Communicating and collaborating with robotic agents

J. Gregory Trafton, Alan C. Schultz, Nicholas L. Cassimatis, Laura M. Hiatt, Dennis Perzanowski, Derek P. Brock , Magdalena D. Bugajska, and William Adams

Introduction

For the last few years, our lab has been attempting to build robots that are similar to humans in a variety of ways. Our goal has been to build systems that think and act like a person rather than look like a person since the state of the art is not sufficient for a robot to look (even superficially) like a human person. We believe that there are at least two reasons to build robots that think and act like a human. First, how an artificial system acts has a profound effect on how people **act toward** the system. Second, how an artificial system thinks has a profound effect on how people **interact with** the system.

How people act toward artificial systems

"Everyone" knows that computers have no feelings, attitudes, or desires. Most people do not worry about hurting a toaster's feelings or cursing at a VCR. However, in a surprising series of studies, Cliff Nass has shown that people in some situations do, in fact, treat computer systems as social entities. Nass has shown that it takes very little "social-ness" for a person to treat computers (including robots, AI programs, etc.) as social creatures.

For example, Nass and Moon (2000) examined people's application of social categories to computers. Nass and Moon (2000) compared users' interactions with two computer systems – a tutor and an evaluator – using different combinations of male and

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **2006**	2. REPORT TYPE	3. DATES COVERED **00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE **Communicating and collaborating with robotic agents**	5a. CONTRACT NUMBER
	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

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)**	**40**	

female voices. Even though the participants indicated that they knew they were interacting with a computer, and explicitly reported that the voice did not relate to the 'gender' of the computer, or even the computer programmer, there were distinct gender-related biases in the experiment data. The evaluator, whose job was to evaluate both the user and the tutor, was said to be less friendly when connected to a female voice than a male. Similarly, the tutor system was evaluated as more competent when praised by a male evaluator than a female evaluator (Nass & Moon, 2000). This application of social rules to computers, and similar studies involving ethnicity, politeness and personality, enforces Nass's hypothesis that humans treat computers as having social properties.

Nass has also conducted experiments showing that not only do humans transfer social properties to computers, but they also treat different computers as distinct social actors (Nass, Steuer, & Tauber, 1994). Nass et al. showed this by injecting notions of "self" and "other" into participants' interactions with different computer boxes and voice output. Interestingly, the participants associated this embodiment with the computer's voice output (i.e. one voice per social actor) as opposed to the physical computer. In other words, two voices on one computer was considered by the user as two different social actors; the same voice on two computers was considered to be the same actor both times.

In other experiments, Nass and his colleagues have shown that computers can elicit social behavior from humans without explicitly displaying emotions. Nass has also shown that people transfer social categories to computational systems, view computers as distinct social entities, and apply social behaviors to their conversations with artificial agents (Nass & Moon, 2000; Nass et al., 1994). In short, Nass and his colleagues have gathered strong evidence that with very minor social cues , people interact with computers the same way people interact with other people.

Nass' overall hypothesis and evidence have at least two implications for how people act toward robots and other artificial systems. First, it means that embodied artificial systems do not have to look like a person in order for people to act in a social manner toward the robot: subtle social cues can cause people to think of computers as social entities. It is not clear how human (or non-human) a robot needs to look in order to elicit social behavior (e.g., would a polite mound of "goo" elicit polite behavior?). Second, if robots act socially, people have a "built in" way of dealing with them – exactly how they would deal with another person.

How people interact with artificial systems

How do people perceive and interact with artificial systems? In most cases, people want the system to help them solve their task or problem while making no mistakes and being polite about it (see above). Our desire for this type of interaction has probably been influenced by popular robots like C3PO (from Star Wars), Data (from Star Trek), and even Robbie the Robot (from Forbidden Planet).

For example, movies and television often portray people interacting with robots as if they were human. They use normal conversation and other modalities of communication associated with humans, such as gestures. These robots refer to objects and have the near-perfect ability of recognizing these objects. Also, they are able to reason about space and time. In reality, however, the interaction humans have with mobile robots is closer to teleoperation – in which humans directly (or in some cases indirectly) control the robot's behavior.

Figure 1 shows the scale of human interaction with the robot as a continuum from teleoperation, where the human directly controls the robot's motions, to dynamic autonomy, where the robot can exercise its own initiative and set its own goals while collaborating with the human.[1]

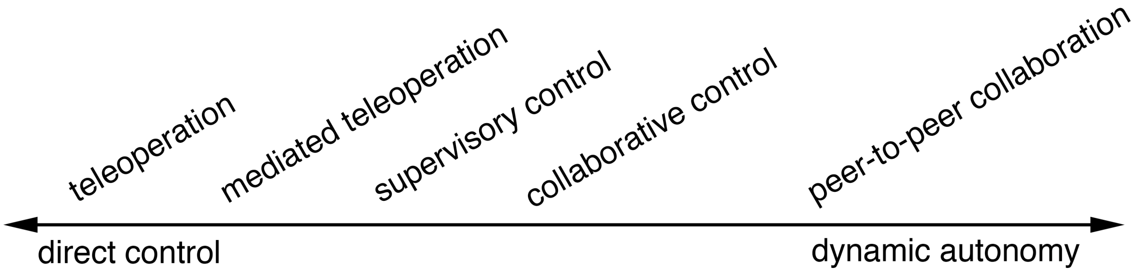

Figure 1: Levels of human interaction with autonomous system

Teleoperation requires that a human attend to the robot one hundred percent of the time. The human is completely responsible for all actions of the robot. Examples of robots that fall into this category include the robots used to help find victims and assess damage in the World Trade Center (WTC) collapse (Casper, 2002), and the small robots used by the U.S. Army in Afghanistan to explore caves. Teleoperation, however, can be very difficult. One of the main problems is ensuring that the human has enough awareness of the environment to understand the robot's position (Blackburn, Everett, & Laird, 2002). For example, rescue workers at the WTC had trouble determining if the robots were right side up with their camera view. Also, teleoperation requires a high-

[1] Various scales have been devised to show the level of autonomy of an unmanned vehicle, the best known being the Sheridan Levels of Autonomy (Sheridan, 1992). Figure 1 deemphasizes the notion of full autonomy that minimizes human interaction, and instead emphasizes the varying levels of collaboration, but in fact implies that the vehicle has the ability to operate autonomously.

bandwidth communications channel between the human and the robot in order to supply the real-time video.

By providing the robot with some basic skills, for example collision avoidance, the human is freed from having to control the vehicle at such a low level. This mode, mediated (also known as safe-guarded) teleoperation, allows the human to concentrate on other, higher-level decision making, such as choosing a path for the robot.

Moving further along the continuum, supervisory control gives the robot even more autonomy. Here the human picks one or more locations and other constraints (such as time), and the vehicle autonomously navigates to those waypoints. Now the human is freed from actually driving the vehicle and can concentrate on analyzing the robot's situation and making higher-level decisions. This level of interaction is particularly suited to very remote operations, such as the exploration of Mars during the Mars Pathfinder and Mars Exploration Rover missions, because the lag in round-trip communications does not support the quick execution of a human's decisions or for scientists and controllers to get real-time video.

Moving along the scale towards collaboration, the interactions become more complex and require that the human and the robot share more common knowledge about the world and about how things within the environment are related. In order to achieve these kinds of interactions and knowledge, the robot and the human must participate in a dialog to achieve common goals. Collaborative control refers to the ability of the robot and the human to ask each other for help in completing a task (Fong, Thorpe, & Bauer, 2002).

This level of interaction requires mixed initiative, or the ability of any agent in a collaborative act to initiate action in solving a task. In other words, each participant takes advantage of unique skills, location, and perspective of the current situation. We believe that at this level and beyond, the robot should utilize representations and procedures that are similar to those used by humans, rather than the other way around, in order to collaborate successfully; this is called the representational hypothesis. There are at least three reasons why a system with human-like representations and procedures will collaborate better with a person than a system that does not have human-like representations and procedures.

First, since algorithms written for traditional real-time robotic systems have to be computationally efficient, they tend to use efficient mathematical representations, such as matrices and polar coordinates, which may not be natural, or at best are extremely cumbersome, for people to use. For example, most position and motion information in robotics is conveyed using position vectors and transformation and rotation matrices. In general, people do not think or reason in this format. Instead, people seem to use a combination of spatial and propositional knowledge (Anderson, Conrad, & Corbett, 1989; Anderson & Lebiere, 1998; Shepard & Metzler, 1971; Taylor, 1992; Trafton et al., 2000; Trickett, Ratwani, & Trafton, under review). Thus, in order to interact with a human, the system must translate the robot's representation to the person's representation. However, because a person's representation of space is so complex (Harrison & Schunn, 2002, 2003a, 2003b; Previc, 1998), this is not a trivial task. Another, more functional argument, is that traditional AI spatial reasoning techniques do not adequately capture how people perform spatial reasoning; a model based on human spatial reasoning will provide some robust advantages over those systems that do not reason as a person would.

Second, if a human is going to collaborate in shared space with a robot, the robot should not exhibit unexpected, unnatural, or "martian" behaviors (Petty, 2001). While the robot may be able to perform a task efficiently, using for example a behavior-based approach, if the resulting behavior is perceived to be unnatural by the human, further interaction suffers as a kind of cognitive disruption. From this it follows to create some robot behaviors by modeling how humans perform such tasks.

Finally, we believe that some tasks for robots can best be programmed not by using more traditional control algorithms, but by understanding how humans solve the task and then creating a computational model of that understanding. So, for example, a robot that could search for hidden snipers would probably perform best if it had been programmed knowledge about how humans hide.

Two reasons for building artificial systems that think and act like a person have been presented. First, systems that act like people will elicit more social behaviors from people and make such systems more natural for people to deal with, and second, artificial systems that think like a person will interact with people with far greater ease than systems that do not. Our specific interest is in how to build robots, so the remainder of our discussion will focus on robotic agents. One issue with working with physically embodied robots is that, because they are physical and move around, people must interact with them in non-trivial ways: social interaction will probably occur, and communication and collaboration should occur. Our overall goal is to build robotic systems that think and act like people do in order to enable natural social behavior and allow better and easier communicative and collaboration. It should be noted, however, that our primary point can be generalized to all types of physically embodied systems.

Task domains

In the following sections, our robotic system will be described and three examples of work in our lab that show humans and robots collaborating and working together on various tasks will be presented. In the first example, the robot is taught how to hide (based on data obtained from a 3 1/2 year old child's behaviors in learning how to hide) and then it is asked to seek using these representations and strategies. The second and third examples use perspective-taking situations to facilitate human-robot communication and interaction. The first model of perspective taking emphasizes a good cognitive model of the representation used by humans, and the second perspective-taking model emphasizes the human process of using mental simulations to imagine another's perspective.

Since robots will be used for all these tasks, our mobile robots and their capabilities and sensors will be described first.

Mobile Robots

The empirical results were obtained by running the computational cognitive models, along with more traditional, reactive control software, on an indoor mobile robot in a laboratory environment.

Hardware

The robot is a commercial Nomadic Technologies Nomad200 suited to operation in interior 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.

Software

A combination of non-cognitive methods (primarily for robot mobility and object recognition), cognitively-inspired interactions (primarily for communicating with a person), and computational cognitive models (primarily for the high-level thinking and reasoning) were used. In previous work the utility of combining low-level reactive systems with cognitive models has been shown (Bugajska, Schultz, Trafton, Mintz, & Gittens, 2001; Bugajska, Schultz, Trafton, Taylor, & Mintz, 2002; Trafton, Schultz, Bugajska, Gittens, & Mintz, 2001).

Non-cognitive Methods:

This project draws on the robot mobility capabilities of the previously developed WAX system (Schultz, Adams, & Yamauchi, 1999), which includes components for map building, self-localization, path planning, collision avoidance, and on-line map adaptation in changing environments. The robot's lowest level of information comes from a dead-reckoning component that integrates motion over time to compute the robot's current location. As the robot moves, it gathers range data from its 16 ultrasonic transducers and a laser-based structured light rangefinder. In a process developed by Moravec and Elfes (Moravec & Elfes, 1985), the range data is interpreted using a sensor model that converts the raw range data to a set of occupancy probabilities for the sensed area. In this manner, data from multiple sensors can be fused into a single short-term occupancy map of the robot's vicinity, represented as a three dimensional array of discrete cells, each containing the probability that it is occupied or empty.

Robot odometry suffers from gradual drift, sometimes punctuated by larger errors from wheel slippage, rough ground, or collisions, so odometry alone is insufficient. Using the process of continuous localization (CL) (Schultz & Adams, 1998), a temporally overlapping progression of short-term perception maps is maintained. At periodic intervals, the oldest short-term map, which has the most data, is registered against a long-term map of the larger environment (typically a room) to determine the correction needed to correct the odometric drift. The long term map can be supplied a priori, or learned through a careful exploration, as was done in (Yamauchi, Schultz, & Adams, 1998). For this work, mapping was not the focus, so an a priori map was used. As a byproduct of correcting odometry, the long-term map can also be adapted to incorporate the now-corrected new readings from the short-term map. Thus, as the robot moves, it not only maintains an accurate estimate of its position but also keeps the long-term map up to date with any changes to the environment.

Because the robot's basic motor system is geometry-based and metric maps can be easily produced, it is a matter of practicality to state goal locations as points in Cartesian space. These goals are passed to the Trulla path planner (Hughes, Tokuta, & Ranganathan, 1992), which uses the long-term map to determine the best path to the goal. Because there may have been changes to the environment that are beyond the robot's sensor range, or recent changes such as people walking near the robot, the paths made by

Trulla cannot be followed blindly. Instead, they are passed as a single vector field to the Vector Field Histogram (VFH) process (Borenstein & Koren, 1991). VFH uses the robot's current position to retrieve from the vector field the direction the robot should move to head toward the goal. This vector is compared to an occupancy histogram built from the short-term map (which has the recent data close to the robot), and the robot is steered in the unblocked direction closest to the one indicated by the vector. In effect, Trulla handles the room-level navigation while VFH provides collision avoidance. If the robot is blocked, VFH prevents collision. CL learns the changes and produces a new adapted long-term map, and Trulla replans around the obstruction.

In addition to general mobility, the robot needs to recognize objects in its environment for the high-level cognition that will be demonstrated below. Rather than providing the robot with a priori information about discrete objects, the robot is instead equipped with limited computer vision in order to detect some objects autonomously. This also allows objects to be rearranged, added, or removed with the robot reacting accordingly. The CMVision package (Bruce, Balch, & Veloso, 2000) was used to provide simple color blob detection using an inexpensive digital camera mounted on the robot.

Relevant objects in the environment are tagged with color markers that are easily distinguished from the surroundings. The marker color is the identifier for the characteristics of an object. For example, all lime green objects are "chairs" and have the same characteristics. The bearing to the object is then determined from its location in the camera image, and the range to it is obtained from a scanning laser rangefinder.

Cognitively Inspired Methods

In order to communicate with a person, several methods that have some basis in human cognition are used. The methods that are used here allow a user to communicate with the robot using spoken language, gestures in the real world, and gestures on a Palm Personal Digital Assistant (PDA).

The human user can interact with the mobile robot using natural language and gestures that are part of our multimodal interface (Perzanowski, Schultz, & Adams, 1998; Perzanowski et al., 2002; Perzanowski, Schultz, Adams, & Marsh, 1999, 2000; Perzanowski, Schultz, Adams, Marsh, & Bugajska, 2001). The natural language component of the interface uses a commercial off-the-shelf 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 (Wauchope, 1994), to produce a regularized expression. This latter representation is linked, where necessary, to gesture information, and an appropriate robot action or response results.

For example, the human user can tell the robot "Coyote, go hide and I'll try to find you." The speech signal is analyzed into a text string which when parsed produces the following representation, simplified here for expository purposes.

```
(and (imperative (p-hide: hide)
             (system: you
 (name: coyote)))
    (future (p-attempt: try)
           (agent: I)
```

```
          (action (p-find: find)
                           (agent: I)
                       (system: you
       (name: coyote)))))
```

Basically, Nautilus parses the utterance into appropriate commands (e.g. the imperative structure in our example) and statements (e.g. the future declaration in our example), and the various verbs or predicates of the utterance (e.g. hide, try, and find) are mapped into corresponding semantic classes (p-hide, p-attempt, and p-find) that have particular argument structures (agent, system) which result in a semantic interpretation of the utterance. With gesture information, where appropriate, a combined representation incorporating both the linguistic and gestural information is then sent to the robotic component whose modules translate the representations into appropriate actions.

In the example above, no further gesture information is required to complete the command. Coyote will, therefore, respond “I will go and hide,” in order to inform the user that it has understood the utterance. The appropriate behavior based on the cognitive model for the hide-and-seek activity is invoked and appropriate robot action according to the model ensues.

If a gesture is required to disambiguate the speech, as in “Coyote, hide somewhere over there,” the gesture information obtained from the laser rangefinder mounted on the top of the robot indicates the desired location, and this information is included in the interpreted utterance for further analysis by the robotic system.

Hide and Seek

The first domain in which robotic agents that think and act like people will be demonstrated will be the children's game commonly known as “hide and seek.” Hide and seek is a simple game in which one child is "It," stays in one place counting to ten with eyes closed, and then goes to seek, or find, the other child or children who have hidden. This game allows us to address our high-level goals of understanding how human representation and processing of spatial information (Skubic, Perzanowski, Blisard, Schultz, & Adams, in press) can aid in designing better human-robot interaction in collaborative spaces. This work is described more fully elsewhere (Trafton, Schultz et al., under review); a summary of our findings is discussed here.

The study had two primary goals: 1) to understand how children learn to play hide and seek via computational cognitive modeling; and 2) to build a system that thinks and acts like people do. This latter point should serve to facilitate human-robot interaction. The first point will be briefly summarized and more fully described to show how our system thinks and acts like children learning how to play.

Hide and seek game-playing behavior was gathered from a 3 1/2 year old child. Previous research suggests that 3 1/2 year old children do not, in general, have perspective-taking ability (Huttenlocher & Kubicek, 1979; Newcombe & Huttnelocher, 1992; Wallace, Allan, & Tribol, 2001), but they are able to play a credible game of hide and seek (supported mostly by anecdotal evidence of the game-playing behavior at local parks and playgrounds, since there are almost no empirical investigations of the naturalistic game of hide and seek). Spatial perspective taking is clearly needed for a “good” game of hide and seek: a good hider needs to take into account where "It" will come into a room, where "It" will search first, and where to hide behind an object taking the perspective of "It" (Lee & Gamard, 2003) so that “It” will not be able to find the hider easily. Additionally, the hider must know that just because the hider can't see "It" doesn't mean that "It" can't see the hider. The research question was to explore how 3 1/2 year old children learned to play hide and seek without perspective taking. The hypothesis (which was supported by computational simulation) was that 3 1/2 year old children were able to learn relationships of objects to play hide and seek. For example, a child may learn that hiding under or inside of an object was a good hiding place. In contrast, hiding behind an object occurred rarely because that required spatial perspective taking. Evidence was obtained from a child learning to play hide and seek; subsequently, computational simulations in ACT-R (Anderson & Lebiere, 1998) were written that learned how to play hide and seek in the same manner as the child did. Additionally, the computational system was put on our robot and hide and seek was played (Trafton, Schultz et al., under review) with it.

In order to show the benefits of a system that thinks and acts like a person, we wanted to show how the computational system could be generalized to a different situation where similar but not exact knowledge would be needed. The most obvious task to explore was the “seeking” part of hide and seek, since the computational cognitive model that was written focused solely on learning how to hide. The seeking system should exhibit several interesting behaviors. First, it should seek according to its own model of hiding. That is, it should search in places that it thinks are plausible for “It” to hide in.[2] Second, it should be able to deal with novel objects or objects that were not in its original environment. Third, it should be able to accomplish this seeking behavior without new learning mechanisms while using its current representations and algorithms. This seeking behavior would be a proof of concept for the representational hypothesis: building a system that thinks and acts like a person would make the system more "natural" in some ways. In this case, a child would presumably find a system that plays hide and seeks like another child more fun than a system that hides or seeks in very odd places (e.g., a robot that hid in a very difficult location would not be much fun to play with).

In order to explore how our existing system would seek for a person after it had learned how to hide, several straightforward steps were gone through. First, the model was run as above, allowing the to learn different pertinent features of objects and object-relations. The model was then "frozen." In order to allow the robot to seek, two more pieces of information were given to it: (1) what a person "looked like" (e.g., the person might wear a blue shirt which was identifiable by CMVision) and (2) how to start the game (e.g., a location to start from; what to count to, etc.). In order to seek for a person, the computational cognitive model determined where it would best hide and then gave those coordinates to the robot where it would then look. If it did not find the person in that location, it searched in the next place that it would have hidden until either it had found the person or it had run out of locations to search. The model's "individual preferences" (e.g., locations that had higher or lower levels of activation) were not

[2] Because our robot cannot bend or change shape like a young child, as a simplification for both the model and the robot, we assumed that our hider is small (approximately the size of a small child) and does not contort itself a great deal or squeeze itself into a location that is smaller than itself.

cleared. The model searched those locations in approximate (because of noise) order of activation. The environment was changed slightly as well (i.e., added additional objects it already knew something about, moved the location of other objects, etc.).

Both the model and robot behaved as expected. The robot systematically searched different locations that it had learned were acceptable hiding places until it found the person hiding. Over multiple games, it searched locations in different orders. Most importantly, it did not attempt to search for a person in locations that would have been very "odd." For example, while it could have found a person hiding out in the open (like children do when they're first learning how to play hide and seek), it did not systematically search all the open space for a person hiding out in the open. Instead, the robot searched where it thought it would have hidden. A full set of movies of the robot seeking can be found at http://www.aic.nrl.navy.mil/~trafton/hideseek.html.

The fact that the robot and computational system were able to find a hiding person successfully by using its own representations and processes supports our representational-level hypothesis; namely, a computational or robotic system that thinks and acts like a person will interact well with the person. This hypothesis was supported by taking the "hiding" model and applying it to seeking. The model successfully searched for a person using the same representations and processes that it had learned and used while learning how to hide. Our hypothesis also states that by using similar representations and processes, alien behaviors could be avoided.   As shown above, our system did not search for or hide in unusual places; instead, it only considered those places that a human would consider.

Clearly, our approach could lead the system to make systematic errors: it would not expect a person to have climbed a rope and clung to it, etc. It also could not use perspective taking for seeking or even assume that the hider would move locations because that information was not built into the original hiding model. However, 3 1/2 year olds do not typically climb ropes or use perspective taking to hide from someone, and they do not typically look for hiders in these types of odd places, either (Trafton, Schultz et al., under review).

Perspective Taking

Our second and third domains for exploring robots that think and act like people involve the basic cognitive skill of perspective taking.

Imagine two astronauts working together on a collaborative construction project. While they might be able to talk and gesture to each other to get their job done, they would be dressed in full spacesuits and consequently have diminished perceptual abilities and decreased freedom of movement. Given these limitations, their work could be facilitated by a robotic system that could hand them tools and follow simple instructions, or perhaps even give them instructional assistance. In order to determine the kinds of instructions and utterances the robots would need to understand and process in this situation, we have analyzed data that was collected during a specific astronaut training session. When astronauts train for missions, part of their training occurs in various simulated microgravity environments, such as the Neutral Buoyancy Laboratory (NBL) at NASA/JSC. In the NBL, astronauts conduct a wide variety of training for extravehicular activity (EVA); i.e., working outside the space shuttle, including working out the procedures and defining roles to perform EVAs.

One issue that astronauts must deal with is spatial language and spatial perspective taking. 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”) (Carson-Radvansky & Logan, 1997; Carson-Radvansky & Radvansky, 1996; Goldin-Meadow, 1997; Levelt, 1984; McNeill, 1992; Mintz, Trafton, Marsh, & Perzanowski, in press). Unfortunately, astronauts must deal with frames of references and spatial situations that people here on Earth do not typically have to deal with. For example, “up” may mean something completely different in space in different situations (i.e., up may mean toward the ceiling of the spaceship rather than with reference to the normal sense of gravity here on Earth). In general, astronauts do not have problems themselves in understanding the spatial language and taking another’s point of view, but one of the challenges for robotic systems is to understand what someone else is talking about from a different spatial perspective.

As part of this project a series of astronaut utterances has been analyzed as they performed a cooperative assembly task for Space Station Mission 9A, specifically the construction of the first right-side Truss segment and the Crew and Equipment Translation Aid (CETA) Cart A in the NBL (Trafton, Cassimatis et al., under review). This analysis project is still under progress, but several critical issues have already surfaced. First, astronauts seem to switch reference frames quite often, just as people do while giving directions (Franklin, Tversky, & Coon, 1992). Second, astronauts in this collaborative process must frequently take another's perspective, even when they cannot see the person whose perspective they are taking. For example, the following conversation (Table 1) occurred between three individuals--two astronauts (EV1 and EV2) in the neutral buoyancy tank at NBL and one person (Ground) outside of the tank in mission control. The latter watched the two astronauts through a video feed of the activity.

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.		

Table 1: Dialog between two astronauts and an observer

Notice several things about this conversation. First, the mission control person mixes reference frames from addressee-centered ("by your right hand") and exocentric ("straight nadir" which means towards the earth) in one instruction, the very first utterance. 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.

This short excerpt shows that an automated reasoning system needs to be able not only to mix perspectives, but to do so in a rather sophisticated manner. One of the most difficult aspects of this problem is the addressee-centered point of view, which happens quite often in the corpus that was examined. Thus, in order for a robotic system to be truly helpful, it must be able to take into account multiple perspectives, especially another person's perspective.

At this point we turn to a discussion of two further projects that show how robots can think and act like people. The first project uses similar processes (specifically simulation) that people use when they take another person's perspective, and the second project uses the same spatial representations that people use.

Perspective taking using similar processes: Polyscheme

Our hypothesis that humans and robots interact better when they share similar representations and when robots can take the perspective of humans has helped determine how to implement the cognitive subsystem of our robots. First, since robots must share similar representations with humans, a cognitive architecture that had cognitively-inspired spatial and logical reasoning mechanisms was used. Second, an architecture that provides a mechanism for simulating alterative states of the world was used so that the robots could reason about the perspective of other people. The Polyscheme (Cassimatis, 2002) cognitive architecture fulfills both requirements.

Polyscheme is a cognitive architecture based on the ability to conduct mental simulations of past, future, distant, occluded and/or hypothetical situations. Our approach has been to use Polyscheme to enable robots to simulate the world from the perspective of people with whom they are interacting and to understand and predict the actions of humans.

Polyscheme uses several modules, called specialists, which use specialized representations for representing some aspect of the world. For example, Polyscheme's space specialist uses cognitive maps to represent the location of and spatial relations among objects. Its physics specialist uses causal rules to represent the causal relationship between events. Using these specialists, Polyscheme's specialists can simulate, i.e., represent the state and predicted subsequent states, of situations it cannot see at present, either because they occurred in the past or future, they are occluded from view and/or they are hypothetical.

Polyscheme modelers have the ability to set strategies for choosing which situations to simulate in what order. Modelers use these strategies to implement reasoning and planning algorithms, including perspective taking. For example, the counterfactual simulation strategy, "when uncertain about A, simulate the world where A is true and the world where A is false", implements backtracking search when used repeatedly. The stochastic simulation strategy, "when A is more likely to be true than false, simulate the world where A is true more often than the world where A is false", implements an approximate form of probabilistic reasoning (often used, e.g., to estimate probabilities in a Bayesian network). Polyscheme's ability to combine multiple simulations from multiple strategies and to share simulations among strategies is the key to its ability to tightly integrate multiple reasoning and planning algorithms (Cassimatis, Trafton, Schultz, & Bugajska, 2004). Since each simulation is conducted by specialists that use multiple representations (e.g., perceptual, spatial, etc.), the integration of reasoning with sensation and multiple forms of reasoning is constant.

Using this framework, we have been able to improve human-robot interaction by giving robots the ability to simulate the world from the perspective of humans they interact with. An important problem when humans and robots communicate using natural language is that most verbal commands or questions have multiple literal meanings. Although humans are normally able to use contextual information to eliminate most possible interpretations and thus identify the speaker's intent, this has remained a difficult problem for computers and hence robots.

By using Polyscheme to implement the perspective simulation strategy, "when a person, P, takes action, A, at time, T, simulate the world at time T from A's perspective", we have given our robots the ability to reason about the world from the perspective of people and to thereby disambiguate their utterances. In many cases, for instance, an utterance is ambiguous given the listener's knowledge, but unambiguous given the speaker's knowledge. Figure 2 is an example. The figure shows a robot and a person facing each other. The robot can see that there are two cones in the room, cone1 and cone2, but the person only knows about cone2 because cone1 is hidden from her. When the person commands, "Robot, go to the cone", the phrase "the cone" is potentially ambiguous to the robot because there are two cones, though unambiguous to the person because he only knows of the existence of one cone. Intuitively, if the robot could take the perspective of the person in this task, it would see that, from that perspective, cone2 is the only cone and therefore "the cone" must refer to cone2.

Polyscheme was used to implement this sort of reasoning on the robot described earlier. The following list outlines the sequence of simulations that enable the robot to properly disambiguate the person's utterance:

- **Simulate current real world (i.e., perceive it):**
    - Perception specialist notices the existence and location of person, cone1, cone2 and obstacle.
    - Language specialist hears "Coyote, go to the cone" and infers that there is an object, C, that is a cone and that the person wants it to go to.
    - Identity hypothesis specialist infers that C can be identical to cone1 or cone2:
        - C = cone1, C = cone2
    - Identity constraint specialist notices a contradiction.
    - This contradiction triggers the counterfactual simulation strategy.
- **Simulate the world where C = cone1**
    - Since in this world Person has referred to cone1, the perspective-simulation strategy is triggered:
    - **Simulate the world where C=cone1 and Robot=Person**.
        - The spatial reasoning perspective indicates that cone1 does not exist in this world since person cannot see it.
        - Thus, C != cone1.
- **Simulate the world where C = cone2**
    - Since in this world Person has referred to cone2, the perspective-simulation strategy is triggered.

        - **Simulate the world where C=cone2 and Robot=Person**
            - Since cone2 is visible in this world, there is no contradiction in this world.
- Infer that C = c2, i.e., that "the cone" refers to cone2.

This example illustrates how robots can use their own mechanisms for reasoning about the world to reason about the beliefs and intentions of other agents without needing elaborate machinery for social reasoning. An online video of this example can be found at http://www.aic.nrl.navy.mil/~trafton/movies/perspective-2objects-mp4.mov.

Polyscheme is able to solve this problem by using mental simulation, a human-level ability that is, in general, not well used in other cognitive architectures. By using mental simulation (a similar mental process to what people do), it greatly increases the human-robot interaction in this situation: without this kind of machinery, the robot would need to ask "Which cone?" which could lead to confusion on the person's part if she did not know there was more than one cone. Other work not only provides a more complete description of Polyscheme, but also provides more details about other tasks, including the perspective-taking examples used here (Cassimatis, 2002; Cassimatis, Trafton, Bugajska, & Schultz, in press; Cassimatis et al., 2004; Trafton, Cassimatis et al., under review).

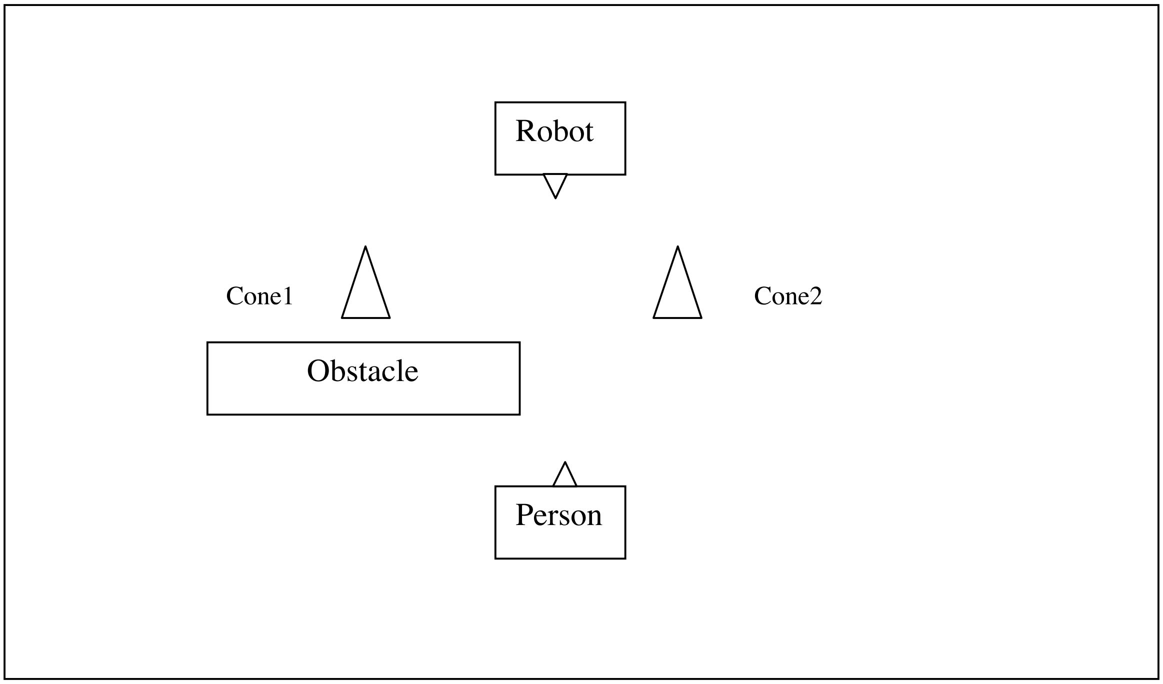

Figure 2. The robot needs to take the perspective of the person to determine to which cone the human has referred.

Perspective taking using similar representations: ACT-R

Polyscheme showed that mental simulation could be used to solve a problem in perspective taking, and shows an example of how to build a robot that thinks and acts like a person. Could a similar perspective-taking task be accomplished by focusing on the spatial representations that people have? We attempted to answer this question by using ACT-R/S (Harrison & Schunn, 2002, 2003a, 2003b).

The cognitive architecture jACT-R is a java version of the ACT-R architecture (Anderson & Lebiere, 1998). To represent declarative memory, it uses chunks of various types of elements. These chunks can be accessed through a memory retrieval buffer. In order to use and manipulate the chunks of memory, ACT-R provides a framework for production rules. A sample chunk and production rule are shown in Figure 3. ACT-R then simulates cognitive behavior and thought based on activation values and propagation of chunks and higher-level goals. ACT-R also includes support for perceptual and motor cognitive tasks, such as Precognitive Remote Perception tasks, by including a second visual buffer for viewing objects in space.

ACT-R/S extends jACT-R to implement a theory about spatial reasoning (http://simon.lrdc.pitt.edu/~harrison/). It posits that spatial representations of objects are temporary, egocentric and dynamically updated (Wang & Spelke, 2002). ACT-R/S has three buffers for spatial cognition: the configural buffer, the manipulative buffer, and the visual buffer. The configural buffer represents spatial extents of objects that are updated during self-locomotion and is used during navigation, path-computation, object-avoidance, etc. The manipulative buffer represents the metric spatial bounds of an object and is used for spatial transformations of objects (Trafton, Marshall, Mintz, & Trickett, 2002; Trafton, Trickett, & Mintz, in press). The visual buffer is the same as the "standard" perceptual-motor buffer in ACT-R/PM (Byrne & Anderson, 1998).

ACT-R/S represents objects using vectors to the visible sides of the object. It has the ability to track these objects through the configural buffer, a data structure analogous to the other buffers of ACT-R that stores each object once it has been identified. The coordinate vectors of the objects in the buffer are then dynamically updated as the agent moves throughout the spatial domain. The configural buffer, unlike the visual and retrieval buffers of ACT-R, can hold more than one object to account for the fact that animals have been shown to track more than one landmark at once while moving through the world (Harrison & Schunn, 2003a). In order to focus on the representational aspects of perspective-taking, our model uses only the spatial representations within jACT-R/S.

Using the configural extension begins with locating and attending to an object via the visual buffer provided by the standard Perceptual-Motor extension to ACT-R. Once an object is found, it is possible to request that the ACT-R/S-visual object at that location, if one exists, be placed in the configural buffer. The model then begins tracking this object, creating the initial vectors and updating them as the agent moves around in the world. The updating transformation is done by adding or subtracting vectors representing the agent's movement to the vectors and object's location.

```
chunk_cone:
        isa: cone
        color: gray
        speaker_can_see: true
        location: (x,y)

production_take_cone:
        if isa cone
                and speaker_can_see
                and (my_x, my_y) = (x,y)
        then take_cone
```

Figure 3: An ACT-R memory chunk and production rule.

In order to demonstrate the results of perspective taking using jACT-R/S, the same perspective-taking task that Polyscheme solved was implemented: disambiguating which cone a person referred to when the robot could see two cones but the person could only see one. For this example, the full system was not implemented on a physical robot. In the simulated world, two agents (hereafter referred to as the 'speaker' and the 'robot') are in a room with two cones and a screen. The screen blocks the view of one of the conees from the speaker, but not the robot. Then, the speaker asks the robot to hand them the cone, using some locative clue such as “in front of me.” If both of the conees match this description, then the robot should hand the speaker the cone that they know the speaker can see.

The model thus uses the ACT-R/S architecture in order to use spatial perspective taking to complete its task. There are several components to the perspective taking that it goes through in order to do so.

Perspective-taking process. The production rules involved in the perspective-taking process are the most important part of the model, as they implement the heart of its theory of spatial perspective-taking. Taking the perspective of someone at position and orientation B, from position and orientation A, the over all procedure is to:

1. Turn to face position B
2. Walk to position B
3. Face orientation B
4. Extract the desired information from the visual knowledge at this position and orientation
5. Face position A
6. Walk back position A
7. Return to orientation A.

The key to this process is that all of these movements – i.e. turning and walking – are mentally done by only transforming the configural buffer contents by the appropriate vector, leaving everything else the same. Thus, the physical location of the robot does not change; it is only its mental perspective that changes.

Initial scan for objects. The model first uses perspective taking to deduce where it should begin looking for the cone. When the speaker says 'in front of me', or ‘to my left’, etc., the robot interprets that information by taking the speaker’s perspective and mentally placing itself in their shoes. It then looks at a location in front of it, or to its left (as indicated by the speaker's initial instructions), and keeps track of that location as it returns to its own perspective. This is where it begins its search for the cone.

Deciding which cone to go to. The model also uses perspective-taking once a cone has been found. When it has located a cone in the desired location, it looks around for obstacles that could possibly block the speaker's view of the cone. If it finds any such obstacles, it takes the speaker's perspective again in order to judge whether or not it can see that particular cone.

This time, however, once the robot has taken the speaker's perspective, instead of turning to match the speaker's orientation, it turns to face the located cone. Determining whether or not the cone is visible by the speaker is then done by comparing the transformed location vectors of the target object with the location vectors of the possible obstacles, making sure that the obstacle's vectors do not completely surround the target object's vectors. This ensures that the speaker has the ability to see at least part of the cone.

If the speaker can in fact see the cone, the robot goes to that cone. If the speaker cannot see the cone, the robot continues to look for a cone that the speaker can Although building a model that completes this task could be done in a variety of ways, what distinguishes jACT-R/S from other spatial cognitive models is that it uses the spatial representation of humans in order to complete the perspective-taking task. Once again, this representation entails creating and updating a set of vectors to the edges of each object currently being attended to. Using this representation allows the cognitive agent to undergo perspective taking by imagining movement throughout the world by simply altering the representation of the objects in the configural buffer. This ultimately results in true perspective taking in the sense that the agent's representation of objects, once it has imagined movement to the second agent's location, roughly matches the second agent's own representation of these objects, truly seeing the world as the second agent does. In the end, this provides a more natural and human-like interaction with the second agent, since the cognitive agent responds as a human plausibly would instead of introducing into the conversation an item (here, a cone), that the second agent might now even know exists.

Summary of Perspective taking

When a task needs perspective taking, there are, of course, many ways to solve the task. For example, a straightforward method of solving the "Go to the cone" problem discussed above would be to simply ask the person "Which cone?" Alternatively, the robot could simply guess and go to a cone. Unfortunately, both these solutions break down under more complex conditions and under conditions where speed and accuracy are critical (like the astronaut construction task discussed earlier). Having a robot ask many questions would quickly get boring, bringing the level of autonomy to a level that hurts team performance. Similarly, if a robot is going to guess frequently, team performance will likely degrade and interaction with the robot will quickly become frustrating.

Using the forms of perspective taking that have been outlined here, we believe that we are building robots that think and act like people (to a limited degree). The main advantage of this approach is that if a robot thinks and acts like a person, not only will a person treat it (approximately) as a person, but also the interaction with the robot will be quite natural for the person.

Future directions in social perspective-taking

Our work on perspective taking attempts to create a robot that thinks and acts like a person; this presents several future research questions and opportunities that fall into two broad categories. The first involves improving robots' abilities to infer and represent the perspective of humans and the second pertains to actions that they can take to ensure that human and robot representations are synchronized and to make corrections should they begin to diverge.

In much of the work that has been described in this chapter, robots infer a human's perspective by observing which objects are currently visible to him from his perspective. There are several other factors robots can use to infer the human perspective, each of which enables them to coordinate their behavior with humans in more complex situations. These factors include perceptual salience, the history of a person's attentional gaze and predictions of future actions formed by predicting the intent of past actions.

Robots must not only be able to represent the perspective of a person, but also be able to identity which aspects of his perspective are most salient. Such a perceptual capacity in a robot would be valuable in many practical circumstances. Studies of human-human interaction have shown that people can make ambiguous references to objects that other people can easily disambiguate by choosing the most salient interpretation. Clark, Schreuder, and Buttrick (1983), for example, found that a group of thirty students individually made the same choice with an average of 70% or better when asked to either choose an ambiguous reference, choose what another person would choose, or simply choose what was most salient in various scenes of similar objects. In addition, it was found that the students' ratings of confidence in their choices correlated highly with the concurrence of their choices. In accord with our theme that robots with human-like representations will generate more predictable behavior and be easier to deal with, we suspect endowing robots with a sense of salience similar to that of humans will lead to more advanced human-robot interaction.

Robots must also be able to infer the perspective of a person, not only from his current spatial location, but also from the history of where he has been and what he has looked at. This kind of inference is such a fundamental part of what humans expect of an interaction, that it has been found to underlie the behavior of infants and very young children. For example, Baldwin (1991) has found that when toddlers are learning the name for an object, they do not merely associate the visual and auditory stimuli they are currently perceiving. Instead, they keep track of what a speaker was looking at while naming an object and attach the word he uttered to his object of attention even if they do not actually see the object until later. Wimmer and Perner (1983) have found that four year old children can predict the actions of another person based on what that person has seen in the past, even if that requires them to represent that another person has an incorrect view of the world. These studies indicate that humans have a basic ability to infer other people's perspective using not only information about what person is currently looking it, but by referring to the history of their interaction. We hope that endowing robots with this ability will enable them to interact in more complex tasks with people by needing less information and time to construct richer models of their joint activities.

In addition to using more information to "see" other people's perspectives, robots must constantly monitor how well synchronized their view of the world is with that of the people they are working with and take actions to correct these views when their views or representations diverge. There is extensive evidence that humans constantly engage in this behavior when interacting among themselves and we assume that they will expect the same of the robots with which they interact.

One simple strategy that people use to communicate that they understand each other is the use of "backchannel responses". For example, during conversations, people will nod their heads, smile or utter make brief utterances such as "uh huh" to indicate that they understand each other. These behaviors are not just occasional conversational ticks but are part of spectrum of behaviors that exhibit understanding that people expect and whose absence can lead to substantial miscommunication (Brennan, 1998; Brennan & Hulteen, 1995; Clark & Brennan, 1991). We believe that recent advances in the expressiveness of robots create an opportunity for the use of backchannel responses to make robots act even more like people than ever before. These types of backchannel responses, in fact, may very well be a primary way that robots can act like people and cause people to act toward robots in a social manner.

On many occasions, people take more overt actions to indicate how well synchronized their representation of the world is with the people they are cooperating with. In cases where a person wants to verify that he understands the intent of a speaker's utterance, he will reformulate the speaker's meaning with another utterance. For example, Clark and Wilkes-Gibbs (1986) found that in scenarios where a speaker attempted to refer to an object, the listener would sometimes find a new way of referring to the object and ask the speaker if this was his meaning, e.g., speaker A says, "Um, third one is the guy reading with, holding his book to the left," speaker B asks. "Okay, kind of standing up?" and speaker A answers. "Yeah." In cases where one person in a conversation detects a mismatch between the representations of the participants, he will initiate "repair" utterances to resynchronize the representations as in this example, again from Clark and Wilkes-Gibbs (1986):

A. Uh, person putting a shoe on.

B. Putting a shoe on?

A. Uh huh. Facing left. Looks like he's sitting down.

B. Okay.

These future research directions indicate that many superficially disparate aspects of interaction are all applications of the principle that humans and robots should share the same kinds of representations and should continually engage in activities to make sure these are synchronized. It also enables the large body of research in human-human interaction, especially including work that indicates what humans expect of those they interact with, to create systems that think and act like people.

Conclusion

The main point of this paper has been to present, explore, and support ways of building robots that think and act like people. The strongest examples have focused on how to build robots that think like people. We also presented a representational hypothesis – using similar representations and processes as a person will improve and facilitate interaction. This paper has shown three strong demonstrations of robots that think and act like people. First, we showed that a model of hiding could be used to seek. The model used the same representations and strategies to seek as to hide. These human-based representations and strategies allowed the robot to interact with a person without violating the person's expectations. Second, we showed two different perspective taking models that solved a complex task in different ways. The first model, written in Polyscheme, focused on mental simulation to solve the task. The second model, written in jACT-R/S, focused on the spatial representations that people are thought to have. Both models successfully solved the perspective-taking problem presented to it.

In sum, the systems presented here take seriously the idea that people can be used as models for computational systems, specifically robots. The two primary advantages that flow from this idea are 1) that people will act socially toward systems that act as a human would; and 2) that people will interact with a system that "thinks" like a person would.

Acknowledgements

This work was partially supported by the Office of Naval Research to Greg Trafton and Alan Schultz under work order number N0001402WX20374. Additional funding was supplied by DARPA IPTO under the MARS program. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U. S. Navy.

References

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science, 13*, 467-505.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Mahwah, NJ: Erlbaum.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development, 62*, 875-890.

Blackburn, M. R., Everett, H. R., & Laird, R. T. (2002). *After action report to the joint program office: Center for the robotic assisted search and rescue related efforts at the world trade center* (No. 3141). San Diego: U.S. Navy SPAWAR.

Borenstein, J., & Koren, Y. (1991). The Vector Field Histogam - fast obstacle avoidance for mobile robots. *IEEE Transactions on Robotics and Automation, 7*(3), 278-288.

Brennan, S. E. (1998). The grounding problem in conversation with and through computers. In S. R. Fussell & R. J. Kreuz (Eds.), *Social and cognitive psychological approaches to interpersonal communication* (pp. 201-225). Hillsdale, NJ: Lawrence Erlbaum.

Brennan, S. E., & Hulteen, E. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems, 8*, 143-151.

Bruce, J., Balch, T., & Veloso, M. (2000). Fast and inexpensive color mage segmentation for interactive robots. In *Proc. of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vol. 3, pp. 2061-2066). Takamatsu, Japan.

Bugajska, M. D., Schultz, A. C., Trafton, J. G., Mintz, F. E., & Gittens, S. (2001). Building adaptive computer generated forces: The effect of increasing task reactivity on human and machine control abilities. In *Late-Breaking Papers at the 2001 Genetic and Evolutionary Computation Conference (GECCO 2001)*. San Francisco, CA.

Bugajska, M. D., Schultz, A. C., Trafton, J. G., Taylor, M., & Mintz, F. E. (2002). A hybrid cognitive-reactive multi-agent controller. In *Proceedings of 2002 IEEE/RSJ International conference on Intelligent Robots and Systems (IROS 2002)*. Switzerland.

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson & C. Lebiere (Eds.), *Atomic Components of thought* (pp. 167-200). Mahwah, NJ:: Lawrence Erlbaum.

Carson-Radvansky, L. A., & Logan, G. D. (1997). The influence of functional relations on spatial template construction. *Journal of Memory & Language, 37*, 411-437.

Carson-Radvansky, L. A., & Radvansky, G. A. (1996). The influence of functional relations on spatial term selection. *Psychological Science, 7*, 56-60.

Casper, J. (2002). *Human-Robot Interactions during the Robot-Assisted Urban Search and Rescue Response at the World Trade Center*. Unpublished Master's, USF, Florida.

Cassimatis, N. L. (2002). *A Cognitive Architecture for Integrating Multiple Representation and Inference Schemes*. Unpublished Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Cassimatis, N. L., Trafton, J. G., Bugajska, M. D., & Schultz, A. C. (in press). Integrating Cognition, Perception and Action through Mental Simulation in Robots. *Robotics and Autonomous Systems*.

Cassimatis, N. L., Trafton, J. G., Schultz, A., & Bugajska, M. (2004). Integrating Cognition, Perception and Action through Mental Simulation in Robots. In *Proceedings of the 2004 AAAI Spring Symposium on Knowledge Representation and Ontology for Autonomous Systems*: AAAI.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, R. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA Books.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior, 22*, 1-39.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Fong, T., Thorpe, C., & Bauer, C. (2002). Robots as partner: Vehicle teleoperation with collaborative control. In A. Schultz, L. Parker & F. Schneider (Eds.), *Multi-Robot systems: From swarms to Intelligent Automata*: Kluwer.

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

Goldin-Meadow, S. (1997). When gestures and words speak differently. *Current Directions in Psychological Science, 6*(5), 138-143.

Harrison, A. M., & Schunn, C. D. (2002). ACT-R/S: A computational and neurologically inspired model of spatial reasoning. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty Fourth Annual Meeting of the Cognitive Science Society* (pp. 1008). Fairfax, VA: Lawrence Erlbaum Associates.

Harrison, A. M., & Schunn, C. D. (2003a). ACT-R/S: Look Ma, No "Cognitive-map"! In *International Conference on Cognitive Modeling*.

Harrison, A. M., & Schunn, C. D. (2003b). Segmented spaces: Coordinated perception of space in ACT-R. In F. Detje, D. Dorner & H. Schaub (Eds.), *The logic of cognitive systems: Proceedings of the fifth international conference on cognitive modeling* (pp. 307). Bamberg, Germany.

Hughes, K., Tokuta, A., & Ranganathan, N. (1992). Trulla: An algorithm for path planning among weighted regions by localized propogations. In *Proc. of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 469-476). Raleigh, NC: IEEE Press.

Huttenlocher, J., & Kubicek, L. (1979). The coding and transformation of spatial information. *Cognitive Psychology, 11*, 375-394.

Lee, F. J., & Gamard, S. J. (2003). Hide and seek: Using computational cognitive models to develop and test autonomous cognitive agents for complex dynamic tasks. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Boston, MA.

Levelt, W. J. M. (1984). Some perceptual limitations on talking about space. In A. J. van Doorn, W. A. van der Grind & J. J. Koenderink (Eds.), *Limits in perception* (pp. 323-358). Utrecht: VNU Science Press.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL, USA: University of Chicago Press.

Mintz, F., Trafton, J. G., Marsh, E., & Perzanowski, D. (in press). Choosing frames of reference: Perspective-taking in a 2D and 3D navigational task. In *Proceedings of the Human Factors and Ergonomics Society, 2004*. New Orleans: HFES.

Moravec, H. P., & Elfes, A. E. (1985). High resolution maps from wide angle sonar. In *Proc. of the IEEE International Conference on Robotics and Automation* (pp. 116-121). St. Louis, MO: IEEE Press.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81-103.

Nass, C., Steuer, J. S., & Tauber, E. (1994). Computers are social actors. In *Proceedings of the CHI 94 Conference* (pp. 72-77).

Newcombe, N., & Huttnelocher, J. (1992). Children's early ability to solve perspective taking problems. *Developmental Psychology, 28*, 654-664.

Perzanowski, D., Schultz, A., & Adams, W. (1998). Integrating Natural Language and Gesture in a Robotics Domain. In *Proceedings of the IEEE International Symposium on Intelligent Control: ISIC/CIRA/ISAS Joint Conference* (pp. 247-252). Gaithersburg, MD: National Institute of Standards and Technology.

Perzanowski, D., Schultz, A., Adams, W., Bugajska, M., Marsh, E., Trafton, J. G., et al. (2002). Communicating with teams of cooperative robots. In A. C. Schultz & L. E. Parker (Eds.), *Multi-Robot Systems: From Swarms to Intelligent Automata* (pp. 16-20). The Netherlands: Kluwer.

Perzanowski, D., Schultz, A., Adams, W., & Marsh, E. (1999). Goal tracking in a natural language interface: Towards achieving adjustable autonomy. In *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation: CIRA '99* (pp. 208-213). Monterey, CA: IEEE Press.

Perzanowski, D., Schultz, A., Adams, W., & Marsh, E. (2000). Using a Natural Language and Gesture Interface for Unmanned Vehicles. In G. R. Gerhart, R. W. Gunderson & C. M. Shoemaker (Eds.), *Proceedings of the Society of Photo-Optical Instrumentation Engineers* (Vol. 4024, pp. 341-347).

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

Petty, M. D. (2001). Do we really want computer generated forces that learn? In *Proceedings of the 10th Conference on Computer Generated Forces and Behavioral Representation (CGF&BR)*. Norfolk, VA.

Previc, F. H. (1998). The neuropsychology of 3-D space. *Psychological Bulletin, 124*(2), 123-164.

Schultz, A., & Adams, W. (1998). Continuous localization using evidence grids. In *Proceedings of the 1998 IEEE International Conference on Robotics and Automation* (pp. 2833-2839). Leuven, Belgium: IEEE Press.

Schultz, A., Adams, W., & Yamauchi, B. (1999). Integrating exploration, localization, navigation and planning wih a common representation. *Autonomous Robots, 6*, 293-308.

Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*: MIT Press.

Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., & Adams, W. (in press). Spatial Language for Human-Robot Dialogs. *IEEE Transactions on Systems, Man, and Cybernetics*.

Taylor, H. A., & Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory & Language, 31*, 261-292.

Trafton, J. G., Cassimatis, N. L., Brock, D. P., Bugajska, M. D., Mintz, F. E., & Schultz, A. C. (under review). Enabling effective human-robot interaction using perspective-taking in robots.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: Extracting and generating information from complex visualizations. *International Journal of Human Computer Studies, 53*(5), 827-850.

Trafton, J. G., Marshall, S., Mintz, F. E., & Trickett, S. B. (2002). Extracting explicit and implicit information from complex visualizations. In M. Hegarty, B. Meyer & H. Narayanan (Eds.), *Diagramatic representation and inference* (pp. 206-220). Berlin Heidelberg: Springer-Verlag.

Trafton, J. G., Schultz, A. C., Bugajska, M. D., Gittens, S., & Mintz, F. E. (2001). An investigation of how humans and machines deal with increases in reactivity. In *The proceedings of the tenth conference on Computer Generated Forces and Behavioral Represenntation (CGF&BR)*. Norfolk, VA.

Trafton, J. G., Schultz, A. C., Perzanowski, D., Adams, W., Bugajska, M. D., Cassimatis, N. L., et al. (under review). Children and robots learning to play hide and seek.

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (in press). Connecting Internal and External Representations: Spatial Transformations of Scientific Visualizations. *Foundations of Science*.

Trickett, S. B., Ratwani, R. M., & Trafton, J. G. (under review). Real-World Graph Comprehension: High-Level Questions, Complex Graphs, and Spatial Cognition. *Cognitive Science*.

Wallace, R., Allan, K. L., & Tribol, C. T. (2001). Spatial perspective-taking errors in children. *Perceptual and Motor Skills, 92*(3), 633-639.

Wang, R. F., & Spelke, E. S. (2002). Human spatial representation: Insights from animals. *Trends in Cognitive Sciences, 6*(9), 376-382.

Wauchope, K. (1994). *Eucalyptus: Integrating Natural Language Input with a Graphical User Interface* (No. NRL/FR/5510-94-9711). Washington, DC: Naval Research Laboratory.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(103-128).

Yamauchi, B., Schultz, A., & Adams, W. (1998). Mobile Robot Exploration and Map Building with Continuous Localization. In *Proceedings of the 1998 IEEE International Conference on Robotics and Automation* (pp. 3715-3720). Leuven, Belgium: IEEE.