interest) between 7 to 9 months of age (Baron-Cohen, 1991). In addition, there are a number of skills that allow children to participate in sharing attention with others. For instance, *referential looking* (i.e., the ability to look to where another is looking) develops from 6 to18 months of age (Butterworth, 1991). Another skill is *proto-declarative pointing* --- the ability to point in order to comment or remark on the world to another person. This ability requires the child to explicitly represent the mental state of attention of another person. Infants are first able to direct their gaze to where another person points at around 9 to 12 months of age, and they begin to point in order to direct the gaze of others at approximately 14 months of age.

In human-human communication, it is important to note that the act of "referring to" is fundamentally a collaborative process (Clark & Wilkes-Gibbs, 1986). The goal is to establish a state of mutual belief (recall section 3 and see section 6.4 for our implementation) among those sharing joint attention. Having both human and robot look at the same thing is important part of this process, but is not sufficient alone to share joint attention (as defined above). To date, robots that do respond to deictic gestures and other attention direction cues can look at the same object as a person, and therefore engage in referential looking (e.g., Scassellati 1998; 2000; Breazeal *et. al*., 2001). However, they do not explicitly represent the attentional state and resulting beliefs of the human collaborator. In this sense, such robots are more like primates than humans---primates can follow gaze but do not seem to relate this to the viewer's state of attention or consequent beliefs (Povinelli, 1996; Baron-Cohen, 1991; Baldwin & Moses, 1994). In making a reference, however, 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.

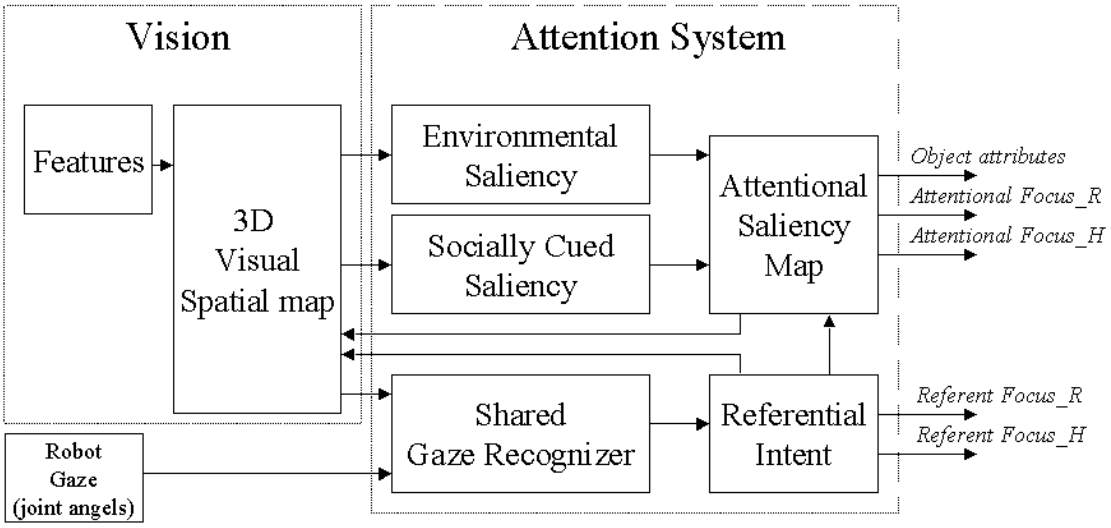


Figure 7: Schematic of the robot's visual attention system for external factors. It enables the robot to share joint attention with people. The saliency based system deals with environmental aspects of saliency (color, movement, etc.). The socially based system deals with social cues that direct attention (speech, head pose, pointing, etc.). Internal factors come from the robot's cognitive system (not shown).

One critical aspect of mutual belief is to establish a common referential focus. The distinction between *attentional focus* (what one is currently attending to) verses the *referential intent* (or *referent focus*) the object or event the communication is *about*) matters a great deal (Baron-Cohen, 1991; Baldwin & Moses, 1994). For instance, a human might direct the robot's attention so that they both look at a particular object to *share* attention and establish the referential focus. As the human continues to discuss the object with the robot, the robot's attentional focus (and therefore what it looks at) will shift. For instance, the robot may attend to the person's face to assess his or her emotional reaction to the object, look back to the object for more detailed visual processing, and so forth. However, despite these shifts in the robot's attentional focus, it must understand that the information being conveyed is attributed to the object that is the referential focus---not to whatever the robot might happen to look at.

To address this issue, 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.

6.3.1 Model of Robot's Attention

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 Figure 7). 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.), and *socially directed reference* (pointing to, 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

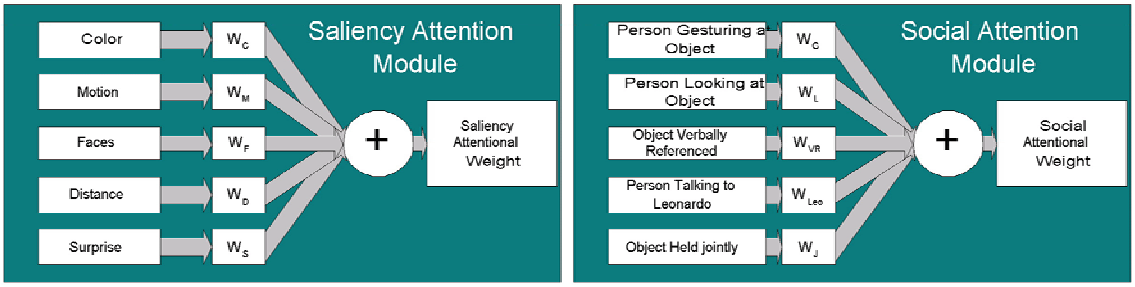


Figure 8: The weighted factors that are considered when determining the overall saliency of environmental (left) and social cues (right) that direct the robot's attention.

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 Breazeal & Scassellati (1999). The left diagram of Figure 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 (Breazeal et. al., 2001). The *referential focus, RefFocus_R* is determined as the last object that was the subject of shared attention between robot and human.

6.3.2 Model of Human Attention

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 Figure 9 where robot and human are sharing joint visual attention via head pose and pointing gesture.

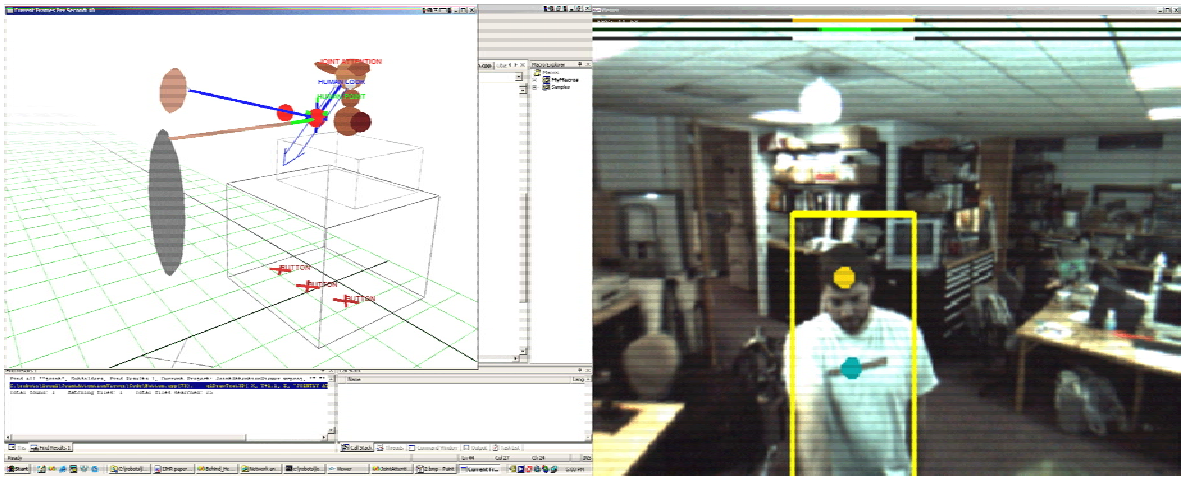


Figure 9: Visualizer showing the robot and a human sharing joint visual attention on the same object. The right image shows the visual input of a person looking at and pointing to the center button. The left image shows the visualization of the robot's internal model. The human's gaze is shown as the 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.

These attention following and directing skills can be accompanied by conversational policies along with gestures and shifts of gaze for repair, elaboration, and confirmation to confirm a shared referential focus and to maintain mutual beliefs between human and robot (as discussed in section 6.1). 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 as discussed in section 1.1. 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.

6.4 Mutual Beliefs

This leads us to the issue of modeling and reasoning about beliefs (it is possible that the human and robot may hold different beliefs due to different visual perspectives). Young children around 3 years of age appreciate the relationship between perception and belief. Whereas *seeing* reflects the state of the world as it is directly perceived, *beliefs* are representational and are held to be true even if they do not happen to agree with immediate perceptual experience (Wellman, 1991). For instance, a child may believe that a hidden object continues to be present even though it cannot be seen at that moment. For collaboration, it is important that Leonardo be able to establish mutual beliefs between itself and its human partner. This entails that the robot be able to represent beliefs for itself as well as model the beliefs of others.

6.4.1 Perceptions and Beliefs

Leonardo's belief system can best be described as a mechanism for performing temporal integration of perceptual input, resulting in coherent sets of objects that are believed to exist in the world and their affiliated properties. It is represented as a dynamic database of "belief" objects, each corresponding to an external object in the world (Burke *et. al,*

2001). For instance, Leonardo's buttons are each represented as the associated features of color, shape, position, and a linguistic label (that as assigned when a person names the button for the robot).

Internally, the belief system is defined by pointers to the percepts feeding into it and rules that govern the merging of new perceptions into the existing set of beliefs. Percepts may correspond to a multitude of sensory data: vision, tactile information, speech commands, audio inputs, and proprioceptive feedback. Whenever a new piece of sensory input arrives in the perception system, it activates a hierarchical structure (called the *percept tree*) that classifies the sensory information into a data structure called a "snapshot" which is then passed over to the belief system. Upon reception of this "snapshot", the belief system applies a set of sequential merging and arbitration rules to either create new beliefs from the sensory information, or to update existing beliefs about objects Leonardo already knows about.

For instance, when a human wishes to teach Leo the name of an object, he or she will point to it and name it for the robot (e.g., "Leo, this is Button 1"). Upon recognizing this gesture, shared attention shifts to the indicated button, and the belief for that button is updated to denote that it is the current attentional focus and the current referent focus of the robot and human. The name for the button is extracted from the speech stream and associated with this belief given that it is the current referent focus. From this point onward, even when the button is physically moved, its location attribute is updated and the name moves with it since they are part of the same belief. Moreover, if a person requests that the robot press that button (e.g. "Leo press Button 1") the belief for that button is queried for its current location so that Leonardo knows where to look for it and upon what object to apply its button-pressing action.

6.4.2 Model of Human Beliefs

Leonardo models the human's beliefs (and assigns the attentional focus and referent focus attributes) using the same database representation and feature association mechanisms. In this way, the robot can compare those beliefs it holds personally with those that it attributes to the human. To update its model of the human's beliefs, the robot follows the person's gaze to nearby objects to update the human's attentional focus attribute for beliefs associated with those objects. Those beliefs that are shared also with the robot are labeled as mutually held, $MutBel$. The robot updates the human's referent focus attribute by using its joint attention mechanism.

7 Learning Motor Skills from "Internal" Demonstration

Leonardo learns how to press buttons from "internal" demonstration by a human operator. In this scenario, the human demonstrator wears a telemetry suit that allows her to control Leonardo's joint angle movements. As she "shows" Leonardo how to perform an action by guiding the robot using the telemetry suit, the robot records these specific actions as they are applied to objects at specific locations. The motion capture suit measures joint angles at over 40 different places of the operator's hips, torso, arms and

neck using potentiometers and gyros at a frame rate of up to 120 hertz. The motion capture software associated with the suit generates accurate 3D models of the human in Euler angles. We convert these suit angle measurements in real-time into the equivalent joint angles, carefully calibrated to the robot, so that the operator can command Leonardo's motors in a natural fashion.

The human demonstrator can "show" Leonardo how to press a button at several different locations in its workspace. This defines the basis set of button-pressing examples that are indexed according to 2D button location. The robot can then interpolate these exemplars (see equation 1) using a dynamically weighted blend of the recorded button pressing trajectories (based on the Verb & Adverb animation blending technique of Bodenheimer *et. al.* (1998).

Equation 1

For each joint angle J_k in the robot:

$$J_k = \sum_{i=1}^{NumExemplars} E_{k,i} \times W_i$$

Where $E_{k,i}$ is the kth joint angle in the ith exemplar, and W_i is the weight of the ith exemplar

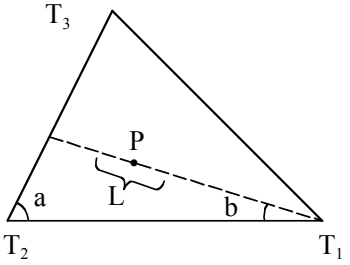
To determine the blend weights, we first precompute the Delaunay triangulation of the target points (also known as the dual of the Voronoi diagram). Then we find the triangle of targets that encloses the new location, and calculate the three weights such that a weighted sum of those targets is equal to the position of the new button location (see equation 2). Once the weights are determined, we can blend these three source animations together according to the calculated weights on a per joint basis.

Equation 2

To calculate blend weights for point P, find weights W_1, W_2, and W_3 such that:

$$P = T_1W_1 + T_2W_2 + T_3W_3$$

We solved this problem geometrically:



$$W_1 = \frac{\left|\overline{L}\right| - \left|\overline{T_1P}\right|}{\left|\overline{L}\right|}$$

Where:

$$\left|\overline{L}\right| = \left|\overline{T_1T_2}\right| \sin(a) / \sin(\pi - a - b)$$

W_2 and W_3 can be obtained similarl

This process is done for each frame of the resulting movement trajectory; thus for each frame each joint angle is computed using a weighted sum of the joint angles from all of the motion-captured source trajectories for that frame. While this type of computation

can result in an end effector position that is not linearly related to the blend weights used, we have found that approximating this relationship as linear has been sufficient for this case. Using this technique, Leonardo does reasonably well at pressing a button located anywhere in its workspace.

8 Learning Names of Things and Following Requests

With the capabilities outlined in the previous sections, Leonardo is able to learn the names of objects from natural human instruction. For the human, teaching Leonardo the names of objects is straight forward---one simply has to point to the desired object and name it for the robot, e.g. “Leonardo, this is a hammer.” 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., “hammer” to the robot’s belief system. In parallel, the vision and spatial reasoning systems pass the visual features of the object referent to the belief system as well. This allows the belief system to bind 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 then looks back the person’s face when he has formed the corresponding belief.

Alternatively, if the robot did not understand the utterance, he communicates this to the human as well. Immediately upon hearing a person say something, 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 confusion (cocks his head to the side and shrugs). This allows the person to correct the misunderstanding. This 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.

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 hammer.” If Leonardo has assigned the “hammer” 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 “hammer” label to an object (perhaps the person did not label anything a “hammer” yet, or perhaps the robot didn’t understand when she did so) the robot communicates that it doesn’t 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 object. Otherwise, if not, Leo will disappointedly shake its head “No…” while shrugging to you show me the hammer?” In this case the robot will enthusiastically nod its head “yes” and then point to the corresponding object if Leo has assigned a “hammer” label to an communicate, “…because I don’t know.”

Therefore, by using natural social cues as feedback, the human instructor can quickly assess what the robot has learned and what it hasn't learned so far, and immediately correct any misunderstandings if needed before moving on to teaching the robot 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 wrench." To respond, the robot directs its gaze from the humans' face to the desired object. After a short period, the robot looks back to the speaker to await her next utterance. Alternatively, the human can direct the robot to operate on an object, such as "Leonardo, press button 1." The robot responds by applying the desired action to the specified object. Video of a sample scenario can be viewed at http://robotic.media.mit.edu/projects/Leonardo/Leo-demo-movie.html.

9 Learning Task Structure from Collaborative Interaction

In addition to requesting that Leonardo perform actions with his buttons, we are able to teach Leonardo more complex tasks made up of these actions. Tasks are represented in a hierarchical structure of actions and sub-tasks. Completing a task involves completing each of its child actions or tasks (unless higher-ranking goals have been achieved otherwise), which are then recursively expanded to their actions. The task representation encodes constraints among the actions, which are currently utilized only as sequential constraints, but the representation is generic and others could be added in the future.

In the learning of a task, a goal is associated with each of the constituent actions as well as the task as a whole. Therefore, the task goal is more than just the conjunction of the goals of its actions and sub-tasks. Additionally, in executing the task, the task goal can be evaluated as a whole rather than evaluating each of its children's goals to determine if the task is done, improving efficiency. We found the goal driven approach crucial for both the tutoring of tasks and collaborating on them. Goals provide a common ground for action segmentation and understanding as well as for coordinating actions as part of a collaborative activity.

Leo encodes the goal state associated with an action performed in the tutorial setting by comparing the world state before and after its execution. The system currently has two different types of task goals. Goals that represent a state change in the world, and those in which the goal is simply to perform the activity. One key difference between the two is that in the state-change goal you have to evaluate whether the goal has been met before doing the activity and thus may not have to perform it, but in the other the activity must always be performed. Another difference is that in the state-change goal just doing the activity once may not have achieved the goal (if there was a mistake) so there is the possibility of needing to try again.

9.1 Teaching Tasks Using Natural Instruction

Teaching is best illustrated with an example; the following goes through teaching Leonardo to turn his buttons on and then off. The teacher begins by asking Leo to do the task, "Buttons-On-and-Of." To this, the robot indicates that it does not know how to do this task and goes into learning mode. The teacher then tells the robot that the task starts with the task "Buttons-On"; again Leonardo indicates that he does not know "Buttons-On" either. As a response, the teacher begins to teach him how to do this subtask by leading him through it. When asked to "Press Button 1," the robot notices that pressing *Button 1* changes the state of the world, such that *Button 1* is now on. Hence, this is encoded as the goal of the press-action for *Button 1*, and this action is stored as part of the "Buttons-On" subtask.

The rest of the "Buttons-On" task is instructed in a similar fashion, and when all of the buttons are on the teacher tells Leonardo that the "Buttons-On" task is done. At this point the robot notices the difference in the state of the world before and after the "Buttons-On" task, and encodes that the goal of this subtask is to have all of the buttons in the ON state. This task is then stored as the first part of the larger "Buttons-On-and-Off" task.

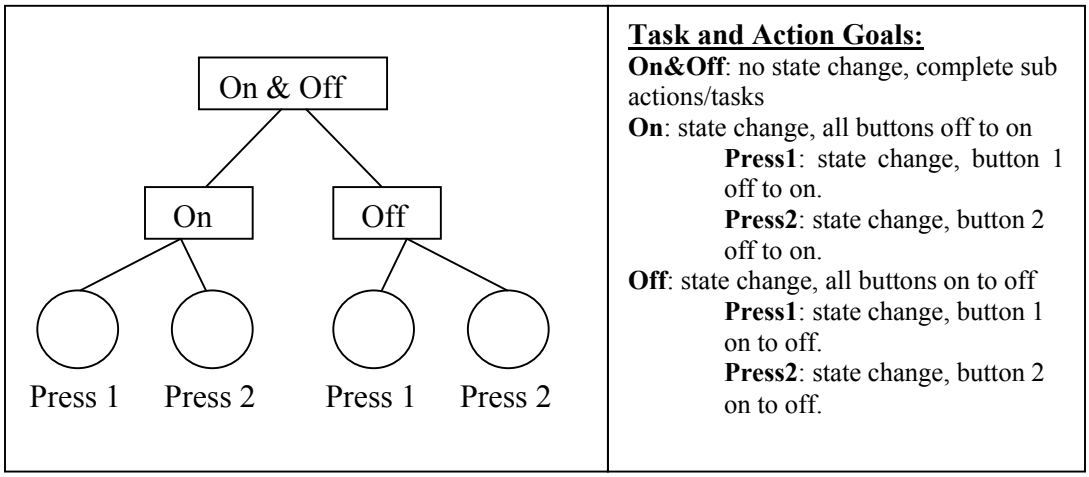


Figure 10: The diagram on the left is the resulting hierarchical task representation from the "Buttons-On-and-Off" learning scenario described in the text. The right box describes the task, subtask, and action level goals that Leonardo learned. The goal associated with the overall task is the completion 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.

The "Buttons-Off" subtask is taught in a similar way, and Leonardo then has a representation of the "Buttons-Off" task with the goal of ending up with all of the buttons in the off state. After this, the robot is told that the original "Buttons-On-and-Off" task is done. Leo sees that the state of the world before and after the "Buttons-On-and-Off" task is the same, so the robot assumes that the goal of the task is simply the act of doing it.

The overall task goal is encoded as a *just-do-it* goal. Figure 10 shows the resulting representation from the above teaching example.

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. Leonardo uses subtle expression to indicate to the human tutor when the robot is ready to learn something new, and its performance of taught actions provides the tutor with immediate feedback about the robot's 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. Natural key words such as "next", "first" are used to indicate task structure and sequencing constraints.

Failure to comprehend an action or its goal is easily and naturally detected by the tutor, and we are currently working to incorporate negative feedback to correct a task representation. As in a natural teacher-student interaction, errors are corrected just as they happen as part of the natural flow of the interaction.

10 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.

We placed a high importance on communicating the robot's perceived state of the world and the task (recall our discussion in section 3). Goals refer to world state as well as to activity state, establishing common ground between the robot and the human. As a result, joint intention, attention and planning is naturally achieved. Throughout the collaboration, the human partner has a clear idea as to Leonardo's current singular intent as part of the joint intent.

Leonardo meshes sub-plans for how to execute the shared task dynamically according to its abilities and the actions of the human partner. If, when Leo is doing one part of the task the human completes a separate element, Leonardo will take this into account and will no longer keep this on the list of things for it to do. Furthermore, before attempting an element of the task, Leonardo negotiates who should complete it. The robot has the ability to evaluate its own capabilities. Thus, if Leonardo is able to complete the task element, the robot will offer to do so. Conversely, whenever Leonardo believes that it cannot do the action, the robot will ask the human for help. Since Leonardo does not speak at the moment, the robot indicates its willingness to perform an action by pointing to its self and adopting an alert posture and facial expression. Analogously, when detecting an inability to perform an action assigned to it, Leonardo's expression indicates helplessness, as the robot points to the human. Leo uses gaze direction to indicate what it is that it needs help with.

Again, a variety of gestural cues have been used in order to communicate Leonardo's internal state (who it thinks is doing an action, whether it believe the goal has been met) with the human. When the human partner changes the state of the world, Leonardo acknowledges that it has detected this change by glancing shortly towards the area of change before redirecting its gaze to the human. We found this particularly valuable when the human completes part of the joint plan synchronously as Leonardo performs a different part of the task (or when the human unexpectedly changes something in the world). Leo's post-factum glance reassures the human collaborator that the robot noticed the human's actions and establishes a mutual belief on the progress of the shared plan. Similarly, Leonardo uses subtle nods while looking at his partner to indicate when he thinks a task or subtask is completed.

11 Discussion

For humans, learning and teaching form a coupled system in which the learner and the teacher work together. Humans can serve as motivated and natural teachers for creatures (natural or artificial) that are rewarding to teach. Given that humanoid robots are targeted for tasks in the human environment, it only makes sense to design them in a way that they can take advantage of natural human instruction.

The social and collaborative style of interaction that characterizes human tutelage plays an important role in constraining and guiding the learning process in a wide number of ways. At its best, the teaching and learning process is intimately coupled, well tuned, and transparent. The teacher and learner read and respond to the other, tuning their behavior to each other, to more effectively guide the learner's exploration. In this way, the instructor becomes a more effective teacher and the pupil a more effective learner---each simplifies the task for the other. The role of a teacher is largely to make the world simpler for the learner and to guide the search by providing timely feedback, luring the learner to perform desired behaviors, controlling the environment so the appropriate cues are easy to attend to, etc.

Our architecture supports the construction of collaborative learners that are easy and rewarding to teach using techniques that are intuitive for people. In Breazeal (2003) we outline several key issues in learning a new task, and describe how social cues, offered during natural human teaching scenarios, can be used to address them. These are meant to supplement existing statistical learning techniques, to make the robot's ability to learn more effective and require fewer examples. Leonardo's cognitive architecture addresses the following issues of learning from people as outlined below.

11.1 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 learning the task. As its perceptual abilities increase, the search space becomes enormous. If the robot could narrow in on those few relevant perceptions, the learning problem would become significantly more manageable.

Knowing what matters when learning a task is fundamentally a problem of determining saliency, which can be guided either externally (the environment), internally (the robot's goals, etc.), or socially (by the instructor) as discussed in section 6.3. Such social and guiding cues help the learner to identify the most relevant items to consider and accelerates 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 learner's attention must be transparent to the instructor so that he or she can easily infer what the learner is attending to, at what time, and what it is about to do.

Leonardo's attention system (in conjunction with its perceptual, discourse and motor systems) support a variety of attentional cues that help the robot to identify the appropriate object to be named (the referential focus), important state-action-result associations, and how to group these associations to acquire a high-level task model. Simultaneously, Leonardo communicates its focus of attention and current referential focus to the human using gaze direction and gestures. By doing so, human and robot can share joint attention over a wide number of possible contexts --- a critical skill for learning via tutelage.

11.2 Knowing What Action to Try

Once the robot has identified salient aspects of the scene, how does it determine what actions it should take? As robots become more complex, their repertoire of possible actions increases. This also contributes to a large search space. If the learner had a way of focusing on those potentially successful actions, the learning problem 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 (Billard *et. al.,* 1998) or another robot (Dautenhahn, 1995). The learner follows the model, thereby sharing a similar perceptual and motor state, to learn a reactive state-action policy (Billard & Dautenhahn, 1998). 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 (Billard *et. al.,* 1998). It has also been applied to learn a control policy for navigating a maze environment (Hayes & Demiris, 1994).

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 robot already knows how to perform, the human instructor can simply show or tell the robot what to do under what circumstances. Friedrich & Dillmann (1995) 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. This approach views teaching as explicit programming of the robot, which might prove burdensome. It also fails to capture the social and collaborative spirit of the teaching-learning process that characterizes human tutelage and our work with Leonardo.

In contrast, we model teaching as a collaborative process 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. At a more fundamental level, the instructor can teach the robot how to perform new motor skills via "internal" demonstration (recall section 7).

Through the process of learning the task model, Leonardo is able to infer the conditions upon when to apply the actions according to the situational context (the observable state of the world) and the task context (what portion of the task the robot is currently performing). Leonardo is also able to infer the desired outcome that is to result from performing each action. Because the robot can assess its own success, Leonardo can diligently bring about the desired situational context (using multiple attempts if necessary) before proceeding to the next portion of the task. The goal representation used for actions and subtasks, as well as the task model, allow the robot to perform the learned task either on its own, or in collaboration with the human.

11.3 Knowing How to Recognize Success and Correct Failures

Once a learner can observe an action and attempt to perform it, how can the robot determine whether it has been successful? How does it assign credit for that success? Further, if the learner 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.

Fortunately, the human instructor can help the robot do this given that he or she has a good understanding of the task and knows how to evaluate the learner's success and progress. One way that a human instructor facilitates the learner's evaluation process is by providing feedback through a number of channels (facial expression, gesture, speech, tone of voice, etc.). For instance, Nicolescu & Mataric (2003) use verbal cues to help their robot identify the incorrect aspects of its task model as it performs the task. 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. Animal training techniques, such as clicker training, have been used to teach a robot dog new tricks (Kaplan *et. al.,* 2002; Blumberg *et. al.,* 2002). In clicker training, the "click" sound is associated with a positive reward. Its distinctive sound and short duration allow the trainer to signal reward at a precise time to reinforce a specific association. In both of these examples, the timing of when the feedback is given (either positive or negative) is extremely important in helping the learner identify what is to be reinforced or corrected.

Feedback can also be provided to the robot to help it measure progress toward the goal (closer or further) or to recognize when the goal has been achieved. In this way, the human can play an important role in helping the robot to evaluate its own success through intuitive communication channels

In our work, Leonardo collaborates with the human instructor to infer his or her intended goal for each step of the instruction process. The robot is able to identify an intended goal on multiple levels---i.e., the desired outcome of an action, a subtask goal, or the overall task goal. The robot can distinguish two different kinds of goals: either to realize a desired state change on the world or simply to perform the action.

Leonardo is able to make these distinctions because of the collaborative process. There is 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 is an assumption that there is a desired outcome for each action, task, or subtask that the human offers to teach the robot. When Leonardo indicates that it does not how to perform something, it tries to infer the desired outcome when the human offers to teach the robot how to do it. Leonardo can make distinctions as to the nature of the goal and whether it is associated with an action, a subtask, or the overall task based on the instructional dialog and the changes it sees in the world state.

To confirm that it properly understood the desired outcome, Leonardo demonstrates this understanding to the instructor and waits for feedback. Positive feedback indicates that the robot is correct. We are currently adding negative feedback if the robot made the incorrect hypothesis and repair is necessary. If the human goes on to the next part of the task, the robot assumes that the inferred goal is correct. For state-change goals, the robot knows that it is not successful until it has achieved this desired outcome and it will reattempt if initially unsuccessful. If no net change on the world is observed after performing the action or subtask, then the robot assumes the goal is to simply perform it.

11.4 Knowing How to Explore

A human instructor plays an important role in helping the learner to explore its state-action space to discover a solution more quickly. Providing a robot with a demonstration of the desired task is one way to do this. For instance, teaching humanoid robots dexterous skills is challenging because their state-action space becomes prohibitively large to search for a viable solution in reasonable time. Imitative learning has been applied as an efficient way to explore this space, using a human's demonstration to help initialize the robot's own search (see Schaal 1999 for a review). 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 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 (Miyamoto *et. al.,* 1996) or hitting a tennis forehand (Miyamoto & Kawato, 1998)

Animal training techniques, such as shaping and luring, guide search by leading the learner through the task. Blumberg *et. al.* (2002) applied a luring technique in conjunction with clicker training to train a virtual dog to perform new tricks, such as "roll-over" or "walk-a-figure-eight." To do so, the virtual dog is given the behavior to try to place its nose as close as possible to the end of a training baton. By moving the baton around the virtual dog, the trainer can leverage this "follow-your-nose" behavior to help the character explore its movement repertoire. When the "dog" does a desired behavior, a clicking sound (already associated with a reward signal) reinforces the action with verbal command and biases the character to repeat that action in the future.

Alternatively, a human can use feedback to guide exploration. Within a reinforcement learning paradigm, Kaiser (1997) explores the use of positive or negative feedback given at the end of a trial after the robot demonstrates the desired skill. Refinement of learned skills is considered where an exploration process alters the performance skill during execution, and the feedback is used to assess the quality of the exploration. It was found, however, that providing a delayed reward at the end of the trial presents problems with credit assignment. As discussed in section 11.3, the timing of when feedback is given is very important to facilitate the learning process.

In our approach, the human instructor collaborates closely with the robot to guide its exploration as it quickly and efficiently learns each step of the task. This is done in a very intuitive way, where the robot leverages 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. The robot immediately demonstrates 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. Instead, 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 can learn a sophisticated and flexible task structure, including learning task components such as actions and subtasks from the bottom-up, very quickly and efficiently from few examples. The overall interaction is natural and intuitive for the instructor.

11.5 Knowing How to Leverage from Provided Structure

As discussed in section 4, little attention has been paid to the question of how a human can guide the learning process of a robot. Instead, 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. As a result, the most powerful techniques rely on the availability of a great deal of data in order to discover this structure. This makes these techniques difficult to apply when the number of examples may be very small or the time to learn the task is limited (e.g., the patience of the human). Also, such techniques are not well suited to rapidly changing environments where earlier examples are no longer relevant. In addition, 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.

In contrast, our approach recognizes that much of human instruction provides structure through social interaction and communication. This socially provided structure frames the learning task for the robot so that the task can be acquired quickly and efficiently. For instance, 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. For instance, Leonardo understands specific dialogs that cue the robot of the human's intention to teach the robot a new task and when the instruction for that task is completed. Other specific dialogs cue the robot as to where these tasks fit within the overall task structure (e.g., those that begin with sequencing words such as "first," "next," "finally," etc.).

To do this well, the instructor needs to maintain an accurate mental model of what the robot understands thus far. With this information, the instructor can control the rate of information exchange --- to either speed it up, to slow it down, repair misunderstandings, or to elaborate as appropriate. Leonardo helps the teacher to form an accurate mental model by communicating its internal state (what it is attending to, when it is confused, etc.) and demonstrating its understanding using natural cues such as mutual gaze, feedback nods, facial expressions, etc. By regulating the interaction in partnership with the robot, the instructor provides relevant instruction at the time the robot needs it (Breazeal, 2002; 2003).

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 (Breazeal 2000; 2002). The instructor demonstrates, the learner performs, and if necessary the instructor demonstrates again, often exaggerating or focusing on aspects of the task that were not performed successfully. Similarly, in ***learning by experienced demonstration,*** Nicolescu & Mataric (2003) view teaching and learning as a mixed-initiative process that interleaves instructive demonstrations with supervised practice trials of the robot. The human instructor leads the robot through the task as the first demonstration, and the robot encodes this into its approximation of the task model. The robot then demonstrates its representation of the task, where the teacher can intervene to correct and refine the robot's performance. Our approach models this as a more collaborative process where the robot shares responsibility for confirming the correctness of what it has learned so far with the human instructor. This allows errors to be repaired quickly at the time they occur, rather than being allowed to persist.

Furthermore, Leonardo and the human instructor collaborate to frame the purpose of the interaction. For instance, is the purpose of the interaction to learn from it or to demonstrate understanding? Also what level of representation is currently being addressed---at the level of an action, a subtask or the overall task? In Nicolescu & Mataric (2003), the instructor takes the sole responsibility by issuing single word commands that places the robot in either demonstration mode or learning mode. This is used by the human to stop the robot's demonstration during an incorrect portion of the

task, then put the robot in learning mode where it follows the human through a corrected sequence to update its task model, and finally restore demonstration mode to allow robot to finish the task. Alternatively, the instructor may provide a new demonstration of the same task in a different way to help the robot generalize its task representation.

In our approach, the robot shares this responsibility with the human. The robot readily confesses to what it does and does not understand, thereby prompting the human for what it would like to be taught next. When the human offers to teach the robot that step of the task, the robot does its best to learn how to achieve the goal specified by the human. Leonardo then takes the initiative to demonstrate this understanding and waits for feedback to either quickly confirm (using short utterances such as "good," "OK," etc.) or correct its performance. If the human goes on to the next step, the robot understands that this implies its performance was satisfactory and proceeds to the next step. In this manner, the robot fluidly switches between learning from the human and demonstrating what it has learned as the interaction unfolds at a natural pace. In addition, the human has confidence that the robot is acquiring the correct task structure because he or she has a good mental model of what the robot has learned so far.

12 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. As argued in section 2, there are many reasons to believe that a social interaction will be the most natural and intuitive way for ordinary people to work with humanoid robots and to teach them. In section 3, we carefully distinguished human-robot cooperation from other forms of human-robot interaction to identify those key issues that must be addressed to build robots that can serve as helpful assistants and effective teammates. Of primary importance is to build robots that understand people as people. In contrast to inanimate objects, our behavior arises from a rich network of mental states. In short, 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.

Toward this ambitious goal, we have presented how our ideas (informed by Joint Intention Theory and Theory of Mind) 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 Theory of Mind competencies, developed for our humanoid robot, and 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---from low level actions to subtasks, and how these relate to model the overall task.

In addition, we have shown how this approach can be extended to allow the robot to perform the learned task cooperatively with a human teammate. The robot collaborates with the human to maintain a common ground from which joint intention, attention, and planning are naturally achieved .The robot is aware if its own limitations and can work with the human to dynamically divide up the task appropriately (i.e., meshing sub-plans), 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 (e.g., using social cues such as glancing to notice the change in state the human just enacted, or giving quick nod to the human) so they are both in agreement as to what has be accomplished so far and what remains to be completed.

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 with which to interact. 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.

Acknowledgements

This work would not be possible for the contributions of many others. The Axiom FFt facial feature tracking software is provided by Nevengineering Inc.. Leonardo's speech understanding abilities are developed in collaboration with Alan Schultz and his group at the Navy Research Lab. In particular, we worked closely with Scott Thomas on the development of Leonardo's task-oriented dialogs. Stan Winston Studio provided the physical Leonardo robot. The real time 3D head pose tracking software was developed by Louis-Phillipe Morency and collaborators of the CSAIL Vision Interface Group under Prof. Trevor Darrell. Geoff Beatty and Ryan Kavanaugh provided the virtual model and animations. Leonardo's architecture is built on top of the C4 code base of the Synthetic Character Group at the MIT Media Lab, directed by Bruce Blumberg. This work is funded in part by a DARPA MARS grant and in part by the *Digital Life* and *Things that Think* consortia.

References

J. Astington and A. Gopnick (1991). Developing understanding of desire and intention. In A. Whiten (ed.) Natural Theories of Mind. Blackwell, Oxford UK, pp.39-50.

W. Bluethmann, R. Ambrose, M. Diftler, E. Huber, M. Goza, C. Lovchik, and D. Magruder (2003). "Robonaut: A Robot Designed to Work with Humans in Space,"*Autonomous Robots*, 14, pp. 179-207.

D. Baldwin, L. Moses (1994). "Early understanding of referential intent and attentional focus: Evidence from language and emotion," in *Children's Early Understanding of Mind* (C. Lewis & P. Mitchell, eds.). Lawrence Erlbaum Assoc., pp 133-155.

S. Baron-Cohen (1991).”Precursers to a Theory of Mind: Understanding Attention in Others,” In A. Whiten (ed.) *Natural Theories of Mind*. Blackwell. Chapter 16.

D. Baldwin & J. Baird (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences*, 5(4), pp. 171-178.

A. Billard & K. Dautenhahn (1998). “Grounding communication in autonomous robots: An experimental study,” *Robotics and Autonomous Systems* 24(1—2) pp 71—81.

A. Billard, K. Dautenhahn, and G. Hayes (1998) Experiments on human-robot communication with Robota, an imitative learning and communicating doll robot, in *Proceedings of Socially Situated Intelligence Workshop as part of the Fifth Conference of the Simulation of Adaptive Behavior. Centre for Policy Modelling technical report series number CPM-98-38.*

B. Blumberg, *et. al*. (2002). Integrated Learning for Interactive Synthetic Characters. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH ’02) (pp. 417 – 426). New York: ACM Press.

B. Bodenheimer, C. Rose and M. Cohen (1998). “Verbs and Adverbs: Multidimensional Motion Interpolation,” *IEEE Computer Graphics and Applications*, pp. 32-40.

M. Bratman (1992). “Shared Cooperative Activity,” *The Philosophical Review*, 101(2) pp. 327-341.

C. Breazeal (2003). “Social Interactions in HRI: The Robot View,” *IEEE Transactions on Man, Cybernetics and Systems: Part C*. Forthcoming.

C. Breazeal, D. Buchsbaum, J. Grey and B. Blumberg (2004). “Learning from and about others: Toward using imitation to bootstrap the social competence of robots. *Artificial Life*. Forthcoming.

C. Breazeal (2002). *Designing Sociable Robots*, MIT Press.

C. Breazeal & B. Scassellati (2002a). “Robots that imitate humans,” *Trends in Cognitive Sciences.* 6, 481-487.

C. Breazeal (2000). “Proto-conversations with an anthropomorphic robot," In *Proceedings of the Ninth IEEE International Workshop on Robot and Human Interactive Communication (Ro-Man2000).* Osaka, Japan, 328—333.

C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati (2001), "Active vision systems for sociable robots," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 31:5, pp. 443-453.

C. Breazeal & B. Scassellati (1999) A context dependent attention system for a social robot, in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99*), Stockholm, Sweden, pp. 1146—1151

R. Burke, D. Isla, M. Downie, Y. Ivanov and B. Blumberg (2001). “Creature Smarts: The art and architecture of a virtual brain,” Proceedings of the Computer Game Developers Conference.

Butterworth (1991). “The ontogeny and phylogeny of joint visual attention,” In A. Whiten (ed.) *Natural Theories of Mind*, Blackwell, Oxford U,pp 223-232.

J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson and H. Tan (2000). “Human conversation as a system framework: designing embodied conversational agents,” In J. Cassell, J. Sullivan, S. Prevost and E. Churchill (eds.) *Embodied Conversational Agents*. MIT Press. Cambridge MA.

H. Clark & D. Wilkes-Gibbs (1986). “Referring as a Collaborative Process,” *Cognition* 22, pp. 1-39.

P. Cohen, H. Levesque, J. Nunes, and S. Oviatt (1990). "Task-Oriented Dialog as a Consequence of Joint Activity," In *Proceedings of the 1990 Pacific Rim International Conference on Artificial Intelligence*, Nagoya Japan, pp. 203-208.

P. Cohen and H. Levesque (1991). "Teamwork," *Nous* **25,** pp. 487-512.

P. Cohen & H. Levesque (1990),"Persistence, Intention, and Commitment," In Cohen, Morgan and Pollack (eds.) *Intentions in communication*. MIT Press, Chapter 3.

T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb (2001). "Plan-view Trajectory Estimation with Dense Stereo Background Models," Proceedings of the International Conference on Computer Vision (ICCV-01), Vancouver, Canada, pp. 628-635.

K. Dautenhahn (1995). "Getting to know each other – Artificial social intelligence for autonomous robots," *Robotics and Autonomous Systems* 16, pp. 333-356.

M. Davies & T. Stone (Eds) (1995). *Mental Simulation*. Oxford: Blackwell Publishers.

D. Dennett (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.

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

M. Downie (2000). Behavior, animation, music: the music and movement of synthetic characters. Master's thesis, Cambridge, MA: MIT Press.

C. Eveland, K. Konolige, R. C. Bolles (1998). "Background Modeling for Segmentation of Video-rate Stereo Sequences," In Proceedings of Computer Vision and Pattern Recognition (CVPR-98), pp. 266-272.

O. Faugeras, B. Hotz, H. Mathieu, T. Viville, Z. Zhang, P. Fua, E. Thron, L Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy (1993). "Real Time Correlation-based Stereo: Algorithm, Implementations and Application," *INRIA* 2013.

C. Fong, C. Thorpe and C. Baur (2001). "Collaboration, Dialogue, and Human-Robot Interaction," In *Proceedings of the 2001 International Symposium of Robotics Research*. Victoria, Australia.

N. Franklin, B. Tversky and V. Coon (1992). Switching points of view in spatial mental models. *Memory & Cognition, 20*(5), 507-518.

H. Friedrich & R. Dillmann (1995). "Robot programming based on a single demonstration and user intentions," In Proceedings of the 3rd European Workshop on Learning Robots (ECML95). Heraklion Crete, Greece.

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

A.Goldman (2001). Desire, Intention, and the Simulation Theory. In F. Malle, L. Moses & D. Baldwin (eds.) *Intention and Intentionality*. MIT Press, Cambridge MA, pp. 207-224.

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

A. Gopnick and A. Moore (1994). Changing your views: How understanding visual perception can lead to a new theory of mind. In C. Lewis and P. Mitchell (eds.) Children's Early Understanding of Mind. Lawrence Erlbaum Press. pp 157-181.

H. Grice (1975). "Logic in Conversation," In P. Cole and J.L. Morgan (eds) *Syntax and Semantics 3: Speech Acts*. New York, Academic Press.

B. Gross (1996). "Collaborative Systems," 1994 AAAI Presidential Address. *AI Magazine* 2(17), pp. 67-85.

N. Habili, C. C. Lim and A. Moini (2001). "Hand and Face Segmentation using Motion and Color Cues in Digital Image Sequences," In *Proc. IEEE Int. Conf. on Multimedia and Expo.*

I. Haritaoglu, D. Harwood, and L. S. Davis (2000). "W4: Real-time Surveillance of People and their Activities," PAMI 22(8), pp. 809-830.

G. M. Hayes & J. Demiris (1994). "A robot controller using learning by imitation," In Proceedings of the International Symposium on Intelligent Robotic Systems, pp. 198-204. Grenoble, France.

J. Heal (2003). Understanding other minds from the inside. *Mind, Reason and Imagination*. Cambridge University Press, pp. 28-44.

R. Hornik, N. Risenhoover and M. Gunnar (1987). The effects of maternal positive, neutral, and negative affective communications on infant responses to new toys. *Child Development*. 58, pp. 937-944.

G. E. Hovland, P. Sikka & B.J. McCarragher (1996). Skill acquisition from human demonstration using a hidden Markov Model. *In Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA '96). (pp. 2706 – 2711). USA: IEEE.

Y. A. Ivanov, A. F. Bobick, and J. Liu (2000). "Fast Lighting Independent Background Subtraction," International Journal of Computer Vision 37(2), pp. 199-207.

M. Kaiser (1997). "Transfer or elementary skills via human-robot interaction," *Adaptive Behavior* 5(3-4), pp. 249-280.

F. Kaplan, P. Oudeyer, E. Kubinyi and A. Miklosi (2002). "Robotic clicker training," Robotics and Autonomous Systems (38), pp. 1987-206.

K. Konolige (1997). "Small Vision Systems: Hardware and Implementation," *Proceedings of the Eighth International Symposium on Robotics Research*, Hayama, Japah, pp. 203—212.

Y. Kuniyoshi, M. Inaba, and H. Inoue (1994). "Learning by watching: Extracting reuseable task knowledge from visual observation of human performance," *IEEE Trans. Robotics Automation* 10, pp. 799—822.

H. Levesque, P. Cohen, J. Nunes (1994), "On Acting Together," In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, pp. 94--99, Boston, MA, 1990.

H. Jones & S. Rock (2002), "Dialog-based human-robot interaction for space construction teams," In *Proceedings of the 2002IEEE Aerospace Conference proceedings*, Big Sky, MT.

M. J. Mataric (2002). Sensory-Motor Primitives as a Basis for Imitation: Linking Perception to Action and Biology to Robotics. In In Dautenhahn & C.L. Nehaniv (Eds.), *Imitation in Animals and Artifacts*. (pp.391 – 422). Cambridge, MA: MIT Press.

A. Meltzoff & A. Gopnik (1993) The role of imitation in understanding persons and developing a theory of mind. In Baron-Cohen S., Tager-Flusberg, H., & Cohen, D.J., (Eds.), *Understanding Other Minds, perspectives from autism*. (pp. 335 – 366). Oxford: Oxford University Press.

A. Meltzoff (1996). The Human Infant as Imitative Generalist: A 20-Year progress report on infant imitation with implications for comparative psychology. In Galef,

B.G. & Heyes, C.M. (Eds.), *Social Learning in Animals: The Roots of Culture. (*pp.347-370). New York: Academic Press.

A. Meltzoff & M.K. Moore (1997) Explaining facial imitation: A theoretical model. *Early Development and Parenting*. 6, 179-192.

H. Miyamoto, S. Schaal, F. Gandolfo, H. Gomi, Y. Koike, R. Osu, E. Nakano Y. Wada, and M. Kawato (1996) "A Kendama learning robot based on bi-directional theory," *Neural Networks* 9, 1181—1302.

H. Miyamoto and M. Kawato (1998) "A tennis serve and upswing learning robot based on bi-directional theory," *Neural Networks* 11, pp. 1131—1344

L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell (2002). "Fast Stereo-based Head Tracking for Interactive Environment," FG 2002, Washington D.C

M. Nicolescu & M. Mataric (2003). "Natural Methods for Robot Task Learning: Instructive Demonstrations, Generalization and Practice," *Proceedings of AAMAS 2003*. Melbourne, Australia.

D. Norman (2001). "How Might Humans Interact with Robots," Keynote address ot the DARPA-NSF Workshop on Human-Robot Ineraction, San Luis Obispo, CA.

D. Perzanowski, A. Schultz, W. Adams, E. Marsh and M. Bugajska (2001). "Building a Multimodal Human-Robot Interface,"*IEEE Intelligent Systems*, pp. 16-20.

D. Povinelli (1996), "Chimpanzee theory of mind? The long road to strong inference," In P. Carruthers and P. Smith (eds.) *Theories of Theories of Mind*. Cambridge University Press. Chapter 18.

D. Premack & G. Woodruff (1978). Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences*. 1:4, 515-526

B. Reeves & C. Nass (1996). *The Media Equation*, CSLI Publications, Stanford, CA.

J. Rickel and W.L. Johnson (2000). "Task-Oriented Collaboration with Embodied Agents in Virtual Worlds," In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (Eds.), *Embodied Conversational Agents*, Boston: MIT Press, pp. 95-122.

D. K. Roy (1999) *Learning from Sights and Sounds*: A Computational Model. Ph.D. Thesis, MIT Media Laboratory

H. Sacks, A. Schegloff, and G. Jefferson (1974). "A simplest systematics for the organization of turn-taking in conversation," *Language* 50, pp. 696-735.

B. Scassellati (1998) Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot, in *Computation for Metaphors, Analogy and Agents* (C. Nehaniv, ed.), Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, Springer-Verlag.

B. Scassellati (2000) *Foundations for a Theory of Mind for a Humanoid Robot*, PhD Thesis. Cambridge, MA: MIT Press.

S. Schaal (1999) Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences.* 3, 233-242.

S. Schaal, A. Ijspeert, and A. Billard (2000). Computational Approaches to Motor Learning by Imitation. *Philosophical Transaction of the Royal Society of London*: *Series B, Biological Sciences*. 358, 537-547.

S. Schaal (1997). "Learning from Demonstration," Advances in Neural Information Processing Systems (eds. Mozer, M., Jordan, M., and Petsche, T.), vol 9, pp. 1040-1046.

F. Strack, L. Martin, and S. Stepper (1988). Inhibiting and facilitating conditions of the human smile: A non-obtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54, 768–777.

H. Wellman (1991). From desires to beliefs: Acquisition of a theory of mind. In A. Whiten (ed.) Natural Theories of Mind. Blackwell, Oxford UK, pp19-38.

J. H. Williams, A. Whiten, T. Suddendorf & D. Perrett (2003). Imitation, mirror neurons and autism. *Neuroscience and Biobehavioral Reviews*.

A. Whiten & W. Byrne (1997). *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge University Press.

A. Woodward, J. Sommerville nad J. Guajardo (2001). How infants make sense of intentional action. In F. Malle, L. Moses & D. Baldwin (eds.) *Intention and Intentionality*. MIT Press, Cambridge MA, pp. 149-169.