Enabling Effective Human–Robot Interaction Using Perspective-Taking in Robots

J. Gregory Trafton, Nicholas L. Cassimatis, Magdalena D. Bugajska, Derek P. Brock, Farilee E. Mintz, and Alan C. Schultz

***Abstract*—We propose that an important aspect of human–robot interaction is perspective-taking. We show how perspective-taking occurs in a naturalistic environment (astronauts working on a collaborative project) and present a cognitive architecture for performing perspective-taking called Polyscheme. Finally, we show a fully integrated system that instantiates our theoretical framework within a working robot system. Our system successfully solves a series of perspective-taking problems and uses the same frames of references that astronauts do to facilitate collaborative problem solving with a person.**

***Index Terms*—Cognitive modeling, human–robot-interaction, perspective-taking.**

I. Introduction

WHAT guidelines should a designer use to create an interface for human–robot interaction? Unfortunately, there are few overarching theories or models that give good advice on how to design the interface between humans and robots. A great deal of work within human–computer interaction suggests that if a designer creates an interface without good guidelines, without paying attention to the way that people perceive, reason, and act, and without evaluation, the interface turns out to be quite poor [1]–[3]. In other words, a "good idea" from a designer could turn out to be idiosyncratic or arbitrary for most users of the system. We suggest that the default approach for designers should be to use person-to-person interaction as the model for human–robot interaction. Other models and techniques will doubtless be better in some situations, but, since people are able to communicate so well with other people, it makes sense to use interactions between people as the default model for designing and implementing human–robot interaction. There are, of course, many facets of human–human interaction, but we will focus here on one of the most important: the basic ability of people to take one another's perspective and reason about interactions and the world from this alternative point of view. Perspective-taking has been shown to occur in a wide variety of situations and tasks, varying from social situations [4], [5] to way finding and navigation tasks [6]–[10]. Spatial perspective-taking seems to occur in children as young as age four [11]–[14] and develops relatively systematically [15].

Fig. 1. When told "give me the wrench," the robot needs to take the perspective of the person to determine which wrench the astronaut has referred to.

As fundamental as perspective-taking is for people, it is not surprising that perspective-taking abilities on robots would be a valuable asset for people working with them. Imagine, for example, how much more effective a robot capable of perspective-taking would be in helping an astronaut with an assembly task, even if the robot's job were something as relatively simple as giving the astronaut various tools and parts as they were needed. Fig. 1 shows one possible scenario. The robot and the person are facing each other. The robot can see that there are two wrenches in the setting, wrenches 1 (W1) and 2 (W2), but the astronaut only sees W2, from his perspective because W1 is occluded by an obstacle. If the astronaut says, "Robot, give me the wrench," the meaning of the phrase "the wrench" is ambiguous for the robot because it knows of two wrenches.

Manuscript received August 7, 2004; revised February 15, 2005 and March 4, 2005. This work was supported in part by the Office of Naval Research under Work Order N0001404WX30001 (Task Unit 8551) and in part by DARPA IPTO under the MARS program. This paper was recommended by the Guest Editors.

J. G. Trafton, M. D. Bugajska, D. P. Brock, and A. C. Schultz are with the Naval Research Laboratory, Washington, DC 20375-5337 USA (e-mail: trafton@itd.nrl.navy.mil; brock@itd.nrl.navy.mil; magda@aic.nrl.navy.mil; schultz@aic.nrl.navy.mil).

N. L. Cassimatis is with Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: cassin@rpi.edu).

F. E. Mintz is with ITT at NRL, Washington, DC 20375-5337 USA (e-mail: mintz@aic.nrl.navy.mil).

Digital Object Identifier 10.1109/TSMCA.2005.850592

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Enabling Effective Human-Robot Interaction Using Perspective-Taking in Robots	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory,Navy Center for Applied Research in Artificial Intelligence (NCARAI),4555 Overlook Avenue SW,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We propose that an important aspect of human-robot interaction is perspective-taking.We show how perspective-taking occurs in a naturalistic environment (astronauts working on a collaborative project) and present a cognitive architecture for performing perspective-taking called Polyscheme. Finally, we show a fully integrated system that instantiates our theoretical framework within a working robot system. Our system successfully solves a series of perspective-taking problems and uses the same frames of references that astronauts do to facilitate collaborative problem solving with a person.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

The phrase is unambiguous to the astronaut, though, because he only sees one wrench. Intuitively, if the robot could take the perspective of the astronaut, it would see that W2 is the only wrench in the astronaut's field of view and could therefore surmise that "the wrench" must refer to W2. Even in this rudimentary scenario, perspective-taking would immediately enhance the human–robot interaction.

If perspective-taking is likely to be a valuable tool for human–robot interaction, why are there so few examples of robots with perspective-taking in the literature? In fact, one of the only computational systems that uses a form of perspective-taking has been Soar [16], [17] within a gaming environment [18]. The Soar system uses perspective-taking and anticipation to predict what an opponent will do. Our system focuses more on human–robot interaction, where there are potentially many possible actions for a robot partner to take. Interestingly, both our approach and the approach taken within Soar have an emphasis on cognition and on how people think. It is likely that a noncognitive system would have a much more difficult time building a system that uses perspective-taking, since not only would they have to model the way people think (which both our approach and Soar do), but they would also need to determine how to use perspective-taking within their system. Additionally, a robot requires substantial computational resources just to represent the world from its own perspective, and clearly, even more resources would be needed to represent the world from the perspective of a human counterpart. Add to this a requirement to quickly react to dynamic factors in the task environment and possibly account for the presence of additional participants, which entails the representation of more perspectives, and the issue of computational resources is only compounded. Second, representing the perspective of humans requires a robot to integrate multiple data structures and algorithms for perceiving, representing, and making inferences about the world from that perspective. For example, in a task where a robot and a person are cooperating to fix a vehicle, aspects of the person's perspective that can affect their interaction include his spatial location (he might be able to see things from his location that the robot cannot and vice versa), his knowledge of the current situation (he may know of a different method for accomplishing the task than the robot), his knowledge of the specific task (he may not know of a problem with one of the parts that the robot knows about), and his linguistic background (he may use words that the robot does not know). Thus, the common problem in robotics of integrating multiple subsystems that utilize different data structures and representations extends to robot perspective-taking as well.

How prevalent is perspective-taking in tasks where robots may be of assistance? To answer this question, we analyzed videos of astronauts as they trained for extravehicular activities (EVAs) in a simulated microgravity environment called the Neutral Buoyancy Laboratory (NBL) at NASA's Johnson Space Center. EVAs are exactly the type of activity researchers at NASA believe robots would be ideally suited for [19]. As astronauts and ground control worked out procedures and defined roles, it was immediately evident that spatial perspective-taking and the use of spatial language are present in astronauts' work in these EVA environments.

In space, astronauts have to deal with frames of reference and spatial situations in ways that people on earth typically do not encounter. Down can easily mean something completely different in a weightless setting than its normal, earth-bound sense of toward the ground. Despite the potential for confusion, astronauts seem to have no problem using and understanding spatial language with each other or in taking one another's point of view. The mixed orientations of weightless environments, though, may well add an additional challenge for spatial perspective-taking in robots and for their interactive comprehension of astronauts' spatial language. However, virtually all of the experimental work on spatial language and perspective-taking to-date has focused on five frames of reference: exocentric (world-based, such as "Go north"), egocentric (self-based, "Turn to my left"), addressee-centered (other-based, "Turn to your left"), deictic ("Go here [points]"), and object-centric (object-based, "The fork is to the left of the plate") [20]–[26]. Thus, in our analysis, we used this framework to explore the type and amount of spatial perspective-taking that arose among the astronauts in training.

II. Human–Human Interaction Study

A. Method

We analyzed a series of video tapes of astronauts training in the NBL for Space Station Mission 9A. Astronaut utterances were collected as they performed a cooperative assembly task, specifically the construction of the first right-side Truss segment and the Crew and Equipment Translation Aid (CETA) Cart A. Throughout the training, three individuals were primarily involved in conversations and working together: Ground (the person in charge, issuing instructions to accomplish) and EV1 and EV2 (the two astronauts performing the task). Ground could see what was happening from multiple perspectives through various cameras. All three individuals could communicate through microphones. The training session lasted over 6 h. The unit of analysis was the Instruction (e.g., "Go forward") and instruction follow-ups (e.g., "OK, going forward"). Off-task utterances (jokes, etc.) were coded as off-task. All on-task utterances were coded using standard protocol analysis techniques [27]. A total of 4000 on-task utterances were coded.

B. Results

Approximately half of the utterances (2113 out of 4000) were instructions and instruction follow-ups. The other half was confirmation ("OK"), general dialog, and so on. There were far more instruction follow-ups than instructions (1590 versus 523 utterances), $\chi^2(1) = 538.8, p < .001$. Interestingly, the pattern of results for instructions and follow-ups was not significantly different, so they were combined for the following analyzes.

Table I shows the five different types of utterances and the overall rate that they occurred in the corpus. In one very real sense, each instruction could be categorized as "addressee-centered," since every instruction (by definition) was a request for someone else to perform a task. Similarly, each follow-up instruction could be categorized as egocentric, since the person

TABLE I
ASTRONAUT UTTERANCE TYPES AND EXAMPLES AS THEY WORKED ON A COLLABORATIVE ASSEMBLY TASK

Frame of Reference	Example	% Utterances
Exocentric	Go straight zenith ("up")	16%
Egocentric	I am right side double tethered	12%
Addressee-centered	Now bend both your legs	11%
Deictic	Put it over [there] (Points)	1%
Object-centered	Put the forward part of the spud into position	60%

was describing his or her own actions. However, each instruction was coded according to the kind of spatial language that was used within the utterance.

As Table I suggests, the most common utterance was object-centered, $\chi^2(4) = 530.1$, $p < 0.001$, Bonferonni adjusted $\chi^2 p < 0.005$. This result is not surprising, since the astronauts were working mostly with objects. Previous researchers have shown that when making an object-based utterance, the object's reference frame is based primarily on its function: the "top" of a cup is where the liquid is poured into, regardless of the orientation of the cup [21], [22]. In our analysis, the same finding seems to be true: astronauts referred primarily to objects' functional relations.

Second, approximately a quarter of the utterances required some perspective-taking; either the speaker needed to take the point of view of the listener, or the listener needed to take the point of view of the speaker.

Third, consistent with other research [10], people switch perspectives quite often, approximately once every other utterance. When a speaker is talking without interruption, they switch perspectives 45% of the time. Similarly, when a new speaker enters into a conversation, that utterance is also likely to be a different from the original speaker's perspective 44% (477 out of 1083 speaker transitions) of the time. The brief conversation fragment shown in Table II accurately illustrates all three of these points.

Notice several things about this conversation. First, Ground mixes reference frames: addressee-centered ("straight down from where you are"), object-centered ("down under the rail"), addressee-centered ("by your right hand"), and exocentric ("straight nadir" which means toward the earth) all occur in the first instruction that ground gives in this fragment. Second, the participants come up with a new name for a unique unseen object ("the mystery hand-rail") and then tacitly agree to refer to it with this nomenclature later in the dialog.

Other researchers have found at least as much evidence for perspective-taking in psychological studies focused on language and spatial settings. In one study, for instance, while describing spatial environments, a range of 25% to 31% of participants' utterances involved perspective-taking [28]; in another, while writing descriptions of spatial environments, use of perspective-taking in participants' sentences ranged from 28% to 31% [29]. And in our laboratory, we found participants' use of perspective-taking in a virtual robot navigation task ranged from 3% to 72% depending on condition [25]. Findings such as these indicate that perspective-taking plays a substantial role in how people communicate about physical spaces and tasks, and support the focus of the work presented in the remainder of this article. In particular, spatial perspective-taking abilities should be a high priority of human–robot interaction research; it is important for good human–robot interaction when collaborating in shared space, and without it, we believe that autonomous robots will be poor collaborators, at best, in many human–robot activities.

TABLE II
DIALOG BETWEEN TWO ASTRONAUTS AND AN OBSERVER (NAMES HAVE BEEN CHANGED TO PRESERVE CONFIDENTIALITY)

EV1	EV2	Ground
		Bob, if you come straight down from where you are, uh, and uh kind of peek down under the rail on the nadir side, by your right hand, almost straight nadir, you should see the uh,
	Mystery hand-rail	
		The mystery hand-rail, exactly
	OK	
There's a mystery hand-rail?		
		Oh, it's that sneaky one. It's there's only one in that whole face.
Oh, yeah, a mystery one.		
		And you kinda gotta cruise around until you find it sometimes.
I like that name.		

III. SIMULATING PERSPECTIVES USING COGNITIVELY PLAUSIBLE MECHANISMS

As we stated in our introduction, we work from the premise that human–robot interaction is best modeled on human–human interaction principles. This view has led us to a general approach for building human–robot interaction tools that embraces three, interrelated conceptual guidelines.

1) Robotic representation, reasoning and perception mechanisms should be as similar to those of humans as possible.
2) Cognitive systems for human–robot interaction should be based on integrated cognitive architectures.
3) The use of heuristics and principles in collaborative activities similar to those ordinarily employed by people is consistent with people's expectations, and so, is consistent with effective human–robot interaction design.

In addition, a corollary guideline for our perspective-taking work can be stated as follows:

To perform collaborative tasks with humans in physical settings, a robot must be able to simulate and reason about the world from the perspective or vantage point of others.

We believe that these are merely guidelines for building good human–robot interaction. A more in-depth description of some of these guidelines can be found in [30]. Before we turn to a description of our current implementation and the status of our perspective-taking work, we first discuss some of the bases for our guidelines.

A. Similar Representations and Processes

When computational systems are designed to reason about collaborative interactions with representations and processes that are functionally similar to those used by people, the goal of intuitive interaction design is arguably facilitated. A clear example of this comes from spatial reasoning, where in general, people seem to use a combination of spatial and propositional knowledge [9], [31]–[36]. As a matter of practice, though, robotic approaches to spatial reasoning must take into account such factors as the variety and limitations of sensor data, the functional structure of this data, its use in path planning algorithms, and so on, little of which is represented internally in ways that are intuitively meaningful to humans. Thus, while it is a straightforward matter to design an interface that requires the input of numerical coordinates for route specification, it is nontrivial to design an interface that allows a user to specify a route with a hand-drawn map. Current work by Skubic and her colleagues [37], [38] is addressing this very issue. Aspects of the problem include the extraction of qualitative information and its translation into a functionally correct route while coping with incomplete map information and various distortions of scale. The goal of this work is to facilitate the design of a system that is able to represent and reason about space in a way that is functionally similar to how people think about it. While it is hardly possible for robots to use human-like mechanisms for all cognition, to the extent that this is possible, it will make robot simulations of human perspective more intuitive for purposes of collaboration and interaction [30].

B. Integrated Cognition

Human cognition is clearly integrated—researchers may disagree over how and where the integration occurs [39]–[41], but virtually all cognitive scientists agree that cognition is integrated. Likewise, we believe that the cognitive aspects of robotics systems—especially thinking and reasoning—should be integrated as well. Another, more speculative benefit is that since humans are such good general-purpose intelligent systems that have many effective mechanisms for interacting with other humans, choosing human-like mechanisms is a design heuristic for bringing robots closer to this ideal [30].

C. Cognitively Plausible Simulations for Perspective-Taking

A robot's ability to predict or resolve ambiguities in the behavior of a person by simulating the world from the person's perspective should greatly facilitate interactions with that person. When a robot simulates the behavior of a person engaged in a task, it can predict and therefore assist with the next action, e.g., by fetching a needed tool or by offering information that might make it possible to execute the action more effectively. A robot can also simulate a person's perspective to disambiguate speech or gestures, such as the earlier wrench example shown in Fig. 1. For these reasons, we have decided to design an architecture for human–robot interaction based on simulations of the perspective of another person (see also [42]).

An important virtue of a simulation-based architecture for human–robot interaction is that it enables a considerable amount of computational parsimony by reusing subsystems, both for reasoning about the world and reasoning about other people's perspectives. Many inference algorithms can be considered strategies for running mental simulations [43], [44]. For example, in backtracking search, a series of counterfactual states are represented and evaluated (or "simulated") until a solution is found. Stochastic simulation algorithms repeatedly conduct simulations of possible worlds to determine the likelihood of propositions being true in those worlds. Thus, because mechanisms for simulating counterfactual worlds are used widely in intelligent systems, we have attempted to use these mechanisms to simulate perspectives saving the expense of adding new reasoning mechanisms only for human–robot interaction. Cognitive scientists have also found strong evidence of mental simulation for counterfactual reasoning [45], [46].

Finally, as was alluded to earlier in our discussion of representations and processes, simulating the perspective of a person requires robots to use multiple data structures and algorithms since different aspects of a person's perspective on the world are best represented using different techniques. We therefore chose to base our work on a cognitive architecture called Polyscheme [43], [44], which was designed to model how humans integrate multiple representational methods to keep track of the world. Polyscheme, to be described in the next section, also has the benefits of having rich facilities for representing counterfactual worlds and thus can naturally implement simulations of people's perspectives.

IV. Details of Implementation

This section provides an architectural overview of our approach to improving human–robot interaction by enabling robots to simulate the world from the perspective of humans. We first describe the Polyscheme cognitive architecture that this work is based on, and then describe how we apply it to robot perspective-taking. We have developed this framework in order to be as general as possible and therefore do not present it in this section in the context of a specific task or domain. Subsequent sections describe the details of actual implementations and results in specific tasks.

A. Polyscheme

Polyscheme is a cognitive architecture that has been designed both to model how humans integrate multiple representations and inference techniques and to produce intelligent systems by combining the benefits of multiple representations, planning, and reasoning methods. Polyscheme has also been integrated onto a robotic architecture to provide symbolic reasoning and planning algorithms while maintaining the flexibility and robustness of reactive control systems [43].

Polyscheme is implemented in Java and runs on most computing systems.

B. Representing the Current State of the World

We first describe the mechanisms Polyscheme uses to represent the current state of the world and then describe how these mechanisms are used to simulate counterfactual worlds, including those that correspond to the perspective of people.

Since different aspects of the world are best represented by different data structures, Polyscheme programs are constructed from modules, called *specialists*, which represent these aspects using their own specialized data structures. For example, a temporal constraint specialist could keep track of constraints among temporal intervals using Allen's temporal constraint propagation algorithm [47], while an object location specialist could keep track of object locations using an evidence grid.

Since the responsibilities of specialists will overlap (e.g., a temporal constraint specialist and a qualitative physics specialist can both make inferences about the temporal relation between two events), and because one specialist can use information from another (e.g., a qualitative physics specialist can use information from a specialist that remembers object locations), Polyscheme has a mechanism for specialists to communicate with each other called the *focus of attention*. At every time step, all specialists "focus" on the same aspect of the world, which is represented as a literal proposition. For example, when the focus of attention is Color(x, red), all specialists focus on the color of the object x. When specialists focus on a proposition, they all indicate the truth value their inter-representation has for that proposition and submit to Polyscheme's *focus manager* propositions on which they would like to focus, either because they follow from the current focus of attention or because they would help determine its truth value. How the focus manager chooses the next focus of attention will be described below.

Polyscheme's representation of the current state of the world therefore is the combination of each specialist's representation of the world. Focus of attention determines to which aspect of the world the specialists will devote their representational and inferential abilities. By including modules based on different representations, Polyscheme resembles many multiagent systems. Its distinguishing characteristics involve how the computations of these specialists are coordinated (the focus of attention), its ability to represent counterfactual worlds, and its ability to implement reasoning algorithms, not by encapsulating them inside a specialists, but through strategies (focus schemes) for guiding the specialists' attention. These last two mechanisms will be discussed in the next two sections.

C. Representing Alternative States of the World

Representing alternative states of the world is a common theme among many otherwise disparate approaches to reasoning and planning. The underpinnings to many reasoning and planning algorithms are search through a state space. Stochastic simulation algorithms for propagating probabilities in Bayesian Networks sample from and simulate possible states of the world. Logics with possible worlds semantics have been used to formalize notions of information, belief, knowledge, and causality. Such formalisms have also been used to formalize aspects of linguistic semantics. The ability to represent alternative states of the world is thus key to Polyscheme's ability to integrate multiple representations and algorithms. In particular, all specialists in Polyscheme are required to be able to focus on and represent alternate states of the world. This is also reflected in the language for expressing propositions that constitute the focus of attention. Every proposition in Polyscheme has a "world" argument. For example, the propositions Color(x, red, w) states that x is red in alternate world w. The "real world"—the state of the world that is actual—is abbreviated R.

An important feature of worlds in Polsycheme is the inheritance relationships among them. When world w is the hypothetical world where P is true, we say that w is *based* on P. For example, the hypothetical world where x is green is based on Color(x, green, R). If P is true in the real world and there is no reason to infer it is false in w, then specialists are to assume that P is true in w as well. Thus, in the hypothetical world where x is green, Boston is still taken to be in Massachusetts unless otherwise assumed or inferred. This relationship between worlds is used in our perspective-taking work to efficiently represent the perspective of people without having to explicitly represent every aspect of it.

D. Choosing Simulations

Since each proposition has a world argument in it, the choice of which proposition to make the focus of attention determines which alternate world Polyscheme will consider. The focus of attention is chosen at each time step, when specialists submit propositions to the focus manager to which they would like to attend, and the focus manager chooses one of these propositions as the next focus of attention. How the focus manager chooses a proposition depends on various factors—including activations associated with each proposition by specialists—that are beyond the scope of this article and are explained elsewhere [43]. What is important for this discussion is that the manner in which specialists suggest propositions for attention, and how the focus manager chooses them, amounts to a strategy, called a *focus scheme*, for guiding the attention of the specialists in Polyscheme. Focus schemes are described here by natural language approximations.

Two focus schemes, one for probabilistic inference and another for search, will illustrate how Polyscheme uses simulations to integrate multiple, disparate inference algorithms. The stochastic simulation focus scheme implements probabilistic inference in Polyscheme:

> When the specialists think P is p/q times more likely than$\neg P$, focus on the world where P is true $\text{N}p$ times and the world where P is false $\text{N}q$ times, where N is some integer.

Repeated application of the counterfactual simulation focus scheme implements search:

> When the specialists are uncertain about P, focus on the world where P is true and the world where P is false.

In this way, reasoning algorithms from different subfields of artificial intelligence are integrated in one system using the simulation of counterfactual worlds. Because each step of each simulation is conducted using all the specialists, which can include statistical and perceptual representations and processes, Polyscheme provides continual symbolic, statistical, and perceptual integration.

V. Using Simulations of Perspective for Human–Robot Interaction

As discussed earlier, our approach has been to enable robots to simulate the world from the perspective of people so that they can interact with them more effectively. Two particular focus schemes, one for communication and one for cooperation in a task, illustrate this approach.

The first focus scheme, called *command simulation*, causes the robot to simulate the world from a person's perspective in order to disambiguate the person's commands:

> When a person, P, gives a command, simulate giving the command from P's perspective.

The effect of this focus scheme is that (elements of) commands given by P that are literally ambiguous will become clear. For instance, in the example shown in Fig. 1 where the robot knows about two wrenches and the astronaut knows about only one, a command simulation focus scheme can disambiguate the utterance.

The second focus scheme, called *action simulation*, causes robots to predict the actions of humans so that they can better understand their commands or help the person without being instructed:

> When a person P is engaged in a task, simulate P's actions (forward into the future) from P's perspective.

After simulating P's actions into the future (i.e., predicting what P will do), a robot can take steps to assist in those actions or more clearly understand commands involving those actions. For example, if a particular kind of wrench is required for the next step in the task, the robot can fetch that wrench, tell the person where it is, or understand which wrench the person intends when he commands, "Give me the wrench."

This approach has two benefits. First, because simulating the perspective of another person allows robots to disambiguate human commands and offer assistance without prompting for additional utterances, the amount of communication between the human and the robot is greatly reduced, while the quality of the communication is greatly increased. This enables humans and robots to cooperate more efficiently in more sophisticated tasks. Second, because the simulations that constitute the robot's reasoning are continually integrated with multiple representations, including those arising from new sensor information, all this interaction can occur with the flexibility and robustness that is required of robot applications.

VI. Implementation

Thus far, we have implemented this architecture on a robotic platform named Coyote, to allow it to more effectively collaborate with people. Details of the full system can be found elsewhere [30], [48]–[53], but a high-level description of the system will be provided here.

The robot is a commercial Nomadic Technologies Nomad200 suited to operation in office environments. It has a zero turn radius drive system, an array of range, image, and tactile sensors, and an onboard network of Linux and Windows computers with a wireless Ethernet link to the external computer network.

In addition to general mobility enabled by sonar and LADAR, the robot recognizes particular objects in its environment by using the CMVision package [54]. This vision system was used to provide simple color blob detection using an inexpensive digital camera mounted on the robot. In our scenarios, the robot only needed to be able to recognize orange traffic cones and boxes, which were used to create occlusions.

The human user could interact with the mobile robot using natural language and gestures that are part of our multimodal interface [48]–[51], [55]. The natural language component of the interface uses a commercial speech recognition engine, ViaVoice, to analyze spoken utterances. The speech signal is translated to a text string that is further analyzed by our in-house natural language understanding system, Nautilus [56], to produce a regularized expression. This latter representation is linked, where necessary, to gesture information, and an appropriate robot action or response results. Note that we use ViaVoice purely for syntactic input.

Polyscheme interacted with the other robot processes through TCP/IP sockets. After receiving an instruction, Polyscheme reasoned about what was needed and integrated perceptual information from the CMVision package. Polyscheme instructed the robot where to go, and the mobility system would then plan a path to that location and perform collision avoidance to get Coyote to within a small epsilon of that location.

In order to address the robot's problem of integrating multiple representation and inference techniques to represent the perspective of a person, several Polyscheme specialists for various data structures and algorithms are used on Coyote. The *perception specialist* uses color segmentation [54] and laser range finding to identify and localize objects, and also receives verbal input. The *temporal perception* specialist keeps track of the order of events using Allen's [47] temporal constraint framework. The *space specialist* keeps track of the location of objects. The *perspective specialist* computes the objects that are visible to a person from the person's current location. It also infers that if a person knows where something is, it is because he has seen it (this is an assumption we have built into the task domain). The *identity hypothesis* specialist, whose role is described below, uses a neural network to guess which object perceived in the past corresponds to an object perceived at present. The identity constraint system propagates identities [e.g., it infers that if $\text{Same}(x, y)$ and $\text{Same}(y, z)$, then $\text{Same}(x, z)$]. The *spatial relationship specialist* has the ability to reason about all types of spatial relationships encountered in the data collected during the NASA astronaut exercise.

Finally, Coyote has several focus schemes for reasoning in addition to the two perspective-taking focus schemes discussed in the previous section. One of these, the *counterfactual simulation* focus scheme, which was described earlier, will be useful in the example below.

VII. Simplifying HRI

In order for our system to simplify human–robot interaction, it must work in a variety of spatial situations and with a variety of frames of reference, as our astronaut data suggest. To fully explore the combined system, we created a number of situations where perspective-taking is needed to varying degrees. In all these scenarios, Coyote and the person are together in a

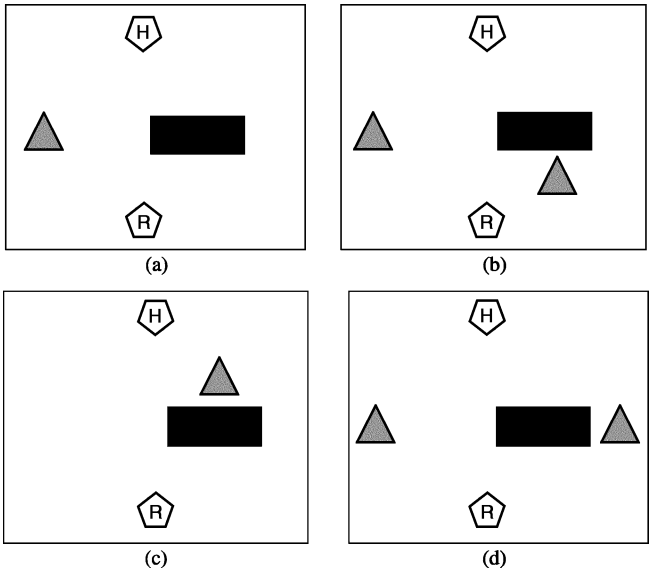

Fig. 2. Scenario diagrams. Triangles are the cones and the rectangles are the occluding boxes. Human (H), on top of each diagram, and robot (R), on the bottom of each diagram, are facing each other. (a) Scenarios 1; (b) 2; (c) 3; and (d) 4.

TABLE III
DIFFERENT SCENARIOS WE GAVE TO COYOTE TO EXAMINE PERSPECTIVE-TAKING

	Environment	What can robot see?	What can human see?	Solution
1	1 Cone, A	Cone A	Cone A	Go to cone A
2	2 Cones, A and B	Both Cones	Cone A	Go to cone A
3	1 Cone, A	No Cones	Cone A	Check hidden location
4	2 Cones, A and B	Both Cones	Both Cones	Request clarification

room with several objects and possible occlusions from either the robot's or human's perspective. The most relevant objects will be two orange traffic cones and a set of boxes. Fig. 2 and Table III describe the different scenarios we examined. In all cases, the human gives the robot the instruction "Coyote, go to the cone."

In all cases, perspective-taking is available to the system (it is the same integrated system throughout). However, in some cases, perspective-taking is a critical component of the reasoning process (scenarios 2 and 3, where there is a hidden cone), while in others it is either not needed (scenario 1, single visible cone) or does not help with the disambiguation process (scenario 4, two cones, visible to both).

Scenario 1, single visible cone, presents the simplest case: The person and Coyote can both see the same cone. When given the instruction "Coyote, go to the cone," Coyote confirms that the human indeed can see Cone A and then it simply navigates to the cone.

In scenario 2 (two cones, one hidden from human's sight, illustrated in Fig. 3), Coyote is initially situated so that it sees the two traffic cones, while the person can only see one, since

Fig. 3. Scenario in which Coyote (in the foreground at the bottom of the picture) can see two orange traffic cones while human can only see one.

TABLE IV
POLYSCHEME PROPOSITIONS AND THEIR MEANING FOR REPRESENTING THE COMMAND "GO TO THE CONE." THE SPEAKERS' OWN WORLD MODEL IS REPRESENTED BY $wSpeaker$

Proposition	Meaning
Exists(ref,E,wSpeaker)	There exists an object, ref, in the speaker's mind, wSpeaker
Category(ref,Cone,E,wSpeaker)	The speaker thinks it is a cone
Location(ref,refLoc,wSpeaker, R)	The cone is at refLoc
WantToGo(robot,refLoc,wSpeaker,E,R)	The speaker wants Coyote to go to the cone's location

the other is occluded by the boxes. In order to perform the task, Coyote must decide which cone the person referred to.

It will help in describing the sequence of Coyote's reasoning to explain how natural language utterances correspond to propositions about what Coyote sees. In this task, the two cones Coyote sees are $c1$ and $c2$. Coyote knows $\text{Category}(c1, \text{Cone}, E, R)$ and $\text{Category}(c2, \text{Cone}, E, R)$. "Go to the cone" is represented with the propositions in Table IV.

Coyote's task is to decide whether $\text{Same}(\text{ref}, c1, E, wSpeaker)$ or $\text{Same}(\text{ref}, c2, E, wSpeaker)$ is true. In other words, Coyote's task is to decide whether the object (ref) to which the speaker refers is identical to cone 1 $(c1)$ or to cone 2 $(c2)$. Thus, the problem of resolving the phrase's reference is represented as an identity problem. To resolve a reference, therefore, is to find which perceived object is identical to the referred object.

Table V shows an outline of the sequence of inferences Coyote makes in order to resolve the following ambiguity.

In scenario 3, the robot cannot see the cone because it is being occluded by the box from the robot's position. The robot must now infer that the cone is in a location that the person can see but the robot cannot. The system uses perspective-taking to choose a location hidden to the robot, but visible to the human, and promptly navigates there. Once Coyote gets to the new location, it repeats the process to find the cone.

Scenario 4 presents the robot with an extremely ambiguous case: There are two cones in the environment which both human and robot can see. When asked to go to the cone, the robot can neither navigate directly to the cone nor determine which cone

TABLE V
OUTLINE OF POLYSCHEME'S REASONING IN ORDER TO SOLVE SCENARIO 2 (TWO CONES, ONE HIDDEN FROM HUMAN'S SIGHT)

The perception specialist makes a request, which the focus manager grants, for the propositions describing the perceived cones and boxes and the person's location. ("pl4-1-0" refers to the place with coordinate (4,1,0).)	Category(c1,Cone,E,R) Location(c1,pl2-1-0,E,R) Category(c2,Cone,E,R) Location(c2,pl4-1-0,E,R) Category(box,Box,E,R) Location(box, pl2-1-0,E,R) Location(speaker, pl2-0-0,E,R)
The perception specialists makes a request, which the focus manager grants, for the propositions describing the person's command:	Exists(ref,E,wSpeaker) Category(ref,Cone,E, wSpeaker) Location(ref,refLoc, wSpeaker,R) WantToGo(robot,refLoc, wSpeaker,E,R)
The command simulation focus scheme keeps the focus on the world of the speaker's mind (*wSpeaker*) and the perspective-specialist infers that if the speaker knows about *ref* (Exists(ref,E,wSpeaker), then he must be able to see it. It therefore asks for the following proposition to be focused on:	CanSee(ref,E,wSpeaker)
The identity hypothesis specialist formulates the hypothesis that the cone the speaker refers to (cone) is identical to *c1* and to *c2*.	Same(c1,ref,E,R) Same(c2,ref,E,R)
The identity constraint specialists infers that since *c1* = *ref* and ref = *c2* that *c1* = *c3*:	NOT Same(c1,c2,E,R) The space specialist indicates that Same(c1,c2,E,R) must be false since they are at different locations at the same time.
The identity constraint specialist therefore decides that either *c1* != *ref* or *c2* != *ref* and asks for each proposition to be focused on:	Same(c1,ref,E,R) Same(c2,ref,E,R)
The counterfactual simulation focus scheme recognizes the conflict and requests that the world, wC1, where *c1* = *ref* and the world wC2, where *c2* = *ref* be imagined. The focus manager picks wC1 first.	Same(c1,ref,E,wC2)
The perspective specialist infers that in the world where *c1* = *ref*, that the person cannot see *c1* because it is blocked by the box.	NOT CanSee(ref,E,wC2)
Since it has already been inferred that the speaker can see *ref*, world wC2 is contradictory and the world specialist infers that Same(ref,c2,E,R) is false since this is what was assumed to create wC2.	NOT Same(ref,c2,E,R)
The identity constraint system infers that *ref* = *c1*, thus resolving the ambiguity:	Same(ref,c1,E,R)

to go to based on the person's perspective. Therefore, it must ask for assistance (e.g., "Which cone?"). In reply, the person will use one of several frames of reference similar to those used by the astronauts: egocentrically (e.g., "the cone to my right"), addressee-centered (e.g., "the cone to your right"), object-centered (e.g., "the cone in front of the box"), or exocentrically (e.g., "the northern most cone"). When such clarification is given, an additional proposition is provided to Polyscheme as shown in Table VI. Previous research has shown how we deal with the fifth deictic case (e.g., "The cone over [there] (points)") [48]–[51], so we will not discuss it further here.

TABLE VI
POLYSCHEME PROPOSITIONS REQUIRED TO RESOLVE SPATIAL RELATIONSHIPS

Right(ref, speaker,t0, wSpeaker)	Egocentrical reference; the reference object is currently to the right of speaker from the speaker's perspective.
Right(ref,robot, t0,wRobot)	Addresse-centered reference; the reference object is currently to the right of the robot from the robot's perspective.
Front(ref,box, t0,wBox)	Object-centered reference; the reference object is currently in front of the box from the box's perspective. Note: It is assumed that a box has a recognizable front.
Northern(ref, room,t0,R)	Exocentrical reference; the reference object is currently the northernmost one. Such references hold for all agents in the environment hence are true in the real world in Polyscheme.

The *spatial relationship specialist* considers the specified relationship with respect to all possible reference objects, i.e., Cones A and B. Based on the location of all objects in the environment and the location of the point of view specified in the relationship, the specialist is able to determine the truth value of each relationship. Given this extra information, Polyscheme is able to come to the correct conclusion.

Why doesn't the model ask for assistance in all situations? Many systems, when they recognize ambiguity and uncertainty, resolve the ambiguity by asking for additional information. However, this explicit request for additional information may be considered extraneous by the human and may reduce the effectiveness of the interaction. In addition, that request for additional information is dissimilar to how humans usually resolve this type of ambiguity. Previous work shows that given the principles of least effort and joint salience [57], the human would not ask for clarification in these cases. Given our emphasis on compatibility with humans, our system only asks for additional information when the situation is truly ambiguous (e.g., scenario 4, when there are two cones visible to both the human and the robot).

Note that there are many different ways to resolve ambiguity. For example, if an astronaut always used a specific wrench for a specific task, and the robot knew that the astronaut was working on that specific task, the robot could always hand the astronaut the correct wrench, regardless of whether the person could see the wrench or not. This type of procedural knowledge is not currently built into the system; here, we only concentrate on the use of visual perspective-taking and frame of reference. In the future, we will consider other methods by which humans resolve ambiguities.

TABLE VII
AMBIGUITY OF DIFFERENT SCENARIOS AS WELL AS THE SUCCESS RATE OVER FIVE TRIALS EACH

Scenario	Scenario Descrpt.	Ambiguity	Ambiguity Ratio	Success Rate
1	1 visible cone	None	1.0	100%
2	2 cones, 1 1 hidden from human	Medium	3.2	100%
3	1 cone, visible only to human	Medium	2.8	100%
4	2 cones, both visible	High	4.2	100%

An online video of scenario 2 (two cones, one hidden from human) can be found at http://www.aic.nrl.navy.mil/~trafton/movies/perspective-2objects-mp4.mov and an online video of scenario 3 (single cone, visible only to human) can be found at http://www.aic.nrl.navy.mil/~trafton/movies/perspective-hidden object-mp4.mov.

VIII. SYSTEM PERFORMANCE

How well does the system perform, and how does the reasoning system do when ambiguity and uncertainty increase? To explore this issue, we coded each scenario as having no ambiguity (scenario 1), a medium amount of ambiguity (scenarios 2 and 3), or a high amount of ambiguity (scenario 4). We computed ambiguity ratios by taking the simplest, least ambiguous case (scenario 1) and examining how much more time it took as complexity increased. As Table VII shows, there is an increase in Polyscheme's runtime as ambiguity and uncertainty increases.

This increase in computational time did not, however, affect the overall success rate. In our analysis of the system, we ran the full system through each of the four scenarios five times each. Of the 20 runs we performed during testing, there were no erroneous situations. Additionally, even though time increased as ambiguity and uncertainty increased, the overall system performance stayed manageable: reasoning time was never more than 36% of the overall system performance. Across all four scenarios and all 20 runs, we examined the amount of time to perform perception (e.g., finding cones and the box), reasoning (e.g., using perspective-taking to determine which cone the person is talking about), and navigation (e.g., moving to the cone). The perception component accounted for 34% of the overall system time, reasoning accounted for 26% of the overall system time, and navigation accounted for 40% of the overall system time.

In summary, when uncertainty and ambiguity increase, computing time also increases to resolve that ambiguity. However, the increase in ambiguity did not affect success rate over 20 trials, nor did it make the overall system excessively slow, even in the most ambiguous case.

Even though our system did not make any errors, there are several possible types of errors that could occur. First, the performance of our perception system is dependent on proper calibration of the color blob tracking. If the light conditions change, the system might experience decreased performance, both due to false positives (mislabeling objects, detecting additional objects, etc.) and false negatives (missing objects). A false representation of the environment could render Polyscheme incapable of reaching a correct decision. A false environmental representation would also interfere with more traditional robotic problems such as obstacle avoidance, path planning, or localization.

We believe that our system can scale up well, as evidenced by the different types of scenarios and the robustness with which it performed. Our system has not been tested with a large number (e.g., hundreds) of objects, however. Many objects would probably cause the system to slow down, so more optimal algorithms may be needed. In other words, in order to scale up 100 orders of magnitude, our current AI algorithms would probably need to be optimized. There are, of course, other methods of modeling perspective-taking (e.g., [42]), which may have different computational properties. However, we believe that the core ideas as well as many of the algorithms will be robust.

IX. CONCLUSION

This paper makes several contributions to human–robot interaction. First, the importance of perspective-taking in human–human interaction was shown in a nontrivial, real-world domain where it is expected that robots will soon be part of the team.

Second, we have outlined three important conceptual guidelines and a corollary for building robotic systems that interact with people. The first is to make the cognitive systems of robots similar to those of humans when it will aid in human–robot interaction. We have supported this guideline by focusing on cognitively plausible simulations for perspective-taking for robots. The second of these guidelines is to build cognitive robotic systems that are integrated across perception, cognition, and action. In fact, almost every current cognitive architecture [16], [17], [33], [44], [58] is integrated across a number of levels (though where that integration occurs is, of course, subject to some debate). Our third guideline is that building models of human–robot interaction based on human–human interaction will result in good design heuristics throughout the project. So far, this principle is still a hypothesis; not enough evidence has been gathered or systems built to adequately evaluate how veridical it is. Our corollary guideline focuses on perspective-taking per se and suggests that, since people use simulations to take others' perspectives, models of perspective-taking should as well. Our computational cognitive models do exactly that in theory as well as in practice.

One of the most difficult aspects of human–robot interaction has been to deal with the collaboration issue: when do you collaborate, when do you ask for help, and how do you respond to assistance. Our system takes a large step for answering these

questions. We collaborate when explicitly asked ("Go to the cone, coyote"). However, we do not request new information about every single decision that must be made: if our system can determine how to help, it does (e.g., it does not ask for assistance if it can resolve its uncertainty on its own). Finally, we have shown that our system can respond to a variety of frames of reference, including egocentric, exocentric, addressee-centered, object-centered, and deictic.

We have also presented a full instantiation of these ideas within a computational system (Polyscheme) and on a working robotic system (Coyote). The system is robust and has been demonstrated on a number of different tasks (additional demonstrations are described in other work [43]). An extremely important aspect of the overall system is that it makes increased complexity of tasks possible between humans and robots because every little detail does not need to be explained or thought-through in advance. Finally, the amount of integration in the full system is substantial. We have a working system that integrates perception, language understanding, problem solving, and spatial reasoning on an embodied robot. This work is a large step toward making a robot a true collaborator.

ACKNOWLEDGMENT

The authors would like to thank W. Adams for his help with the robot experiments and three anonymous reviewers for their comments about the research and suggestions for improving this paper.

REFERENCES

[1] D. Norman, *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Cambridge, MA: Perseus, 1993.
[2] H. Petroski, *The Evolution of Everyday Things*. New York: Vintage, 1994.
[3] J. Rubin, *Handbook of Usability Testing*. New York: Wiley, 1994.
[4] A. D. Galinsky and G. Ku, "The effects of perspective-taking on prejudice: the moderating role of self-evaluation," *Personal. Social Psych. Bull.*, vol. 30, pp. 594–604, 2004.
[5] D. J. Laible and R. A. Thompson, "Attachment and emotional understanding in preschool children," *Develop. Psych.*, vol. 34, pp. 1038–1045, 1998.
[6] B. Tversky, P. Lee, and S. Mainwaring, "Why do speakers mix perspectives?," *Spatial Cogn. Computat.*, vol. 1, pp. 312–399, 1999.
[7] K. Emmorey, B. Tversky, and H. A. Taylor, "Using space to describe space: Perspective in speech, sign, and gesture," *Spatial Cogn. Computat.*, vol. 2, pp. 157–180, 2000.
[8] M. Hegarty and D. Waller, "A dissociation between mental rotation and perspective-taking spatial abilities," *Intelligence*, vol. 32, pp. 175–191, 2004.
[9] H. A. Taylor and B. Tversky, "Spatial mental models derived from survey and route descriptions," *J. Memory Lang.*, vol. 31, pp. 261–292, 1992.
[10] N. Franklin, B. Tversky, and V. Coon, "Switching points of view in spatial mental models," *Memory Cogn.*, vol. 20, pp. 507–518, 1992.
[11] J. Huttenlocher and L. Kubicek, "The coding and transformation of spatial information," *Cogn. Psych.*, vol. 11, pp. 375–394, 1979.
[12] J. Huttenlocher, N. Newcombe, and E. H. Sandberg, "The coding of spatial location in young children," *Cogn. Psych.*, vol. 27, pp. 115–147, 1994.
[13] N. Newcombe and J. Huttnelocher, "Children's early ability to solve perspective-taking problems," *Develop. Psych.*, vol. 28, pp. 654–664, 1992.
[14] R. Wallace, K. L. Allan, and C. T. Tribol, "Spatial perspective-taking errors in children," *Perceptual Motor Skills*, vol. 92, pp. 633–639, 2001.
[15] J. H. Flavell, E. R. Flavell, F. L. Green, and S. A. Wilcox, "The development of three spatial perspective-taking rules," *Child Develop.*, vol. 51, pp. 356–358, 1981.
[16] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: an architecture for general intelligence," *Artif. Intell.*, vol. 33, pp. 1–64, 1987.
[17] A. Newell, *Unified Theories of Cognition*. Cambridge, MA: Harvard Univ. Press, 1990.
[18] J. E. Laird, "It knows what you're going to do: adding anticipation to a Quakebot," in *Proceedings of the Fifth International Conference on Autonomous Agents*, J. P. Muller, E. Andre, S. Sen, and C. Frasson, Eds. Montreal, QC, Canada: ACM Press, 2001, pp. 385–392.
[19] W. Bluethmann, R. D. Ambrose, M. S. Askew, E. Huber, M. Goza, F. Rehnmark, C. Lovchik, and D. Magruder, "Robonaut: A robot designed to work with humans in space," *Auton. Robots*, vol. 14, pp. 179–197, 2003.
[20] W. J. M. Levelt, "Some perceptual limitations on talking about space," in *Limits in Perception*, A. J. van Doorn, W. A. van der Grind, and J. J. Koenderink, Eds. Utrecht: VNU Sci. Press, 1984, pp. 323–358.
[21] L. A. Carson-Radvansky and G. A. Radvansky, "The influence of functional relations on spatial term selection," *Psych. Sci.*, vol. 7, pp. 56–60, 1996.
[22] L. A. Carson-Radvansky and G. D. Logan, "The influence of functional relations on spatial template construction," *J. Mem. Lang.*, vol. 37, pp. 411–437, 1997.
[23] S. Goldin-Meadow, "When gestures and words speak differently," *Current Directions Psych. Sci.*, vol. 6, pp. 138–143, 1997.
[24] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: Univ. Chicago Press, 1992.
[25] F. Mintz, J. G. Trafton, E. Marsh, and D. Perzanowski, "Choosing frames of reference: Perspective-taking in a 2-D and 3-D navigational task," in *Proc. Human Factors Ergon. Soc.*, New Orleans, LA, 2004.
[26] J. G. Trafton, S. B. Trickett, C. A. Stitzlein, L. D. Saner, C. D. Schunn, and S. S. Kirschenbaum, "The relationship between spatial transformations and iconic gestures," *Spatial Cogn. Computat.*, submitted for publication.
[27] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*, 2nd ed. Cambridge, MA: MIT Press, 1993.
[28] K. Emmorey, B. Tversky, and H. A. Taylor, "Using space to describe space: Perspective in speech, sign, and gesture," *Spatial Cogn. Computat.*, vol. 2, pp. 157–180, 2000.
[29] H. A. Taylor and B. Tversky, "Perspective in spatial descriptions," *J. Mem. Lang.*, vol. 35, pp. 371–391, 1996.
[30] J. G. Trafton, A. C. Schultz, N. L. Cassimatis, L. M. Hiatt, D. Perzanowski, D. P. Brock, M. D. Bugajska, and W. Adams, "Communicating and collaborating with robotic agents," in *Cognition and MultiAgent Interaction: From Cognitive Modeling to Social Simulation*, R. Sun, Ed. Cambridge, U.K.: Cambridge Univ. Press, to be published.
[31] E. M. Altmann and J. G. Trafton, "Memory for goals: An activation-based model," *Cogn. Sci.*, vol. 26, pp. 39–83, 2002.
[32] J. R. Anderson, F. G. Conrad, and A. T. Corbett, "Skill acquisition and the LISP tutor," *Cogn. Sci.*, vol. 13, pp. 467–505, 1989.
[33] J. R. Anderson and C. Lebiere, *Atomic Components of Thought*. Mahwah, NJ: Erlbaum, 1998.
[34] R. Shepard and J. Metzler, "Mental rotation of three-dimensional objects," *Science*, vol. 171, pp. 701–703, 1971.
[35] J. G. Trafton, S. S. Kirschenbaum, T. L. Tsui, R. T. Miyamoto, J. A. Ballas, and P. D. Raymond, "Turning pictures into numbers: Extracting and generating information from complex visualizations," *Int. J. Human Comput. Studies*, vol. 53, pp. 827–850, 2000.
[36] S. B. Trickett, R. M. Ratwani, and J. G. Trafton, "Real-world graph comprehension: High-level questions, complex graphs, and spatial cognition," under review.
[37] M. Skubic and G. Chronis, "Robot navigation using qualitative landmark states from sketched route maps," in *Proc. IEEE Int. Conf. Robot. Autom.*, New Orleans, LA, 2004, pp. 1530–1535.
[38] M. Skubic, S. Blisard, C. Bailey, J. A. Adams, and P. Matsakis, "Qualitative analysis of sketched route maps: translating a sketch into linguistic descriptions," *IEEE Trans. Syst., Man, Cyber. B, Cybern.*, submitted for publication.
[39] D. E. Meyer and D. E. Kieras, "A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms," *Psych. Rev.*, vol. 104, pp. 3–65, 1997.
[40] J. A. Fodor, *Modularity of Mind*. Cambridge, MA: MIT Press, 1983.
[41] M. D. Byrne and J. R. Anderson, "Serial modules in parallel: The psychological refractory period and perfect time-sharing," *Psych. Rev.*, vol. 108, pp. 847–869, 2001.

[42] L. M. Hiatt, J. G. Trafton, A. Harrison, and A. Schultz, "A cognitive model for spatial perspective-taking," in *Proc. 6th Int. Conf. Cogn. Model.*, M. Lovett, C. D. Schunn, C. Lebiere, and P. Munro, Eds., Pittsburgh, PA, 2004.
[43] N. L. Cassimatis, J. G. Trafton, M. D. Bugajska, and A. C. Schultz, "Integrating cognition, perception and action through mental simulation in robots," *J. Robot. Auton. Syst.*, vol. 49, no. 1–2, pp. 12–23, 2004.
[44] N. L. Cassimatis, "A cognitive architecture for integrating multiple representation and inference schemes," in *Media Laboratory*. Cambridge, MA: Mass. Inst. Technol., 2002.
[45] D. Gentner and A. Stevens, *Mental Models*. Mahwah, NJ: Erlbaum, 1983.
[46] P. N. Johnson-Laird and R. M. J. Byrne, *Deduction*. Mahwah, NJ: Erlbaum, 1991.
[47] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, pp. 832–843, 1983.
[48] D. Perzanowski, A. Schultz, and W. Adams, "Integrating natural language and gesture in a robotics domain," in *Proc. IEEE Int. Symp. Intell. Contr.: ISIC/CIRA/ISAS Joint Conf.*, Gaithersburg, MD, 1998, pp. 247–252.
[49] D. Perzanowski, A. Schultz, W. Adams, and E. Marsh, "Using a natural language and gesture interface for unmanned vehicles," in *Proc. Soc. Photo-Opt. Instrum. Eng.*, vol. 4024, G. R. Gerhart, R. W. Gunderson, and C. M. Shoemaker, Eds., 2000, pp. 341–347.
[50] D. Perzanowski, A. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human–robot interface," *IEEE Intell. Syst.*, vol. 16, no. 1, pp. 16–20, Jan./Feb. 2001.
[51] D. Perzanowski, A. Schultz, W. Adams, M. Bugajska, E. Marsh, J. G. Trafton, D. P. Brock, M. Skubic, and M. Abramson, "Communicating with teams of cooperative robots," in *Multi-Robot Systems: From Swarms to Intelligent Automata*, A. C. Schultz and L. E. Parker, Eds. Amsterdam, The Netherlands: Kluwer, 2002, pp. 16–20.
[52] J. G. Trafton, A. C. Schultz, D. Perzanowski, W. Adams, M. D. Bugajska, N. L. Cassimatis, and D. P. Brock, "Children and robots learning to play hide and seek," under review.
[53] A. Schultz, W. Adams, and B. Yamauchi, "Integrating exploration, localization, navigation and planning with a common representation," *Auton. Robots*, vol. 6, pp. 293–308, 1999.
[54] J. Bruce, T. Balch, and M. Veloso, "Fast and inexpensive color mage segmentation for interactive robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, vol. 3, Takamatsu, Japan, 2000, pp. 2061–2066.
[55] D. Perzanowski, A. Schultz, W. Adams, and E. Marsh, "Goal tracking in a natural language interface: Toward achieving adjustable autonomy," in *Proc. IEEE Int. Symp. Computat. Intell. Robot. Autom.*, Monterey, CA, 1999, pp. 208–213.
[56] K. Wauchope, *Eucalyptus: Integrating Natural Language Input With a Graphical User Interface*. Washington, DC: Naval Res. Lab., 1994.
[57] H. H. Clark, *Using Language*. New York: Cambridge Univ. Press, 1996.
[58] D. E. Kieras and D. E. Meyer, "An overview of the EPIC architecture for cognition and performance with application to human–computer interaction," *Human–Comput. Interaction*, vol. 12, pp. 391–438, 1997.

J. Gregory Trafton received the B.S. degrees in computer science and psychology from Trinity University, San Antonio, TX, in 1989 and the M.S. and Ph.D. degrees in psychology from Princeton University, Princeton, NJ, in 1994.

He is a Cognitive Scientist with the Naval Research Laboratory, Washington, DC. His research focuses on human–robot interaction, interruptions and resumptions, and the cognition of complex visualizations.

Nicholas L. Cassimatis received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 2002

He is an Assistant Professor of cognitive science at the Rensselaer Polytechnic Institute, Troy, NY. He was a National Research Council Postdoctoral Associate at the Naval Research Laboratory through the summer of 2004. In his research, he designs intelligent systems and cognitive models that combine the benefits of multiple kinds of computational and representational mechanisms.

Magdalena D. Bugajska received the B.S. degree in mathematics and computer science with a minor in artificial intelligence and robotics from Colorado School of Mines, Golden, CO and the M.S. degree in computer science from George Mason University, Fairfax, VA.

She is a Computer Scientist with the Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC. Her research interests include evolutionary computation, cognitive modeling, and robotics.

Derek P. Brock received the M.S. degree in computer graphics and multimedia systems from George Washington University, Washington, DC.

He has been a Computer Scientist who has specialized in human–computer interaction research with the Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, since 1991. His current interests lie in the application of sound and models of human language use to the design of standard and novel user interfaces for computational systems.

Farilee E. Mintz received the B.A. degree in psychology from the University of Michigan, Ann Arbor, and is currently pursuing the M.S. degree in human factors and applied cognition at George Mason University, Fairfax, VA.

She is a Systems Analyst in the Intelligent Systems section, Naval Research Laboratory, Washington, DC. She works for ITT Industries, AES Division. Her research interests include perspective-taking and human–robot interaction.

Alan C. Schultz received the M.S. degree in computer science from George Mason University, Fairfax, VA, in 1988.

He is Head of the Intelligent Systems section, Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC. His research is in the areas of human–robot interaction, machine learning, autonomous robotics, and adaptive systems.