Intersubjectivity in human–agent interaction

Justine Cassell and Andrea Tartaro
Northwestern University, Center for Technology and Social Behavior

What is the hallmark of success in human–agent interaction? In animation and robotics, many have concentrated on the *looks* of the agent — whether the appearance is realistic or lifelike. We present an alternative benchmark that lies in the dyad and not the agent alone: Does the agent's behavior evoke intersubjectivity from the user? That is, in both conscious and unconscious communication, do users react to behaviorally realistic agents in the same way they react to other humans? Do users appear to attribute similar thoughts and actions? We discuss why we distinguish between appearance and behavior, why we use the benchmark of intersubjectivity, our methodology for applying this benchmark to embodied conversational agents (ECAs), and why we believe this benchmark should be applied to human–robot interaction.

Keywords: embodied conversational agents, human–robot interaction, intersubjectivity, nonverbal behavior, psychological benchmarks

Intersubjectivity in human–agent interaction

In Spielberg's *Artificial Intelligence*, neither husband nor wife at first responds to their brand new boy robot David the way they respond to their son Martin. They are wary of interacting with something that, despite its human-like appearance, is for them just a piece of tin. Because they don't know how to act with a robot child, their behavior is unnatural. But when David puts himself at risk by eating real food in imitation of his human brother, the adults' unconscious parental response to his child-like behavior wins the day. They excoriate David and Martin equally. Later the mother takes David's hand to comfort him and suddenly realizes that she has responded to him as if he were a real boy — as if this piece of tin were *like her*. His behavior rather than his appearance has won her over; and it is her behavior that signals the shift.

Since 1994 when we presented the first autonomous communicating virtual character (Cassell, Pelachaud et al., 1994; Cassell, Stone et al., 1994), we have

Interaction Studies 8:3 (2007), **391–410**.
ISSN 1572–0373 / E-ISSN 1572–0381

continued a research program that consists of implementing virtual humans (embodied conversational agents, or ECAs) on the basis of a micro-analysis of human behavior, and then evaluating the success of those agents. Success is assessed both in terms of the ECA's status as scientific visualization of the micro-analysis and as a human–computer interface, and is evaluated on the basis of a micro-analysis of the humans interacting with the agents. Rather than looking solely at the ECAs for clues to their success, we look to the dyad. Rather than concentrating on whether the ECAs look realistic or lifelike, we concentrate on the realism of their behavior. Their behavioral realism leads to 'realism' (smooth, unself-conscious naturalness) in the user's behaviors as well. A realistic interaction between user and agent consists of each producing the appropriate contingent response. Smiling and nodding when the interlocutor makes an assertion, for example, as opposed to staring at the interlocutor without giving behavioral evidence of understanding what one is being told.

Because our goal is natural interaction and engagement, our criterion for success lies in the communication between the two agents (one real and one virtual) rather than in the visual features of the virtual agent. Every aspect of the interaction that differs from the human interaction is a reason to go back to our micro-analyses of human behavior to understand what aspects of human behavior we missed when we implemented the ECA.

When users act, in their unconscious behaviors, as if the ECA is like them, we call that *intersubjectivity*, after Trevarthen's (1987) description of the infant's growing understanding of the motives and intentions of the parent, based on verbal and nonverbal social interaction (Cassell, 2001). Intersubjectivity comes into play in many of our interactions, as we feel what others feel, or function as if they have the same motives and intentions we do (Beebe, 2005). Although we have used this benchmark to evaluate graphical agents, we believe it is general enough to be applied to all kinds of humanoid agents (e.g., ECAs, avatars, robots) as it places the criterion for success in the unconscious reaction of the user. It relies on a built-in gold standard, which is human–human interaction — a gold-standard that derives from our two goals for ECAs: (a) scientific visualizations of micro-analytic theories of human behavior and (b) intuitive and natural humanoid interfaces. This is not to say that human–human interaction should be the gold standard for all interactive systems, or even all robots or agents. However, if one makes the effort to build a humanoid interface, it stands to reason that one should follow the metaphor of embodiment and anthropomorphism to its logical conclusion, and one should rely on the affordances of human–human behavior. Those affordances, we believe, reside mainly in *behavior* and not in appearance (Cassell, Bickmore, Campbell, Vilhjalmsson, & Yan, 2001).

In what follows, we explain why we rely on a benchmark of natural reactions rather than natural looks, and how this benchmark is derived from our design

focus on embodied linguistic behaviors in people. We then describe the iterative methodology we use to design ECAs, followed by a discussion of some of what we've learned about human–machine communication using this methodology. We describe how we believe this benchmark can be used in studies on human–robot interactions (HRI) and how we believe it can be applied to HRI.

Motivation

It is no coincidence that the ECA has a body, and that it has a full body rather than just a head. The ECA arose out of the study of embodied language and the dyadic nature of communication. Through that study of human behavior, our ECAs have come to behave more realistically, and through the ECA we have learned more about how people use their bodies in communication.

Increasingly, language and communicative behavior is viewed through the lens of social practice, or interpersonal action, situated in the space between two or more people, emergent and multiply-determined by social, personal, historical, and moment-to-moment linguistic contexts, and expressed verbally and nonverbally, through body gestures and eye gaze. By "located between people" we mean that every behavior in communication, both conscious and unconscious, is a function of the interaction. This is particularly evident and striking in the millisecond-quick choices that are made by speakers and listeners as they copy one another's accents, converge on particular ways of referring to the world, and modify their gestures mid-stream as they respond to evidence of lack of understanding in the other person. We study these choices and their effects on the course of communication by investigating the relationship between visible nonverbal behaviors (e.g., eye gaze, posture shifts, gestures, head nods and eyebrow raises) and a set of underlying discourse functions (such as emphasizing new information, exchanging turns, structuring topics and determining what is shared information). For example, our study of when people shifted their weight during conversation led to the conclusion that posture shifts mark the beginning of new discourse topics, and that posture shifts are most likely to occur when topic shifts and shifts in speaker coincide; in turn, this understanding led us to build ECAs that shifted their bodies as a function of the underlying discourse structure (Cassell, Nakano, Bickmore, Sidner, & Rich, 2001). Likewise, a study of eye gaze demonstrated that people look up towards the other person when they do not understand (Nakano, Reinstein, Stocky, & Cassell, 2003), and we implemented this in an ECA who was then able to use the user's eye gaze to determine whether to continue the dialogue or go back over the previous point. And it was in observing an ECA give directions that we realized that we had hithertofore neglected the key role that redundancy

(repetition, explanation, elaboration) plays in the discourse (Kopp, Tepper, Ferriman, Striegnitz, & Cassell, in press).

The discourse functions that structure information also play a social role. Thus, each discourse choice that speakers and hearers make in the unfolding conversation also marks whether they share goals, whether they are friends or strangers, and whether they view one another as fundamentally similar. Familiarity between participants, for example, is marked by an increase in coordination, such as overlapping speech and nonverbal synchrony, and a reduction in politeness.

The development of the ECA was born out of a desire to better understand the relation between verbal and nonverbal aspects of conversation. But the ECA has also enabled a number of new applications for computers, filling traditionally human roles, such as peer tutor, realtor and life trainer. Over the last decade, we have used ECAs to implement and test models of human–human communication. We have also used ECAs as interfaces to databases of information about houses, ways of communicating directions, and peer tutors for young children learning how to read and write.

However, while an ECA has a body, a number of researchers have demonstrated that it is not the realism of the body's appearance that results in successful human–agent interaction (Koda & Maes, 1996). In fact, a number of researchers have demonstrated the dangers of a mismatch between the behavior and appearance of an agent (Bailenson et al., 2005; Garau et al., 2003; Nass, Isbister, & Lee, 2000). Likewise, a number of roboticists are beginning to investigate the relative contributions of movement and appearance to the naturalness of interaction (cf. MacDorman et al., 2005).

Methodology for applying our intersubjectivity benchmark

Imagine that you the reader are sitting next to us, the authors of this article, in a small control room separated from the experimental room by one-way glass. We are watching a young man ask an ECA for directions. The ECA is a graphical picture on a huge screen, but it has the appearance of a very large blue robot — with a clanky jaw and ham-like hands, and a head that moves jerkily. Nevertheless, the young man is intent on getting directions. The robot points at a map between them and tells him to turn right at a large potted plant, and the young man looks at the map. The robot finishes speaking and then looks at his interlocutor, and in response the young man nods and smiles. Compare this interaction with our next participant, a young woman who is also asking directions of the blue robot on the screen, but in an experimental condition in which all of the robot's conversational feedback mechanisms have been turned off. Strikingly, this woman is

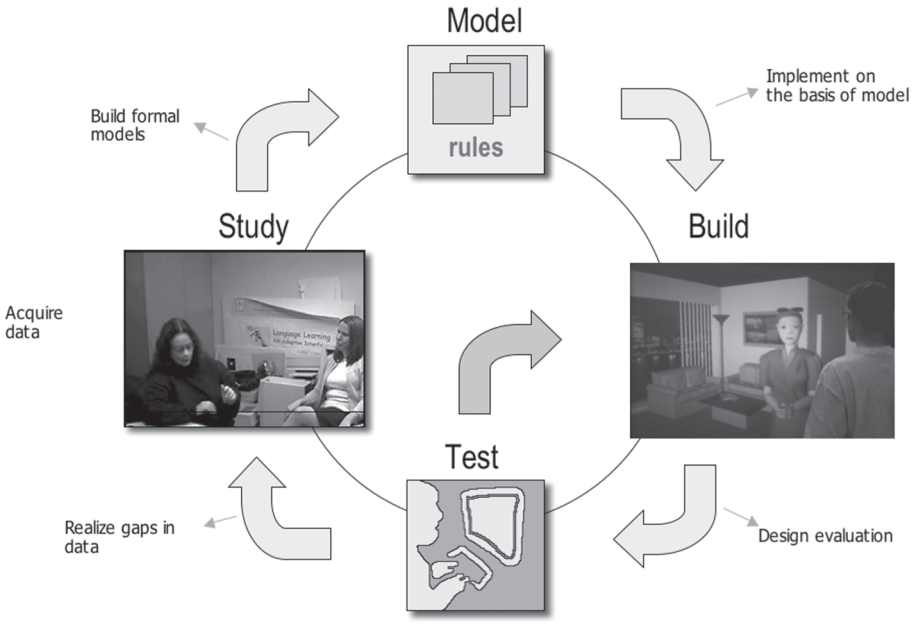

Figure 1. Iterative design methodology

neither nodding nor smiling nor looking in the robot's direction. She started out looking at him, but within an utterance or two she began to stare at the map. These two participants have shown which of the two ECAs is more human-like (Nakano, Reinstein, Stocky, & Cassell, 2003). We have carried out many similar experiments as a way of assessing which version of an agent evokes a more natural interaction style, and we have also used it to determine which model of human behavior (as instantiated in which of two versions of an ECA) is more accurate. And in each ECA, the original model of human–human interaction serves both as input and benchmark.

To investigate the relationships between behaviors, discourse functions, interactional structures, and rapport, we use an iterative design process that enables us to ask questions about communicative behaviors, build formal models of interaction, and test those models (see Figure 1). (a) First we acquire data of people communicating in natural face-to-face settings to understand how the surface level behaviors carry discourse and social meanings. In some cases this work has been done for us by researchers in social psychology or psycholinguistics, but many phenomena have not yet been studied in ways that can provide *formal* models of them. (b) A formal model is a predictive model. Because we are going to turn the model into algorithms for an embodied agent, our analysis of human behavior must be able to predict the context in which a given behavior will occur. (c) The

formal model is then translated into a computational architecture for an embodied conversational agent, and the ECA is implemented. (d) The implementation is then evaluated by having people interact with it, as we described above. This evaluation pushes us to go back to the data and extract further models of conversational interaction so as to improve the ECA.

Human behavior studies

We start with a question about human–human interaction, and begin to answer this question by looking at videotapes of people interacting. The domain in which we investigate human communicative behavior is key, as it will also serve as the domain in which our ECA will operate. We have found that attending to the social structure of domains (and beginning with domains that are highly scripted) is important, because it allows us to concentrate on the micro-analysis of behavior. That is in looking at domains such as direction-giving, real-estate sales and rental, life coaching and children's collaborative storytelling, we pay particular attention to the social constraints of the role so that the ECA can be a partner in a collaborative task. The task can be anything from making friends to making pasta; the constraints of the role tell us how the partnership will progress.

Figure 2. Direction-giving data

In Figure 2 participants engage in a dialogue on how to find their way around Northwestern University. This study arose from our interest in how listeners integrate the gestures they see into their understanding of what they heard given that there are no consistent form-meaning mappings in gesture. We therefore asked people to give directions around a route that we knew well so we could focus on the shape of people's hands as they gave directions, and described each landmark and path along the route. The interaction was videotaped using four cameras trained on different parts of the scene and then each gesture was transcribed, along with the speech that accompanied it.

Building and implementing a predictive model

Our analysis of these videos informs the design of a predictive model of meaning — gesture mappings — that is then implemented into an ECA. This predictive model describes when gestures refer to landmarks and when they refer to paths, and how to predict when the gesture will demonstrate a flat handshape, and when a curved handshape. Table 1 shows a model derived from another study of human interaction, concentrating on the relation between embodied behaviors and storyteller roles among children (Wang & Cassell, 2003). While this is not a comprehensive model of collaborative storytelling, it represents the collaborative speech acts that result in turns being exchanged, and the nonverbal behaviors children use to exchange turns. Our study of children's collaborative storytelling arose out of the project of building a virtual child to collaborate with real children. Our choice of behaviors to document arose, in turn, from the realization that careful coordination of both turn-taking behaviors and speech acts are essential for a virtual peer to participate successfully in a collaborative storytelling task.

Implementation

We try to make formal models that accurately represent how human communication functions. But formal models do not take into account what is possible computationally. When translating the formal model into an implementation, we come up against the constraints of our computational platform. This is why a formal model is not an architecture in itself, and why our computational architectures are not unmediated representations of how cognition works. It is important to be clear about both the model and the architecture so that as computation progresses we can come to closer approximations of how humans interact among themselves. We can use human architectures, so to speak, as a way of pushing engineering prowess.

Table 1. Model of collaborative storytelling among children

Roles	Speech act	Speaker	Function	Turn-taking behaviors
Critics and authors	Suggest	Critic	To suggest an event or idea to the story	Eye gaze towards author, author may use paralanguage drawls and socio-centric sequences like "uhh"
	Correct	Critic	To correct what's been said	Eye gaze towards author
	Question	Both	To seek clarification or missing information	Eye gaze towards other, lack of backchannel feedback like head nods, increased body motion, author stops gesturing
	Answer	Both	To clarify or supply missing information	Eye gaze towards other, rising pitch, question syntax, author stops gesturing
	Acknowledge	Author	To acknowledge a suggestion or correction	Eye gaze towards critic, backchannel feedback like "mm-hmm", author stops gesturing
Facilitator and collaborator	Direct	Facilitator	To suggest storylines and designate roles	Eye gaze towards collaborator, socio-centric sequences like "OK", both stop gesturing
	Acknowledge	Collaborator	To acknowledge a role designation or storyline suggestion	Eye gaze towards facilitator, backchannel feedback like head nods, both stop gesturing
	Elaborate	Both	To narrate following suggested script	Eye gaze towards other, may start gesturing
Co-authors	Role-play	Both	Play the role of characters in the story	Eye gaze towards action, prosody of in-character voice, gesture with prop
	Simultaneous turns	Both	Compete for turn	

In the example of virtual peers for children, we have not yet found a solution for speech recognition for children, and so we cannot base the virtual peer's performance on an understanding of what the child said. We have devised ways for Sam to take turns with children relying on noise threshold to detect turns, and the location of the toys to detect content. To model collaboration with children, we chose three speech acts (one from each role) that maximize the variety of collaborative interactions, and minimize the strain on Sam's understanding.

The direction-giving study described above resulted in the direction-giving robot shown in Figure 3. In this system, the ECA does use speech recognition to understand the user, and it has a complete dialogue system architecture. In this

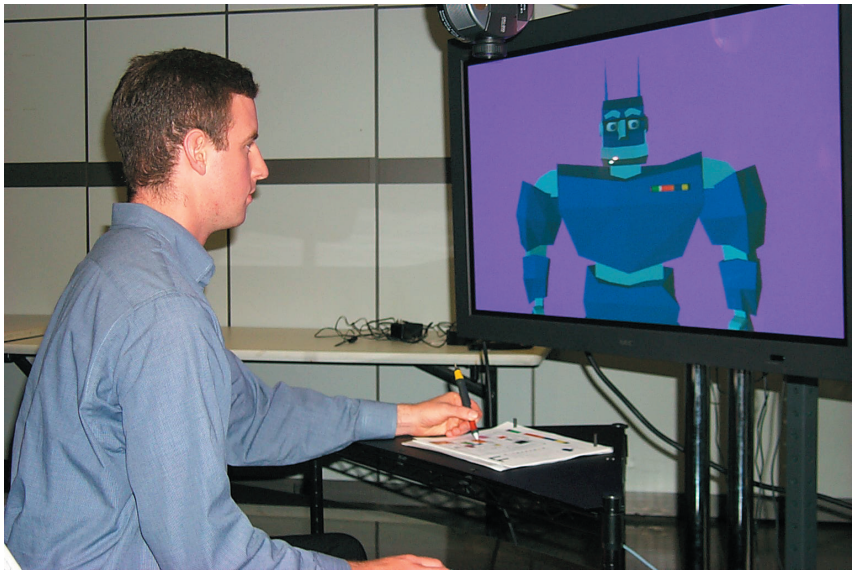

Figure 3. Person interacting with Robot ECA

system, however, the linearity of the architecture, coupled with the fact that utterances must be generated from start to finish before they are realized, makes this architecture incapable of dealing with some of the quick interactive phenomena we find in human–human communication. Resolving this mismatch between the phenomena that we have described in our models, and the implementations that we can currently build, is an active topic for our current research.

Evaluation

Finally, to evaluate our implementation both as a model of human behavior and a successful interface, we watch our ECA interact with real people to see whether it successfully evokes intersubjectivity. In this context, we operationalize intersubjectivity as the extent to which the details of the users' communicative behavior when interacting with the ECA resembles behavior between humans. ECAs give us the flexibility to implement a model of communicative behaviors and to turn on and off aspects of the model, and then observe people's responses and compare them to human–human interaction. We use a combination of observations of the human's behaviors in the different interactions, and questionnaires where participants assess the ECA. We use questionnaires because self-reports can help in designing ECAs as interfaces by revealing information about the users' experience with the system. In their meta-analysis of realism in ECAs, Yee and colleagues found that effect sizes were larger in studies where subjective measures were used

(Yee, Bailenson, & Rickertsen, 2007). In our evaluation of ECAs, we announce success when a person turns to a virtual human, unconsciously nods, and carries on an interested and engaged conversation. Success is equally at hand when people confirm through their questionnaire responses that we have implemented the ECA as intended.

Perhaps more interesting, however, is observing where the communication breaks down — when the participant becomes uncomfortable or the interaction looks wrong. These instances reveal what we do not yet know about communicative behaviors and lead to new questions about human–human interaction.

In this way, virtual humans are tools to think with. They allow us to understand where our models of human communication are flawed. This only works because mechanical beings that seem human make us attribute humanness and aliveness to them, and that makes us act human and alive. Thus, when they are successful, virtual humans evoke distinctly human characteristics in our interaction with them. This is our benchmark of success, and enables us to test the success of our ECA, and our own understanding of human–human communication.

An example of how this evaluation is carried out comes from Bickmore and Cassell (2005) who implemented in a real estate agent ECA (called REA) a model of how small talk might have an effect on trust. Initial comparisons of "small-talk REA" and "task-only REA" showed that extroverted users were more likely to trust REA when she engaged in chitchat in addition to completing the task. For introverted users there was no difference in trust between the two conditions. Trust in this experiment was assessed by questionnaire. Participants in the study made it clear that we had indeed implemented a REAL "small-talker."

We also studied the user's own communicative behavior during their interactions with REA and discovered that some users were quite passive during the interaction, while others initiated parts of the conversation, and that this division did not correlate with introversion and extroversion. More active users preferred the small-talk version of REA.

To further investigate the origin of these differences, we added a condition where users spoke to REA by phone. In this 2x2 design, we compared a REA that engaged in small talk to one that only engaged in task-oriented talk, and an embodied system to a phone-based system. We predicted that trust would be higher in the social condition for extroverts, as we had found previously, and higher in the embodied conditions where the full effect of small talk would be found. What we found, however, was that REA was judged as more friendly, warm, comfortable, informed and knowledgeable on the phone, and there was an interaction such that Phone REA was more tedious when she engaged in only task-oriented talk, while embodied REA was more tedious when she used social chitchat. We interpreted these results to mean that REA did well in single-channel formats, but

her body language conflicted with the cues projected by her voice. That is, while REA's choice of words in the social condition projected a social extrovert, her body language projected introversion.

Results of applying our intersubjectivity benchmark

Applying the iterative design methodology described above, and comparing human interaction with ECAs in different conditions, has led to insights about human–human and human–machine communication.

Human communication

The methodology we have described here has allowed us to adduce four general properties of human conversational behavior: (1) the distinction between the *interactional* and *propositional* functions of language and conversation; (2) the distinction between conversational *behaviors* (such as eyebrow raises) and conversational *functions* (such as turn taking); (3) the importance of *timing* among conversational behaviors; and (4) the deployment of each *modality* to do what it does best. These properties are summarized below and elaborated in (Cassell, 2007).

Propositional and interactional functions. Propositional information is the 'content' of the interaction and moves the conversation forward, while interactional information regulates the process of the conversation. Both propositional and interactional information can be represented in verbal and nonverbal forms. Thus when REA says that the property is located five minutes from campus while making a walking gesture with her fingers, she is expressing propositional information. When she nods and uses sociocentric sequences such as "uh huh," she is performing interactional functions.

Conversational behaviors and functions. As we described above, human interaction is composed of various verbal and nonverbal behaviors that can serve different functions in conversation. Each function can be filled through a number of different behaviors, in one or several modalities. Likewise, the same behavior can serve different conversational functions depending on the context. The form given to a particular discourse function depends on, among other things, the current availability of modalities such as the face and the hands, type of conversation, cultural patterns and personal style. Thus, listeners can nod to indicate they understand, or say "uh huh." Alternatively, REA's walking gesture, described in the example above, may add propositional information to the conversation, or indicate she wants to talk, depending on the context.

Timing. The relative timing of conversational behaviors plays a large role in determining their meaning. Thus, although it has long been known that the most effortful part of a gesture co-occurs with prosodic stress, research using ECAs (Cassell, Stone et al., 1994) revealed content gestures are most likely to co-occur with the *rhematic* or new contribution part of an utterance. For example, if a speaker is pointing to her new vehicle and saying "this car is amazingly comfortable. In fact, its comfort comes from the fact that it has reclining seats," the phrase "amazingly comfortable" would be the rheme in the first sentence, because car is redundant (since the speaker is pointing to it) and "reclining seats" would be the rheme in the second sentence, because comfort has already been mentioned. Therefore, the speaker would be most likely to produce hand gestures with "amazingly comfortable" and "reclining seats."

Using modalities to do what they do best. In face-to-face conversation, we dispose of multiple modalities of expression. We depend on each modality, and their combinations, to communicate. We may use gestures to indicate things that may be hard to represent in speech, such as spatial relationships among objects (Cassell, Stone, & Yan, 2000), and we use our ability to simultaneously produce speech and gesture to communicate quickly. In this sense, face-to-face conversation may allow us to accomplish more than information transmission. We may use the body to indicate rapport with others, while language is getting task work done.

Human–machine communication

Our research on ECAs has taught us not only about aspects of human–human interaction, but also human–computer interaction. ECAs may be useful in situations where keyboard and mouse interactions are difficult or impossible, such as when driving a car, using a small device like a cellphone, or for the elderly and children who don't have the desire or literacy skills to use desktop computers. ECAs may also be useful where the kinds of rapport entrained by interaction with the system are useful in and of themselves. In our recent work on virtual peers, we examine the kinds of rapport evoked by interacting with somebody who is similar culturally. Storytelling practices differ according to cultural background (Champion, 1998), and yet in schools, typically only one cultural practice is used to bootstrap literacy. Children from other cultural backgrounds may feel ignored and have trouble making a bridge from home to school language (Guiterrez & Rogoff, 2003). However, narrative structures from an individual's own tradition can make children feel welcome, and act as a bridge to formal content (Pinkard, 1999). In this context, we are developing a virtual peer (Alex) to act as a learning partner for African-American children. In the real world, peers are powerful learning partners in part because of the rapport they establish with one another (Pellegrini, Galda, Bartini, & Charak,

1998). Rapport can exist on a cultural level and on a micro-interpersonal level. In order to implement a culturally-specific rapport-building virtual peer, we are studying the verbal and nonverbal behavior of African-American children while they tell stories. We believe that a better understanding of how to implement rapport, both behaviorally and linguistically, will be a major contribution of the Alex project, as will a better understanding of the micro-linguistic ways that culture can be demonstrated in an ECA.

Alex and our other virtual peer project, Sam, described earlier, use language and nonverbal behavior in a narrative task as a way of teaching children critical literacy skills. We became interested in autism spectrum disorders (ASD) because ASD is characterized by impairments in exactly those areas that are used for storytelling: social interaction, communication and imagination. Children with ASD are not capable of engaging in reciprocal social interaction and appropriate verbal and nonverbal behavior in conversation. This seriously limits their access to learning opportunities because of the important role social interaction plays in learning. We are developing a new kind of virtual peer to help children with ASD develop reciprocal social interaction and communication skills. This research contributes to an active area of research using humanoid technologies as an instructional aid for children with ASD (Goldsmith & LeBlanc, 2004). Recent approaches include robot therapies (inter alia Feil-Seifer & Mataric, 2005; Robins, Dickerson, Stribling, & Dautenhahn, 2004), virtual tutors (Bosseler & Massaro, 2003) and virtual environments (Parsons, Mitchell, & Leonard, 2004).

Collaborative narrative with a virtual peer is an ideal task for investigating the pragmatic deficits of ASD. And storytelling enables children with ASD to practice turn-taking behaviors, address the beliefs of their peers, take on conversational roles, and invent narrative content. However, storytelling between children with ASD and typically-developing children is hard because typically-developing children don't have much patience for it. In addition, while children with autism avoid social interaction, they love interacting with computers (Goldsmith & LeBlanc, 2004). We hypothesize that virtual peer technology can enable children with ASD to understand the function of the behaviors involved in reciprocal social interaction.

To do this, we are implementing an "authorable" virtual peer (AVP) that can be used in three interaction modes. First, children interact with the virtual peer by telling stories with the system, and thereby rehearse verbal and nonverbal interaction skills with an indefatigable peer. In a second mode, children control the virtual peer by using a "Wizard of Oz" interface to select predefined responses. Using the interface they can select head and body gestures, utterances and story segments for the virtual peer to perform, and observe the outcomes of the interaction. Third, children can use authoring tools to create new behaviors and responses, and construct their own interaction examples. The AVP has its roots in research

on instructional technology systems and extends the constructionist tradition in education — the use of technology as "objects to think with" (Harel & Papert, 1991) — to learning about language and social interaction through building communicating virtual humans (Tartaro & Cassell, 2006).

Studying ECAs informs both our understanding of human–human interaction and human–computer interaction. But as should be clear from the complexity of the data we have described, and the partial nature of our implementations, what we really are learning from studying ECAs is that humanness is infinitely complex, and that the target moves off into the distance as we approach — which is a good thing. As we gain knowledge about an aspect of human behavior that we can model and incorporate into an ECA, we realize that the goal is part of a larger behavior. This opens new doors for using virtual humans to understand human communication. In addition, as we extend the domains in which ECAs function as interfaces — as direction-giving kiosks and learning partners — we learn more about the possibilities and limitations of using ECAs for human–computer interaction. One of those limitations comes from the 2D nature of ECAs.

Robots as conversational agents

Robots offer both new opportunities and challenges as both interlocutors and simulations of communicative behaviors. A decade and a half of research on ECAs can inform these research challenges. In this section, we first examine the affordances and challenges of the physicality of robots. We then describe some techniques for applying to them our benchmark of intersubjectivity.

Affordances and challenges

While an ECA is trapped on a screen, robots exist in three-dimensional space. We believe that the three-dimensionality of robots represents both a promise and a danger. The promise is that they can touch us. They make gestures that touch the user, they can sit side-by-side — in sum, they can share our space. There are gestures, such as leaning forward, that don't translate well to a projection on the screen. Some researchers in small group dynamics have claimed that "trust needs touch" (Handy, 1995). Leaning forward can indicate interest and attention to what someone is saying. All these three-dimensional gestures contribute to rapport and intimacy.

This physical existence can vary based on how much a robot can sense and affect the environment. In fact, Fong et al. (Fong, Nourbakhsh, & Dautenhahn, 2003) describe conversational agents as essentially a type of robot with limited

ability to affect the environment. A robot's physicality both constrains and sets expectations for the form and context of its social interaction.

The danger is that we can touch them. But the touch of metal or rubber skin may break the illusion of intersubjectivity, and subtle flaws are even more apparent when robots are very human-like. The bar is higher for touch than for sight — or perhaps touchable robots are simply a more distant goal. We react to touch so instantaneously — would we react well to touching a robot, and would the robot react well to being touched, over the medium and long-term? For example, analysis of children interacting with a robot dog (AIBO) and a live dog illustrates that children will interact with AIBO in ways that are similar to how they will interact with a live dog. However, they spent more time touching and within arms distance of the live dog (Melson et al., 2005).

The articulation of mechanical parts is currently not as flexible as computer animation. In particular, robotic facial expressions are often not lifelike, and computer animation currently holds more power for expressiveness of eyes, for instance (Fong, Nourbakhsh, & Dautenhahn, 2003). The limitations of robotic articulation could be turned into a strength: What is the minimum representation of a gesture needed to evoke the desired response? For example, the robot Kismet's face uses fifteen degrees of freedom, and is able to express numerous, easily interpreted expressions including anger, fatigue, fear, disgust, excitement, happiness, interest, sadness and surprise (Breazeal & Scassellati, 1999). When using a robot as an interface, or as a tool for understanding models of human interaction, there are opportunities for exploring the use of 3D space to affect social response, while the physicality and limitations of the robot could break the illusion of the interaction.

Applying the intersubjectivity benchmark

Considerable work in robotics has concentrated on implementing theory of mind in human–robot interaction (c.f. Scassellati, 2002). A robot with theory of mind is able to attribute to its interlocutor beliefs, desires and goals other than its own. The hope is that when the robot has a theory of mind, then the user will also attribute different beliefs, desires and goals to the robot. Intersubjectivity, however, is just as important; it is, in a sense, the other side of the theory of mind coin. Whereas theory of mind asks that the two participants in an interaction to attribute different intentions to one another (and therefore exert effort to make themselves understood to the other), intersubjectivity asks that the two participants in an interaction consider each other to be similar enough so that they can map each other's actions and words onto their own motivations for similar actions and words.

In our own work we find that the benchmark of intersubjectivity moves us away from asking whether the robot is realistic or lifelike. This is a relief as ultimately

we don't want to spend our time implementing a robot that sneezes, or becomes whiny when it's tired. Instead we aim for a robot or ECA that is adequate to the task of evoking human-like responses from the user; that has behavior realistic enough to evoke from the user an unconscious sense of sameness.

The interesting part, then, is to implement (rather than sneezes or whines) the tiniest of behaviors that, in the human world, would identify somebody as truly human, from the role they have taken at that moment, all the way down to each behavior that makes up that role. If it's a doorman, choose obsequiousness, and break down obsequiousness into its tiniest parts. In a bottom-up perspective on implementation, good interaction comes from the implementation of the correct moment-to-moment embodied behaviors. Obsequiousness may be made up of downward gaze, reticence to initiate new topics of conversation, many phatic responses (e.g., uh huh, yes ma'am). If the robot is to be a companion, then investigate how companions function: what do companions or friends do during interaction? Do they orient their bodies towards one another, or assume a parallel stance towards a shared task?

Believability in and of itself should not be the goal of human–robot interaction. In the early days of ECA research, there was much focus how we get people to believe in an agent. However, believability should not be the goal of human–robot interaction. Ultimately, ECA researchers realized that believability was tied to function — believable *as what*? A teacher? A peer? A butler? A mechanic? A micro-analysis of behavior should obviate the need to worry about *believability*. More important than believability is automatic suspension of disbelief — automatic ascriptions of "like-us-ness" to the agent. How do we make an agent we react to in human-like ways despite ourselves? Agents that live alongside us in our world, to which we react immediately and unconsciously, will be too much a part of our lives for us to spend time worrying about whether they are believable.

Kahn and colleagues (2006) propose six psychological benchmarks for building successful robots, and they argue against the approach of taking all findings from all psychological scientific disciplines and replicating the results in human–robotic interaction. While we agree that the replication of the extensive psychological research is problematic, we do believe that focusing on psychological findings encourages researchers to look for success within the context of a specific interactional phenomenon. While striving for perfect psychological realism is ridiculous, implementing the minutiae of psychological behavior in the robot — the micro verbal and nonverbal behaviors, and their relationship to functions in the conversation — is critical.

Conclusion

In our work on ECAs, we measure success of the agent by looking at the interaction between human and machine rather than at the appearance of the agent. Our benchmark for a successful agent is when users act as if the ECA is like them, and thus unconsciously interact with the ECA using verbal and nonverbal behaviors that they would use with another human. In this paper, we have described this benchmark as *intersubjectivity*, and motivated why we look for natural reactions from a user rather than natural actions from an agent. Ultimately, we believe that the goal of human–agent interaction, which includes both human–robot interaction and human–ECA interaction, should not be a believable agent; it should be a believable interaction between human and agent in a given context.

Acknowledgements

Many thanks to all of the students, staff, and colleagues who participated in the research reported here, both at MIT and at Northwestern. Research funding comes from NSF, Autism Speaks, and CAN.

References

Bailenson, J., Swinth, K., Hoyt, C., Persky, S., Dimov, A., & Blascovich, J. (2005). The independent and interactive effects of embodied agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. PRESENCE: Teleoperators and Virtual Environments 14(4), 379–393.

Beebe, B. (2005). Forms of intersubjectivity in infant research and adult treatment. New York: Other Press.

Bickmore, T., & Cassell, J. (2005). Social dialogue with embodied conversational agents. In J. van Kuppevelt, L. Dybkjaer, N. Bernsen (Eds.), Advances in multimodal dialogue systems (pp. 23–54). New York: Kluwer Academic.

Bosseler, A., & Massaro, D. W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. Journal of Autism and Developmental Disorders, 33(6), 653–672.

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 858–863). Kyonjiu, Korea.

Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interface. AI Magazine, 22(3), 67–83.

Cassell, J. (2007). Body language: Lessons from the near-human. In J. Riskin (Ed.), Genesis redux: Essays in the history and philosophy of artificial life (pp. 346–374). Chicago: University of Chicago Press.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H. H., & Yan, H. (2001). More than just a pretty face: Conversational protocols and the affordances of embodiment. Knowledge-Based Systems, 14(1-2), 55–64.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., & Rich, C. (2001). Non-verbal cues for discourse structure. Thirty-ninth Annual Meeting of the Association of Computational Linguistics (pp. 106–115). Toulouse, France.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., et al. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. SIGGRAPH (pp. 413–420). Orlando, FL.

Cassell, J., Stone, M., Douville, B., Prevost, S., Achorn, B., Steedman, M., et al. (1994). Modeling the interaction between speech and gesture. 16th Annual Conference of the Cognitive Science Society (pp. 153-158). Atlanta: Lawrence Erlbaum Associates.

Cassell, J., Stone, M., & Yan, H. (2000). Coordination and context-dependence in the generation of embodied conversation. International Conference on Natural Language Generation (pp. 171–178). Mitzpe Ramon, Israel.

Champion, T. (1998). "Tell me somethin' good": A description of narrative structures among African-American children. Linguistics and Education, 9(3), 251–286.

Feil-Seifer, D., & Mataric, M. J. (2005). Defining socially assistive robots. Paper presented at the IEEE 9th International Conference on Rehabilitation Robotics (pp. 465–468). Chicago, IL.

Fong, T., Nourbakhsh, I., & Dautenhah, K. (2003). A survey of socially interactive robots. Robotics and autonomous systems, 42, 143–166.

Garau, M., Slater, M., Vinayagamoorhty, V., Brogni, A., Steed, A., & Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (pp. 529–536). Ft. Lauderdale, FL.

Goldsmith, T. R., & LeBlanc, L. A. (2004). Use of technology in interventions for children with autism. Journal of Early and Intensive Behavior Intervention, 1(2), 166–178.

Guiterrez, K., & Rogoff, B. (2003). Cultural ways of learning: Individual styles or repertoires of practice. Educational Researcher, 35(5), 19–25.

Handy, C. (1995). Thinking about : Trust and the virtual organization. Harvard business review, 73(3), 15.

Harel, I., & Papert, S. (1991). Constructionism : Research reports and essays, 1985–1990. Norwood, N.J.: Ablex.

Kahn, P., Ishiguro, H., Friedman, B., & Kanda, T. (2006). What is a human? Toward psychological benchmarks in the field of human-robot interaction. Proceedings15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06) (pp. 364–371). Hatfield, UK.

Koda, T., & Maes, P. (1996). Agents with faces: The effects of personification of agents. Proceedings 5th IEEE International Workshop on Robot and Human Communication (pp. 189–194). Tsukuba, Japan.

Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K., & Cassell, J. (in press). Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida (Ed.), Engineering approaches to conversational informatics. New York: Wiley.

MacDorman, K. F., Minato, T., Shimada, M., Itakura, S., Cowley, S., & Ishiguro, H. (2005). Assessing human likeness by eye contact in an android testbed. 27th Annual Conference of the Cognitive Science Society. Stresa, Italy.

Melson, G. F., Kahn, P. H., Beck, A. M., Friedman, B., Roberts, T., & Garrett, E. (2005). Robots as dogs: Children's interactions with the robotic dog AIBO and a live australian shepherd. Proceedings of CHI Conference on Human Factors in Computing Systems (pp. 1649–1652). Portland, OR.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. Annual Meeting of the Association for Computational Linguistics (pp. 553–561). Sapporo, Japan.

Nass, C., Isbister, K., & Lee, E.-J. (2000). Truth is beauty: Researching embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), Embodied conversational agents (pp. 374–402). Cambridge, MA: MIT Press.

Parsons, S., Mitchell, P., & Leonard, A. (2004). The use and understanding of virtual environments by adolescents with autistic spectrum disorders. Journal of Autism and Developmental Disorders, 34(4), 449–466.

Pellegrini, A., Galda, L., Bartini, M., & Charak, D. (1998). Oral language and literacy learning in context: The role of social relationships. Merrill-Palmer Quarterly, 44(1), 38–54.

Pinkard, N. (1999). Lyric reader: An architecture for intrinsically motivating and culturally relevant reading environments. Interactive Learning Environment, 7(1), 1–30.

Robins, B., Dickerson, P., Stribling, P., & Dautenhahn, K. (2004). Robot-mediated joint attention in children with autism: A case study in robot–human interaction. Interaction Studies, 5(2), 161–198.

Scassellati, B. (2002). Theory of mind for a humanoid robot. Autonomous Robots, 12(1), 12.

Tartaro, A., & Cassell, J. (2006). Authorable virtual peers for autism spectrum disorders. Combined Workshop on Language-Enabled Educational Technology and Development and Evaluation fo Robust Spoken Dialogue Systems at the 17th European Conference on Artificial Intelligence (ECAI06). Riva del Garda, Italy.

Trevarthen, C. (1987). Sharing makes sense: Intersubjectivity and the making of an infant's meaning. In R. Steele & T. Threadgold (Eds.), Language topics: Essays in honour of Michael Halliday (pp. 177–200). Amsterdam: John Benjamins.

Wang, A., & Cassell, J. (2003). Co-authoring, corroborating, criticizing: Collaborative storytelling for literacy learning. Vienna Workshop: Educational Agents — More than Virtual Tutors. Vienna, Austria.

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. SIGCHI Conference on Human Factors in Computing Systems (pp. 1–10). San Jose, CA.

Authors' address

Justine Cassell and Andrea Tartaro
Center for Technology and Social Behavior
Northwestern University
2240 Campus Drive
Evanston, IL 60208

About the authors

Justine Cassell is Director of the Center for Technology and Social Behavior, and a Professor in the departments of Communication Studies and Electrical Engineering and Computer Science at Northwestern University. Justine Cassell holds undergraduate degrees in comparative literature from Dartmouth and in *lettres modernes* from the Université de Besançon, France. She holds an M.Phil in Linguistics from the University of Edinburgh and a Ph.D. from the University of Chicago in Linguistics and Psychology. Her research interests include embodied conversational agents, natural language generation, and learning technologies.

Andrea Tartaro is a doctoral student in the Technology and Social Behavior joint Ph.D. program in computer science and communication studies at Northwestern University. She received her M.S. in computer science from Northwestern University in 2005, her M.A. in instructional technology from Teachers College, Columbia University in 2003, and her B.A. in computer science from Brown University in 1999. Her research interests include human–computer interaction, technology for special populations, and learning technologies.