WORKING COLLABORATIVELY WITH HUMANOID ROBOTS

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

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

Received 31/5/2004, passcode 27X-A9G7G6J9A4
Revised 30/09/2004
Accepted dd/mm/yyyy

This paper presents an overview of our work towards building humanoid robots that can work alongside people as cooperative teammates. We present our theoretical framework based on a novel combination of Joint Intention Theory and Collaborative Discourse Theory, and demonstrate how it can be applied to allow a human to work cooperatively with a humanoid robot on a joint task using speech, gesture, and expressive cues. Such issues must be addressed to enable many new and exciting applications for humanoid robots that require them assist ordinary people in daily activities or to work as capable members of human-robot teams.

Keywords: Human-robot interaction; teamwork; dialog and gesture; collaboration; social robots.

1. Introduction

Many new applications for autonomous robots in the human environment require them to help people as capable assistants or to work alongside people as cooperative members of human-robot teams[1,2]. For instance, humanoid robots are being developed to provide the elderly with assistance in their home. In other applications, humanoids are being developed to serve as members of human-robot teams for applications in space exploration, search and rescue, construction, agriculture, and more. In the future, we expect to see more applications for robots that share our environment and tools and participate in joint activities with untrained humans. This poses the important question of how robots should communicate and work with us.

1.1 Beyond Robots as Tools to Robot Partners

Robots today treat us either as other objects in the environment, or at best they interact with us in a manner characteristic of socially impaired people. For instance, robots are not really 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 generally 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 deeply 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. Although there have been initial strides in these areas[2], there remains significant shortcomings in the social intelligence of robots. As a

result, robots cannot cooperate with us as teammates or help us as assistants in a human-like way. Consequently, human-robot interaction often is reduced to using social cues merely as a natural interface for operating (supervising) the robot as a sophisticated tool. This sort of master-slave arrangement does not capture the sense of partnership that we mean when we speak of working "jointly with" humans.

Rather than viewing robots as semi-autonomous tools that are directed via human supervision, we envision robots that can cooperate with humans as capable partners. For instance, 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 using the same tools and components. To work as a team, 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 and components 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 out of sequence). 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

Human-robot collaboration of this nature is an important yet relatively unexplored kind of human-robot interaction[3]. This paper describes our efforts to move beyond robots as tools or appliances to robots that interact with humans as capable and cooperative partners. We apply our theoretical framework based on joint intention theory[4] and collaborative discourse theory[6,20] to enable our expressive humanoid robot, *Leonardo* (Figure 1), to work shoulder-to-shoulder with a human teammate on a joint task.

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 left picture shows the robotic structure, the center picture shows the robot when cosmetically finished, the right shows a simulated version of the robot.

2. Theoretical Framework

For applications where robots interact with people as partners, it is important to distinguish **human-robot collaboration** from other forms of human-robot interaction. Whereas interaction entails action *on* someone or something else, collaboration is inherently *working with* others[5,6,7]. Much of the current work in human-robot interaction is thus aptly labeled given that the robot (or group of robots) is viewed as a tool capable of some autonomy that a remote human operator commands to carry out a task[8,9,10].

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[4,7]. According to collaborative discourse theory[6,20], 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

Bratman[5] defines certain prerequisites for an activity to be considered shared and cooperative; he stresses the importance of *mutual responsiveness, commitment to the joint activity* and *commitment to mutual support*. Cohen and his collaborators[4,7,11] support these guidelines and but also predict that an efficient and robust collaboration scheme in a changing environment with partial knowledge commands an open channel of *communication*. Communication plays an important role in coordinating teammates' roles and actions to accomplish the task. It also serves to establish and maintain a set of mutual beliefs (also called common ground) among the team members.

What happens when things go wrong? According to Grosz[6], 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[6,12]. Specifically, the actions and goals that each team member adopts to do their part should not 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. Our work integrates these ideas to model and perform collaborative tasks for human-robot teams.

3. A Collaborative Task Scenario

In our experimental collaborative scenario, there are several buttons in front of Leonardo (see Figure 2). The human stands facing the robot across from the buttons to perform tasks with the robot using natural social cues (e.g., speech, gesture, head pose, etc). The buttons can be switched ON and OFF (which changes their color). Occasionally, a button that is pressed does not light up is considered a failed attempt.

To test our collaborative task execution implementation, we use tasks comprised of speech recognition and understanding (section 4), vision (section 5) and simple manipulation skills (section 6). We have designed a set of tasks involving a number of sequenced steps, such as turning a set of buttons ON and then OFF, turning a button ON as a sub-task of turning all the buttons ON, turning single buttons ON, and others. This task set represents simple and complex hierarchies and contains tasks with shared goals (section 7). (Please refer to the work of Lockerd & Breazeal[13,33], where we present how the robot learns a generalized task representation from human tutelage). Section 8

presents how we apply collaborative discourse to address a number of issues that arise within a collaborative setting – such as how the task can (and should) be divided between the participants, how the collaborator's actions need to be taken into account when deciding what to do next, how to provide mutual support in cases of one participant's inability to perform a certain action, and how to maintain a clear channel of communication to synchronize mutual beliefs and maintain common ground for intentions and actions.

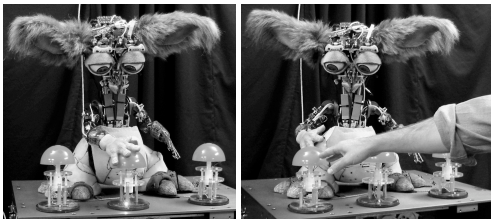

Figure 2: Leonardo following the human's request to activate the middle button (left). Leonardo learns the labels for each of his buttons by having a person point to a specific button and name it (right). In this picture, Leonardo and the human are both attending to the same button as the robot learns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

4. 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 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[9] (with respect to my reference frame versus your reference frame, etc.). 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. We have been developing perceptual, cognitive, and motor systems to support these dialogs.

5. The Vision System

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

below) used by humans to refer to objects and to direct the robot's attention to important aspects of the shared context.

5.1 Perceiving Objects

Each button is detected and tracked via saturated color matching on the intensity data. Once the appropriately saturated pixels have been extracted, a pixel labeling, a clustering and a morphology classification analysis is run over the results. In addition, an adaptive binary classification system is used to detect whether a button is on, using a colored LED at the center of each button whose lighting is toggled when the button is pressed.

5.2 Recognizing Deictic Gestures

We have implemented the ability to recognize 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, following a person's direction of gaze allows people to establish joint attention with others (see Figure 3). We have also implemented visual routines for recognizing pointing gestures (see Figure 4). In addition, we have developed a suite spatial reasoning routines that allow the robot to geometrically follow a human's head pose or pointing gesture to the indicated object referent.

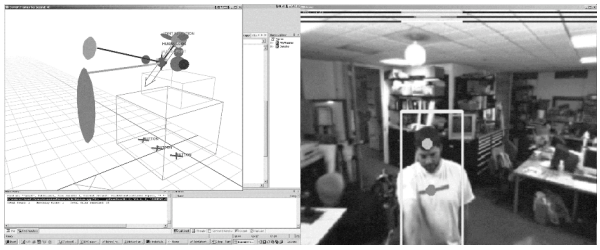

Figure 3: Visualizer showing the robot and a human sharing joint visual attention on the same object. The right image shows the visual input of a person looking at and pointing to the center button. The left image shows the visualization of the robot's internal model. The human's gaze is shown as the dark gray vector and his pointing gesture is shown by the medium gray vector. The robot looks at the same button (robot's dark gray vector) to establish joint attention.

Pointing gestures are detected by the overhead stereo camera, by employing background subtraction in the intensity and disparity domain. To that end, our background models are continuously updated with a two-element IIR lowpass filter. 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.

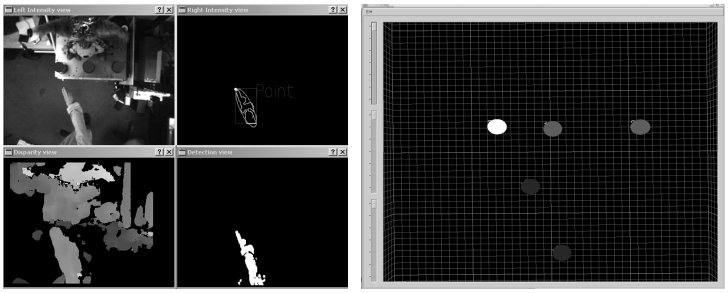

Figure 4: Computing the deictic reference to an object in the visual scene. Left, an overhead stereo camera identifies the locations of the buttons in the scene and recognizes when a pointing gesture occurs, estimating the location of tip of the finger and the angle of the forearm. This is passed to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

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.

Once the arm has been extracted, we recognize whether the hand is configured in a pointing gesture or not (see Figure 4). We accomplish this by estimating the kurtosis of the hand. 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. Object Manipulation Skills

Leonardo acquires the ability to press buttons from "internal" demonstration via a telemetry suit worn by a human operator. In this scenario, the operator "shows" Leonardo how to perform an action by guiding the robot using the telemetry suit. Meanwhile, the robot records these specific actions as they are applied to objects at specific locations. The calibrated mapping between the robot's joint angles and the telemetry suit's Euler angles is learned via an imitative interaction where the human mimics a repertoire of poses led by the robot[14].

Using this approach, the human demonstrator can "show" Leonardo how to press a button at several different locations in its workspace (typically less than 10 examples are needed). This defines the basis set of button-pressing examples that are indexed according to 2D button location provided by the robot's vision system. While the

robot runs autonomously, it can then interpolate these exemplars (see Eq. 1) using a dynamically weighted blend of the recorded button pressing trajectories, based on the Verb & Adverb animation blending technique[15].

For each joint angle $\boldsymbol{J_k}$ in the robot,

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

Where Ek,i is the kth joint angle in the ith exemplar, and Wi is the weight of the ith exemplar.

To determine the blend weights, we first precompute the Delaunay triangulation of the target points. We then 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. Once the weights are determined, we can blend these three source animations together according to the calculated weights on a per joint basis.

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. We are currently working on improving the accuracy that will be necessary for more demanding dexterous manipulations[16].

7. Task Representation to Support Collaboration

In this section we present our task representation for collaborative action. We argue that a goal-centric view is crucial in a collaborative task setting, in which goals provide a common ground for communication and interaction. Humans are biased to use an intention-based psychology to interpret an agent's actions[17]. Moreover, it has repeatedly been shown that we interpret intentions and actions based on goals, not specific activities or motion trajectories[18].All of this argues that goals and a commitment to their successful completion must be central to our intentional representation of tasks, especially if those should be performed in collaboration with others.

7.1 Intention and Task Representation

We represent tasks and their constituent actions in terms of *action tuples*[19] augmented with goals that play a central role both in the *precondition* that triggers the execution of a given action tuple, and in the *until-condition* that signals when the action tuple has successfully completed.

Our task representation currently distinguishes between two types of goals: (a) *state-change* goals that represent a change in the world, and (b) *just-do-it* goals that need to be executed regardless of their impact on the world. These two types of goals differ in both their evaluation as preconditions and in their evaluation as until-conditions. As part of a precondition, a state-change goal must be evaluated before doing the action to determine if the action is needed. As an until-condition, the robot shows commitment towards the state-change goal by executing the action, over multiple attempts if

necessary, until the robot succeeds in bringing about the desired new state. This commitment is an important aspect of intentional behavior[7]. Conversely, a just-do-it goal will lead to an action regardless of the world state, and will only be performed once.

Tasks are represented in a hierarchical structure of actions and sub-tasks (recursively defined in the same fashion). Since tasks, sub-tasks, and actions are derived from the same *action tuple* data structure, a tree structure is naturally afforded. It should be noted that goals are also associated with the successful completion of an overall task or sub-task, separate from the goals of each of the task's constituents.

7.2 Intention and Decision-Making

When executing a task, goals as preconditions and until-conditions of actions or sub-tasks manage the flow of decision-making throughout the task execution process. Additionally, overall task goals are evaluated separately from their constituent action goals. This top-level evaluation approach is not only more efficient than having to poll each of the constituent action goals, but is also conceptually in line with a goal-oriented hierarchical architecture. For example, consider a task with two actions. The first action makes some change in the world (and has a state-change goal), and the second action reverses that change (also a state-change goal). The overall task goal has no total state change and becomes a just-do-it goal although its constituent actions both have state-change goals.

7.3 Task manager

The *task manager module* maintains a collection of known task models and their associated names. Given this set of tasks, the robot listens for speech input that indicates a task-related request from the human partner. If the robot does not recognize the name of the requested task, or if the robot does not know how to perform it, he looks puzzled or shrugs his shoulders "I don't know."

In the case of an unknown task, the robot will learn the task's structure, its constituent actions and goals. In case of a task that's already known, the robot will attempt to perform the task, while at the same time refining its model of the task by responding to human instruction and feedback. This combined learning/execution approach allows for continuous and efficient refinement of the robot's abilities, employing an intuitive social interface to the human teacher. For a detailed discussion of our system's socially guided learning architecture, please refer to Lockerd et al.[13,33]

In the case of a known task, the task manager distinguishes between requests for autonomous task completion and invitations to task collaboration, and starts the appropriate execution module. If Leo is asked to do a known task on his own, then the task manager executes it autonomously by expanding the task's actions and sub-tasks onto a focus stack (in a similar way to Grosz & Sidner[20]). The task manager proceeds to work through the actions on the stack popping them as they are done and, upon encountering a sub-task, pushing its constituent actions onto the stack. The robot thus progresses through the task tree until the task's goals are achieved.

8. Collaborative Interaction

To make the collaboration a natural human interaction, we have implemented a number of mechanisms that people use during collaborative discourse. In particular, we have focused on task-oriented dialogs (section 8.1), flexible turn taking (section 8.2), communication acts to support joint activity (section 8.3), and self- assessment and mutual support (section 8.4) for joint activity.

8.1 Gestures and Expressions for Task-Oriented Dialogs

Dialog is fundamentally a cooperative[21], and we have implemented a suite of a collaborative task-oriented conversation and gestural policies for Leonardo. Cohen *et al.*[11] argue that much of task-oriented dialog can be understood in terms of Joint Intention Theory (see section 2.1). 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.*[11] 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 by the expert 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 the robot's gestures and facial expressions can be used to serve this purpose as well as speech acts (especially since Leonardo does not speak yet). A summary of Leonardo's cues are provided in Table 1. 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 and is ready to relinquish its turn (see Table 2). 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 the red button?" This confirms that both human and robot share the same belief regarding which the button is called by what name.

Table 1: Robot's gestures and expressions to support transparent communication of robot's internal state to human.

Social Cue	Communicated Intention	Interaction Function
Follows gesture to Object of Attention (OOA)	Establish OOA common ground	OOA set & ready for labeling
Point to object, look to object	Identify a particular object as referential focus (e.g., demonstrate correct association of name with object).	Confirm mutual belief about a particular object referent (e.g., successful identification of the target)
Confirming Nod (short)	Confirmation (e.g., OK, got it)	Update common ground of task state (e.g., attach label, start learning, etc.)
Affirming Nod (long)	Affirm query (e.g., Yes, I can)	Affirmation to query
Leaning forward and raising one ear towards human	Cannot understand (unable to recognize/parse speech)	Cues the human to repeat what was last said
Cocking head and shrugging (express confusion)	Cannot perform the request (lack of understanding)	Cues the human to add information or rectify shared beliefs (request clarification or elaboration)
Shake head	Cannot perform the request (lack of ability)	Cues that robot is not able to perform the request

Attention following and attention 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. 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 the introduction. 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.

In addition, back-channel signals (such as quick head nods) are given by the robot to let the human 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.

8.2 Turn Taking Skills

We have supplemented our models of collaborative dialog and gesture with flexible turn-taking skills modeled after those used by humans[22]. 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[22]. 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[23,24] as well as expressive robots[25,26].

Table 2: Implemented suite of envelope displays for flexible turn taking skills.

Social Cue	Communicated Intention	Interaction Function
Small ear perk and slight lean forward	Attention to human voice	Cues that robot is listening and attending to human
Break gaze, perform action	Acquire floor and begin turn	While the robot looks away, its turn is in progress
Looks back at human, arms relaxed	Turn is completed	Relinquish turn back to human

A number of envelope displays have been implemented on Leonardo to facilitate the exchange of turns between human and robot (see Table 2). 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).

8.3 Communication to Support Joint Activity

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

Table 3: Leonardo's cues to support joint activity with human.

Social Cue	Communicated Intention	Interaction Function
Looks back at the human, points to himself.	Gaze shift used to set turn taking boundaries. Gesture indicates perceived ability to perform an action.	Self-assessment and negotiating sub-plan meshing.
Glances to the OOA, and opens arms to the human.	Detects inability to perform needed action on OOA, asking for help.	Request human partner completes the step.
Looks at workspace.	Checks and updates change in task state due to own or other's act.	Acknowledge change in task state to other.
Eyes follow human action.	Acknowledges partner's action, maintains common ground.	Acknowledge that action is completed by other agent.

We have implemented a variety of gestures and other social cues to allow the robot communicate his internal state during collaboration – such as who the robot thinks

is doing an action, or whether the robot believes the goal has been met (Tables 1-3). For instance, when the human partner unexpectedly changes the state of the world, Leo acknowledges this change by glancing briefly towards the area of change before redirecting his gaze to the human. This post-action glance lets the human know that the robot is aware of what she has done, even if it does not advance the task.

If the human's simultaneous action contributes in a positive way to the task, such as turning ON a button during the buttons-ON sub-task, then Leonardo will glance at the change and give a small confirming nod to the human. Similarly, Leo uses subtle nods while looking at his partner to indicate when the robot thinks a task or sub-task is completed. For instance, Leo will give an acknowledgement nod to the human when the buttons-ON sub-task is completed before starting the buttons-OFF sub-task (in case of the buttons-ON-then-OFF task). All of these play an important role in establishing and maintaining mutual beliefs between human and robot on the progress of the shared plan.

8.4 Self Assessment and Mutual Support

At every stage of the interaction, either the human should do her part in the task or Leo should do his. Before attempting an element of the task, Leo negotiates who should complete it. For instance, Leo has the ability to evaluate his own capabilities. In the context of the button task, Leonardo can assess whether he can reach each button or not. If he is able to complete the task element (e.g., press a particular button) then he will offer to do so (see Table 3). Conversely, whenever he believes that he cannot do the action (e.g., because he cannot reach the button) he will ask the human for help..

Since Leonardo does not have speaking capabilities yet, he indicates his willingness to perform an action by pointing to himself, and adopting an alert posture and facial expression (Figure 5a). Analogously, when detecting an inability to perform an action assigned to him, Leonardo's expression indicates helplessness, as he gestures toward the human in a request for her to perform the intended action (Figure 5b). Additionally, Leo shifts his gaze between the problematic button and his partner to direct her attention to what it is that the robot needs help with.

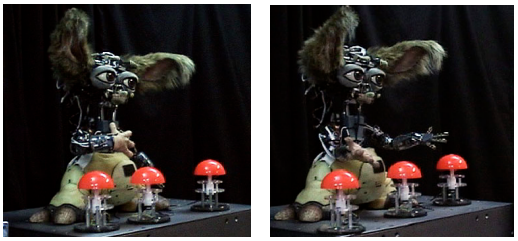

Figure 5: (a) Leonardo participating in a collaborative button-pressing task. (b) Leonardo negotiating his turn for an action he is able to perform.

8.5 The Task Collaboration Module

The above social skills are then introduced into a turn-taking mechanism we call the *Task Collaboration* module. The role of this module is to translate the joint intention-

represented by a task data structure as described in Section 7 – to the robot's individual intention, which in turn drives its actions.

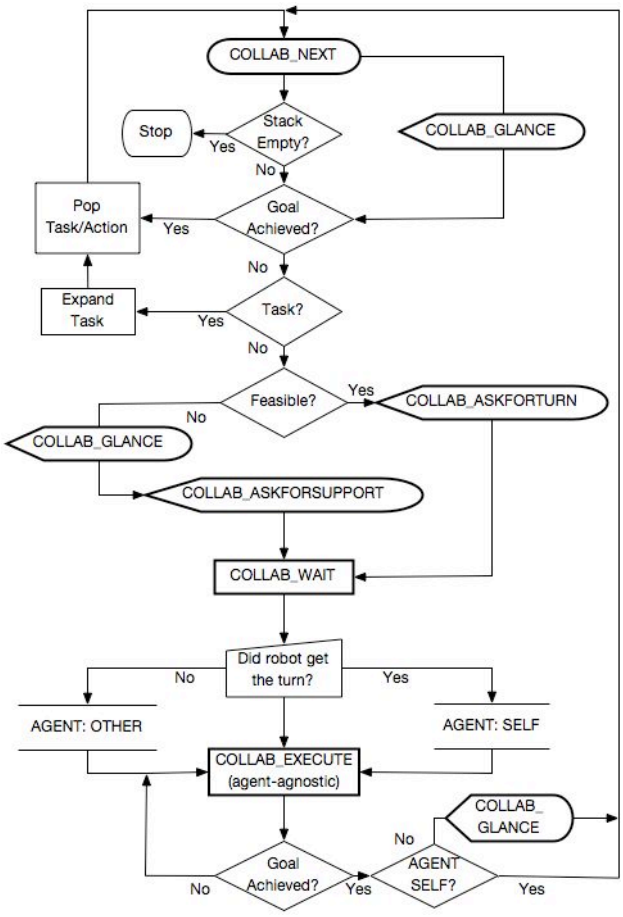

Figure 6: A schematic view of the Task Collaborator module. Note that the `COLLAB_WAIT` state can be terminated by both explicit and implicit turn taking on the human's part; and that in the agent-agnostic `COLLAB_EXECUTING` state the robot may be acting, but may also merely be following the teammate's progress on the action goal.

At its core, the Task Collaboration subsystem is implemented as a state machine (Figure 6) commanding the interaction flow throughout the collaboration, and triggering the appropriate social behaviors. The collaborator module's states are defined, in rough order of a typical interaction, as follows:

- `COLLAB_NEXT` – The initial state of the system, in which the robot evaluates the currently pertinent goal, and acts upon this evaluation.

- COLLAB_ASKFORTURN – If the robot is capable of performing the next step of the task, it will offer to take its turn.
- COLLAB_ASKFORSUPPORT – If the robot is not capable of performing the next step of the task, it will ask for support from the teammate.
- COLLAB_WAIT – Waiting for a response from the other participant in the interaction.
- COLLAB_EXECUTE – An agent-agnostic execution step. If the robot is executing the current action, this happens in this state; if the human teammate is executing a step, the robot waits in this state. In both cases the until-condition of the current action is continually evaluated.
- COLLAB_GLANCE – Establishing common ground by glancing at an object of attention, both for grounding in-sequence action and for joint closure on out-of-turn action.

Since our architecture is goal-based, a typical operative step begins by evaluating the goal of the currently pertinent task. If it has not been achieved, the robot decomposes the task into its constituent sub-tasks and sub-actions and recursively adds them to the focus stack. If an atomic action reaches the top of the focus stack, it is assigned to an agent in the team. Currently this can be either AGENT_SELF or AGENT_OTHER, but the framework allows for any number of agents. The robot then tracks the performance on the current task step, and dynamically adjust its plan according to the changes in goal satisfaction throughout the collaboration, as described above.

9. Performing a Task In Collaboration with People

In sum, 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 2). 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.

We have conducted a few early experiments using the framework described herein, and have found these cues to play a significant role in establishing and maintaining mutual beliefs between the teammates on the progress of the shared plan, and in increasing the efficiency of the human-robot collaboration process. Table 4 shows a sample transcript describing typical task collaboration between Leonardo and a human teammate. We chose to display the following simple tasks for reasons of transcript brevity: *BUTTON-ONE* – Toggle button one, *BUTTON-ONE-AND-TWO* – Turn buttons one and two ON. While these do not illustrate the Leonardo's full range of goal-oriented task representation capabilities, they offer a sense of the joint intention and communicative skills fundamental to the collaborative discourse stressed in this section.

Table 4: Sample task collaboration on single-level task.

T	Human	Robot	Notes
1	"Leo, let's do task BUTTONS"	Shrugs "I don't know"	Leo does not know this task.
2	"Let's do task BUTTON-ONE"	Looks at the buttons	Leo acknowledges that he understands the task, and visibly establishes mutual belief on the task's initial conditions.
3		Points to himself	He can do the first (and only) part of the task, and suggests doing so.
4	"OK, you go"	Presses button one, looking at it	Looking away from the partner while operating establishes turn taking boundaries.
5		Looks back at his partner	Gaze shift is used to signal end of turn
6		Nods shortly	Communicates the robot's perceived end of task
7	"Leo, let's do task BUTTON-ONE"	Looks at the buttons; points to himself	As in steps 2-3
8	"I'll go "	Looks at his partner	
9	Presses button one	Looks at button one	Acknowledges partner's action, creates mutual belief
10		Nods shortly	Communicates perceived end of task.
11	Moves button one out of Leo's reach		
12	"Let us do task BUTTON-ONE"	Looks at buttons	Leo acknowledges that he understands the task, and visibly establishes mutual belief on the task's initial conditions.
13		Looks at button one, then back at the human partner; extends his arms in "Help me" gesture.	Leo assesses his capabilities and consequently requests support.
14	Presses button one	Looks at button one; looks back at human; nods shortly.	Glance acknowledges partner's action and creates mutual belief as to the task's completion.
15	"Let us do task BUTTON-ONE-AND-TWO"	Looks at buttons	Leo acknowledges that he understands the task, and visibly establishes mutual belief on the task's initial conditions
16		Points to himself	He can do the first part of the task, and suggests doing so.
17	"OK, you go"	Presses button one, looking at it	
18	At the same time as 17, presses button two		
19		Looks at button two; looks back at the human; nods shortly	Acknowledges partner's simultaneous action, creates mutual belief as to the task's completion.

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

Note that in Trial 4, there is a case of simultaneous action handling, in which the human changes the world state in opposition to Leo's perceived task goal. In this case,

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

In summary, during the trials for the collaborative button task, Leonardo displayed successful meshing of sub-plans based on the dynamic state changes as a result of his successes, failures, and the partner's actions. Leo's gestures and facial expressions provided a natural collaborative environment, informing the human partner of Leo's understanding of the task state and his attempts to take or relinquish his turn. Leo's requests for help displayed his understanding of his own limitations, and his use of gaze and posture served as natural cues for the human to take appropriate action in each case.

As future work, we would like to improve the complexity of the task representation as well as the interaction and dialog. Although Leonardo's gestures and facial expressions are designed to communicate his internal state, combining this with an ability to speak would give the robot more precision in the information that he can convey. We would also like to implement a richer set of conversational policies to support collaboration. This would be useful for negotiating the meshing of sub-plans during task execution to make this process more flexible and efficient. We continue to make improvements to Leonardo's task representation so that he can represent a larger class of collaborative tasks and more involved constraints between the tasks' action components.

10. Discussion

In viewing human-robot interaction as fundamentally a collaborative process and designing robots that communicate using natural human social skills, we believe that robots will be intuitive for humans to interact. Toward this goal, we have presented our ability to coordinate joint intentions via collaborative dialog to perform a task jointly with a robot. We have shown how we incorporate social acts that support collaborative dialog – the robot continually communicates its internal state to the human partner and maintains a mutual belief about the task at hand. This makes working together more efficient and transparent. In this section, we discuss our approach in the context of related work. Viewed in the context of joint intention and collaborative discourse framework, our approach is significantly different than other approaches to human-robot interaction. Our goal is broader than interaction; we try to achieve *collaboration* between human and robot partners.

10.1 Collaboration vs. Interaction

As discussed in section 2, human-style cooperative behavior is an ongoing process of maintaining mutual beliefs, sharing relevant knowledge, coordinating action, and demonstrating commitment to doing one's own part, helping the other to do theirs, and completing the shared task. Using joint intention theory and collaborative discourse theory as our theoretical framework, we have incorporated the notions of joint intentions and collaborative communication in our implementation. Our goal oriented task representation allows the robot to reason about the task on multiple levels, easily sharing

the plan execution with a partner and adjusting to changes in the world state. The robot acts in accordance with joint intentions, and also works to communicate and establish mutual beliefs about the task state as the interaction progresses (e.g. confirming when a particular step is complete, and negotiating who will complete a portion of the task).

In related work, Kimura *et al* explore human-robot collaboration with vision-based robotic arms.[27] While addressing many of the task representation and labor division aspects necessary for teamwork, it views the collaborative act as a planning problem, devoid of any social aspect. As such, it does not take advantage of the inherent human expertise in generating and understanding social acts. As a result, the interaction requires the human teammate to learn gestures and vocal utterances akin to programming commands.

Fong *et al*. consider a working partnership between human and robot in terms of *collaborative control*, where a human and a robot collaborate in vehicle teleoperation.[10] The robot maintains a model of the user, can take specific commands from the operator, and also has the ability to ask the human questions to resolve issues in the plan or perceptual ambiguities. The role of the human in the partnership is to serve as a reliable remote source of information. In contrast, our work explores collaboration where the human and robot work together on a collocated task where both the human and the robot can complete steps of the plan. Because the human and robot act upon a shared environment, the robot must therefore notice changes made by the human and dynamically reassess the plan and coordinate actions accordingly.

Some work in the field of human and virtual agent teams also has the notion of shared plans that must be continually maintained and updated according to changes in the world state. For instance, Traum *et al* have a system in which a human is part of a team of agents that work together in a virtual world.[28] Their system addresses plan reassessment and uses dialog models and speech acts to negotiate a plan as a team. Roles are attached to various steps of the plan, and an authority structure helps in negotiating control. Our work differs in two respects from this virtual teamwork system. First, in our physically embodied scenario, we explore the issues of face-to-face gestures and socially relevant communication acts that facilitate collaboration. Second, we do not utilize an authority structure; instead, the robot and the human negotiate turns in the context of a shared plan.

Employing social cues for dialog and collaboration has been investigated in the field of embodied conversational agents (ECA). The agent's verbal and nonverbal communication, including gesture, gaze, and head pose has been explored in tutorial systems, e.g. by Rickel and Johnson[24] and in embodied dialog systems, e.g. Thórisson.[30] On the opposite side of the spectrum, Nakano et al. have studied human means of face-to-face grounding and implemented their findings in the design of an ECA for a collaborative dialog system. Their agent detects and analyzes head pose and gaze of the human and offers appropriate elaboration when needed.[31] By the very nature of virtual agents, the tasks in both cases have been primarily informational, and could therefore not capture the physical aspects of shoulder-to-shoulder collaboration between a human and a robot, in particular with regard to object manipulation.

Moreover, this work has been predominantly concerned with grounding in dialog with regards to discourse contributions in a sequential dialog and neither with joint intention, joint action, a shared workspace and hierarchical tasks. In case of the ECA discussed above, the roles of the human and the artificial agents were clearly separated, whereas our work stresses the case of a shared action in which the task must be divided

between the human and robotic team members. To enable this, we believe that social skills need to be applied at every level of the interaction, and our robot maintains grounding on the perceptual level, the object reference level and the task progression level. Finally, it is important to note that attempts to transfer these important concepts to the realm of autonomous agents, and in particular robotic agents have so far been rare, rudimentary, and also curiously often information-centric.[32]

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

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

11. Conclusion

This paper presents an overview of our work to build sociable humanoid robots that work cooperatively with people using natural dialog, gesture, and social cues. We have shown how our approach allows our robot to perform a given 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 of 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 been 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 and work with humanoid robots in the future. These implications reach far beyond making robots appealing, entertaining, or easy with which to interact. Human-robot collaboration is a critical competence for robots that will play a useful, rewarding, and long-term role in the daily lives of ordinary people – robots that will be able to cooperate with as capable partners rather than needing to be operated neither manually or by explicit supervision as a complicated tool.

Acknowledgements

This work is made possible by the contributions of many collaborators. 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. Geoff Beatty and Ryan Kavanaugh provided the virtual model and animations. Leonardo's architecture is built on top of the C5M 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

1. C. Breazeal, Social Interactions in HRI: The Robot View, *IEEE Transactions on Man, Cybernetics and Systems: Part C*, 34(2), pp. 181-186, (2004).
2. T. Fong, I. Nourbakshsh, & K. Dautenhahn. A Survey of Social Robots. *Robotics and Autonomous Systems,* 42, pp. 143 – 166, (2002).
3. C. Breazeal, G. Hoffman, A. Lockerd. Working and Teaching Robots as a Collaboration. *Proceedings of AAMAS-04, to appear* (2004).
4. P. Cohen and H. Levesque. Teamwork, *Nous* 25,. 487-512 (1991).
5. M. Bratman, Shared Cooperative Activity, *The Philosophical Review*, 101(2) 327-341, (1992).
6. B. Grosz, Collaborative Systems, 1994 AAAI Pres. Address, *AI Magazine* 2(17),67-85 (1996).
7. P. Cohen and H. Levesque, Persistence, Intention, and Commitment, in Cohen, Morgan and Pollack (eds.) *Intentions in communication* (MIT Press, 1990), Chapter 3.
8. H. Jones and S. Rock, Dialog-based Human-robot Interaction for Space Construction Teams. *IEEE Aerospace Conference* (2002).
9. D. Perzanowski, A. Schultz, W. Adams, E. Marsh and M. Bugajska, Building a Multimodal Human-Robot Interface, *IEEE Intelligent Systems*, pp. 16-20, (2001).
10. T. Fong, C. Thorpe and C. Baur, Collaboration, Dialogue, and Human-Robot Interaction, *Proceedings of the International Symposium of Robotics Research* (2001).
11. P. Cohen, H. Levesque, J. Nunes, and S. Oviatt , Task-Oriented Dialog as a Consequence of Joint Activity, In *1990 Pacific Rim International Conference on Artificial Intelligence*, Nagoya Japan, pp. 203-208, (1990).
12. H. Levesque, P. Cohen, J. Nunes, On Acting Together, In *Eighth National Conference on Artificial Intelligence (AAAI-90)*, (Boston, MA, 1990), pp. 94—99.
13. A. Lockerd and C. Breazeal, Robot Learning Through Collaborative Dialog, In *AAAI-04 Workshop on Supervisory Control of Learning and Adaptive Systems, to appear* (2004)
14. A. Brooks, J. Gray, G. Hoffman, A. Lockerd, H. Lee, C. Breazeal. Robot's Play: Interactive Games with Sociable Machines, *Proceedings of ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE 2004), to appear* (2004).
15. B. Bodenheimer, C. Rose and M. Cohen. "Verbs and Adverbs: Multidimensional Motion Interpolation," *IEEE Computer Graphics and Applications*, pp. 32-40, (1998).
16. J. Lieberman and C. Breazeal. Improvements on Action Parsing and Movement Interpolation for Learning through Demonstration, submitted to *Humanoids 2004*. In review.
17. D.A. Baldwin and J.A. Baird, Discerning Intentions in Dynamic Human Action, *Trends in Cognitive Sciences*, 5(4), 171-178 (2001).
18. B. Gleissner, A. N. Meltzoff and H. Bekkering, Children's coding of human action: cognitive factors influencing imitation in 3-year-olds, *Developmental Science*, 3(4), 405-414 (2000).

19. B. Blumberg, *et. al*., Integrated Learning for Interactive Synthetic Characters. In *29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '02)* (New York: ACM Press, 2002), pp. 417 – 426
20. B. J. Grosz and C. L. Sidner, Plans for discourse in P. R. Cohen and J. Morgan and M. E. Pollack (eds.), *Intentions in communication* (MIT Press,1990), pp. 417—444.
21. H. Grice, Logic and Conversation, In P. Cole and J.L. Morgan (eds.) *Syntax and Semantics 3: Speech Acts (*New York, Academic Press, 1975).
22. H. Sacks, A. Schegloff, and G. Jefferson, A simplest systematics for the organization of turn taking in conversation, *Language* 50, 696-735, (1974).
23. J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson and H. Tan, 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, 2000).
24. J. Rickel and W.L. Johnson, 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, 2000) pp. 95-122.
25. C. Breazeal, Proto-conversations with an anthropomorphic robot, In *Ninth IEEE International Workshop on Robot and Human Interactive Communication (Ro-Man2000)* (Osaka, Japan, 2000), pp. 328—333.
26. C. Breazeal, *Designing Sociable Robots*, (MIT Press, 2002).
27. H. Kimura, T. Horiuchi and K. Ikeuchi, Task-Model Based Human Robot Cooperation Using Vision, in *Int'l Conference on Intelligent Robots and Systems (IROS'99)* (1999), pp. 701-706.
28. D. Traum, J. Rickel, J. Gratch, and S. Marsella, Negotiation over Tasks in Hybrid Human-Agent Teams for Simulation-Based Training, in *Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, (Melbourne, Australia, 2003) pp. 441-448.
29. C. Breazeal, A. Brooks, D. Chilongo, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, and A. Lockerd. Teaching Humanoid Robots via Socially Guided Learning. To appear *International Journal of Humanoid Robots* (2004).
30. K. R. Thórisson. Gandalf: an embodied humanoid capable of real-time multimodal dialogue with people. In the*1st int'l conference on Autonomous agents*, pages 536-537., 1997.
31. Y. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Association for Computational Linguistics*, Sapporo, Japan, July 2003.
32. Hideaki Kuzuoka, Keiichi Yamazaki, Akiko Yamazaki, Jun'ichi Kosaka, Yasuko Suga, and Christian Heath. Dual ecologies of robot as communication media: thoughts on coordinating orientations and projectability. In *Proceedings of the 2004 conference on Human factors in computing systems (CHI2004)*, pages 183–190. ACM Press, 2004.
33. A. Lockerd and C. Breazeal. Tutelage and socially guided robot learning. In *International Conference on Intelligent Robots and Systems (IROS'04),* Sendai, Japan, September 2004.