(in press) in Jessica Riskin (ed.)
The Sistine Gap: Essays on the History and Philosophy of Artificial Life.

Body Language: Lessons from the Near-Human[i]

Justine Cassell
Northwestern University

> The story of the automaton had struck deep root into their souls and, in fact, a pernicious mistrust of human figures in general had begun to creep in. Many lovers, to be quite convinced that they were not enamoured of wooden dolls, would request their mistresses to sing and dance a little out of time, to embroider and knit, and play with their lapdogs, while listening to reading, etc., and, above all, not merely to listen, but also sometimes to talk, in such a manner as presupposed actual thought and feeling. (Hoffmann 1844)

1 Introduction

It's the summer of 2005 and I'm teaching a group of linguists in a small Edinburgh classroom. The lesson consists of watching intently the conversational skills of a life-size virtual human projected on the screen at the front of the room. Most of the participants come from formal linguistics; they are used to describing human language in terms of logical formulae, and usually see language as an expression of a person's intentions to communicate and from there issued directly out of that one person's mouth. I, on the other hand, come from a tradition that sees language as a genre of social practice, or interpersonal action, situated in the space between two or several people, emergent and multiply-determined by social, personal, historical, and moment-to-moment linguistic contexts, and I am as likely to see language expressed by a person's hands and eyes as mouth and pen. As a graduate student pursuing a dual Ph.D. in linguistics and psychology in the 1980s I had felt profoundly inadequate in the presence of these scholars: their formalized theories belong to a particular kind of technical discourse that is constructed in opposition to every-day language (Agre 1992) and that had seemed more scientific than my messy relational and embodied understanding of how language looks. Those feelings of inadequacy – paired with the real-life experience of having articles rejected from mainstream journals and conferences – led me to try to formalize or ‘scientify’ my work, undertaking a collaboration with computer scientists in 1993 to build a computational simulation of my hypotheses, a simulation that took the form of virtual humans who act off of a “grammar” of rules about human communication. In turn, that simulation has, in the manner of all iconic representations, turned out to both reveal and obscure my original goals, depending on what the technical features of the model can and cannot handle. And the simulation has, like many scientific instruments, taken on a life of its own – almost literally in this instance – as the virtual human has come to be a playmate for children, a teaching device for soldiers, and a companion on cell phones, a mode of interacting with computers as well as a simulation that runs on computers.

But back to the classroom in Edinburgh. In the intervening 15 years since graduate school, I have armed myself with a “sexy” demo to show other scientists and, doubtless quite independently, times have changed and the notion that language is embodied is somewhat more accepted in linguistics today. And so these formal linguists have chosen to attend the summer school class on “face-to-face pragmatics” that I am co-teaching. In the conversation today I’m trying to convince them of two points: that linguists should study videotapes and not just audiotapes, and that we can learn something important about human language by studying *embodied conversational agents* - fake humans who are capable of carrying on a (very limited) conversation with real humans -- such as the one we call NUMACK, who is depicted in Figure 1.

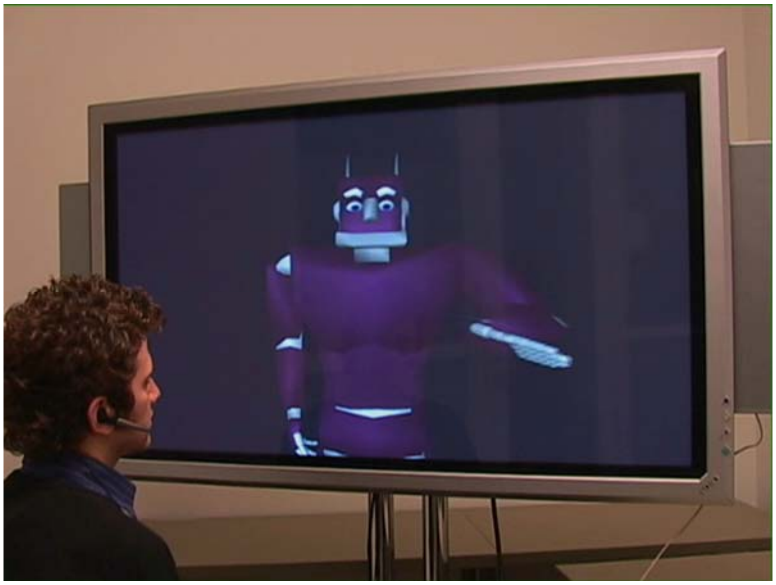


Figure 1: NUMACK, the Northwestern University Multimodal Autonomous Conversational Kiosk, giving directions to a real human

I show the students a brand new video of NUMACK (the **N**orthwestern **U**niversity **M**ultimodal **A**utonomous **C**onversational **K**iosk) interacting with a real human, a simulation of our very latest work on the relationship between gesture and language during direction-giving. On the basis of an examination of 10 people giving directions to a particular place across campus, my students and I have tried to extract generalities at a fine enough level of detail to be able to understand what the humans are doing, and to use that understanding to program our virtual humans to give directions in the same way as humans do. The work exemplified in this particular video has concentrated on the shape of the people's hands as they give directions, and on what kind of information they choose to give in speech and what kind in gesture. I'm excited to share this work, which has taken over a year to complete - moment by moment investigations into the minutiae of human gesture and language extracted from endless examinations of videotapes that show four views of a conversation (as shown in Figure 2), followed by complicated and novel implementations of a computer system that can behave in the same way. In fact, this will be the first time I see the newly updated system myself, as I’ve been traveling and my graduate students finished up the programming and filmed the demo.

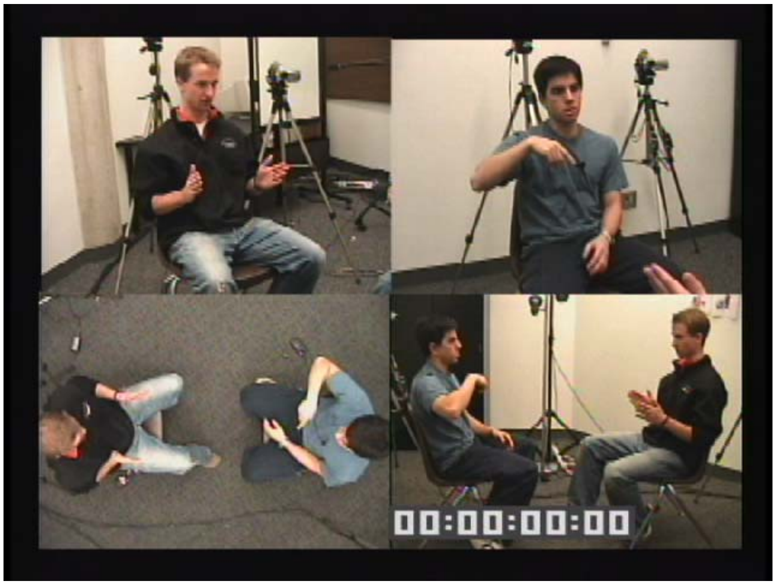


Figure 2: Analysis of Videotapes allows us to draw generalizations about human – human direction-giving

The Edinburgh linguists and I together watch the video of NUMACK giving directions to a person and it looks terrible! The small group of students tries to look down so as not to reveal that they don't think this is a fitting culmination of one year of work. I break the silence and say, "it looks ridiculous! Something is really off here. What is wrong with it? Can anybody help me figure out why it looks so non-human?" The students look surprised – after all, NUMACK looks non-human along hundreds of dimensions (starting with the fact that it is purple). Used as they are to seeing impeccably animated characters in movies and on webpages, they have expected to hurt my feelings by criticizing the virtual human’s poor rendering of reality. But, as we watch the video over and over again, what becomes salient is the way in which NUMACK’s interaction violates our intuitions about how direction-giving should look. After 3 or 4 viewings one of them suggests that the two hands of the computer-programmed agent operate independently in giving directions. The virtual human says, "take a right" and gestures with his right hand. He then says, "Take a left" and gestures with his left hand. I've never thought about this before, but in looking at the robot I am struck by the fact that we humans don't do that - we must have some kind of cohesion in our gestures that makes us use the same hand over and over for the same set of directions. Another student points out that the virtual human describes the entire route (roughly 14 “turn left”, “turn right”, “go straight ahead” kinds of segments) at once, with only a “uh huh” on the part of the real human – no real human would do that – the directions are too long and couldn’t possibly be remembered in their entirety.

I am thrilled and once again amazed at how much I learn about human behavior when I try to recreate it – in particular when, and because, my imitations are partial and imperfect. Only when I try to reproduce the processes in the individual that go into making embodied language, do I get such a clear picture of what I don't yet know. For example, here I have realized that we will need to go back to our 10 real human direction-givers and look at their choice of hands – can I draw any generalizations about the contexts in which they use their right hand or their left? When is the same hand used over and over, and when do they switch to a different hand? Likewise, we will need to go back to our real human direction-givers and look further at the emergent properties of the directions. What behaviors signal to the direction-giver when to pause and when to continue, when to elaborate and when to repeat? What embodied and verbal actions serve to alert the two participants to that the message has been taken up and understood, and the next part of the message can be conveyed? I am also struck once again at the extent to which people are willing to engage with the virtual human, both as participants in a conversation about how to get to the campus chapel, and as participants in a conversation about the holes in our theory of the relationship between verbal and nonverbal elements in conversation.

I have learned something about the particularities of human communication here despite the fact that what I am viewing is a freak of artificial nature - a virtual human that is both generic and very particular, general and very detailed. In fact, for the experiment to work, we depend in part on the not-so-laudable schemas and expectations of our viewers and ourselves – that there can be such a thing as the unmarked or generic human, which probably entails, for a direction-giving robot, that it is male and humanoid (albeit purple) and that its voice is Caucasian and American. As Nass & Brave (2005) point out, violating cultural assumptions about expertise and gender or race produces distrust on the part of users. In the art world, Lynn Hershman Leeson, among others, has violated exactly these assumptions by synthesizing an infinitely smart female bot whose body is present only in certain contexts, and who reproduces herself. But, in the current case, these largely unconscious assumptions on the part of the scientists examining the simulation are what allow them to identify as failings not a lack of personality or cultural identity in the virtual human, but simply that the hands are not synchronized. And thus, in this simulation, I have learned something about human communication despite all of the ways in which this virtual human is not very human at all. Below I will return to question these assumptions, but for the moment let us return to the fundamental questions that guide this work.

AI investigators and their acolytes, like automata makers before them, ask, “Can we make a mechanical human? (or, in the weaker version “a human-like machine”)” I would rather ask “what can we learn about humans when we make a machine that *evokes* humanness in us – a machine that acts human enough that we respond to it as we respond to another human? (where I mean both responds to us in our status of interlocutor, or of scientist)” Some researchers are interested in stretching the limits and capabilities of the machine, or interested in stretching the limits of what we consider human by building increasingly human machines. Such is the case for the work described by Evelyn Keller in this volume (Keller in press). In my own work, at the end of the day I’m less interested in the properties of machines than in the properties of humans. For me there are two kinds of “ah ha!” moments: to learn from my successes by watching a person turn to one of my virtual humans and unconsciously nod and carry on a conversation replete with gestures and intent eye gaze. And to learn from my failures by watching the ways in which the real human is uncomfortable in the interaction, or the interaction looks wrong, as I illustrated in the Edinburgh classroom. These simulations serve as sufficiency proofs for partial theories of human behavior – what Keller has described as the second historical stage in the use of simulation and computer modeling (Keller 2003) – and thus my goal is to build a virtual human to whom people can’t stop themselves from reacting *in human-like ways*, about whom people can’t prevent themselves from applying native speaker intuitions. And key to the enterprise is the fact that those theories of human behavior and those native speaker intuitions refer to the whole body, as it enacts conversations with other bodies in the physical world.

In the remainder of this chapter I’m going to talk about my work on one particular kind of virtual human called an Embodied Conversational Agent (ECA), in terms of its duality as a simulation and as an interface. That is, I will describe how these virtual humans have allowed me to test hypotheses about human conversation, and what they have taught me by their flaws. But I will also describe the life that ECAs have acquired when they leave the lab – the kinds of functions that companies and research labs have put them to. In this way, I hope to illuminate the kinds of conversations that these virtual humans engage when scientists use them as tools to study conversational phenomena, and when everyday people use them as tools to access information, dial phone numbers, learn languages etc.

2 Embodied Conversational Agents as Conversational Simulations

Just to be clear about our terms, *Embodied Conversational Agents* (ECAs) are cartoon-like, often life-size, depictions of virtual humans that are projected on a screen. They have bodies that look more-or-less human-like, they are capable of initiating and responding in (very limited) conversations (in pre-set domains) with real humans, and they have agency in the sense that they behave autonomously, in the moment of their deployment, without anybody pulling the strings. Of course, this agency relies on a prior pre-set network of interactions between their inventors, their users, and the sociotechnical context of their deployment. As a point of contrast, consider chat bots or chatterbots. Chat bots (such as the popular Alice, http://www.alicebot.org/, which readers can try out for themselves) rely on a mixture of matching input sentences to templates, stock responses and conversational tricks (such as “what makes you say X [where X is what the user typed in]” or “I would need a more complicated algorithm to answer that question” when they don’t understand). Chat bots are increasingly employed by artists such as Lynn Hershman Leeson, STELARC or Kirsten Geisler because they are relatively easy to program and thus allow the artist to concentrate on the aesthetic experience s/he wishes to provoke in the viewer. Chat bots often communicate with viewers only through text, but when embodied, generally they only have a head, and a head that displays only the most rudimentary of behaviors (blinking, looking left and right). Embodied Conversational Agents, on the other hand are by definition models of human behavior, which means that at least along some dimension they must function in the same way humans do. Thus, Wang et al’s (Wang et al. 2005) pedagogical agent and Walker et al’s (Walker, Cahn, and Whittaker 1997) virtual actor both rely on Brown and Levinson’s theory of politeness and language use (Brown and Levinson 1987). Poggi and Pelachaud (Poggi and Pelachaud 2000) base the facial expressions of their ECA on Austin’s theory of performatives (Austin 1962). Likewise, ECAs are full functioning Artificial Intelligence systems in the sense that they understand language by composing meanings for sentences out of the meanings of words, they deliberate over an appropriate response, deliver the response, and then remember what they said so as to make the subsequent conversation coherent. They mostly have both heads and bodies, and their behavior is based on an observation of human behavior.

Figure 3 shows an ECA named REA (for Real Estate Agent) who was programmed on the basis of a detailed examination into the behavior of realtors and clients. Over a period of roughly 5 years, various graduate students, post-docs and colleagues in my research group studied different aspects of house-buying talk, and then incorporated their findings into the ECA. Hao Yan looked at what features of a house description were likely to be expressed in hand gestures, and what features in speech (Yan 2000). Yukiko Nakano discovered that posture shifts were correlated with shifts in conversational topic and shifts in whose turn it was to talk (Cassell et al.

2001). Tim Bickmore examined the ways in which small talk was employed to establish trust and rapport between realtor and client (Bickmore and Cassell 1999). Earlier work by Scott Prevost on intonation (Prevost 1996), and by Obed Torres on patterns of eye gaze (Torres, Cassell, and Prevost 1997) also went into the implementation. As our research into human conversation progressed, we also came to better understand some of the overall properties of human conversation, and those insights were also incorporated.

The result was a virtual woman who tried to sell a home to whoever approached her. A small camera on top of the screen allowed her to detect the presence of real humans and initiate a conversation with them. Her knowledge of appropriate realtor conversation led her to ask questions about the person’s housing needs and then nod, seem to reflect, and pull up data on appropriate properties, describing them using a combination of descriptive hand gestures, head movements and spoken language.

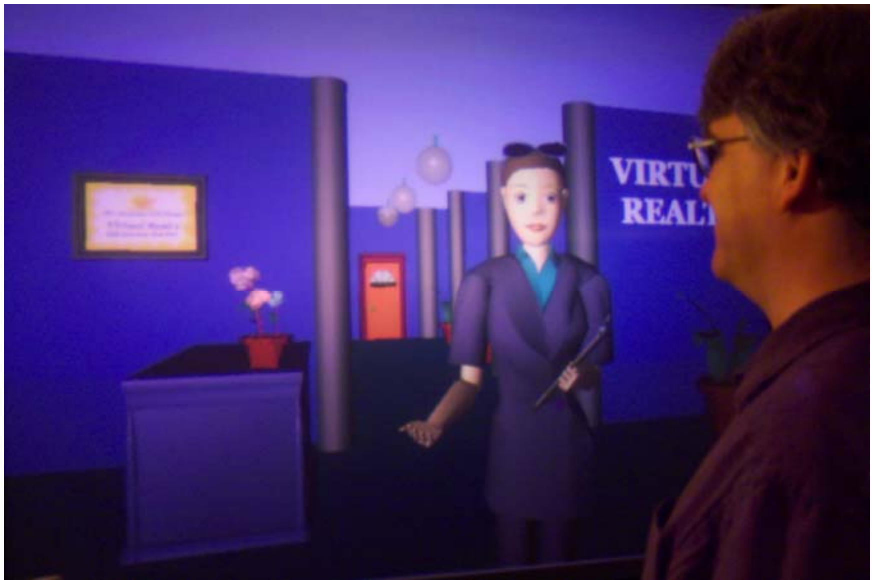


Figure 3: REA, the virtual Real Estate Agent

3 A Sample Interaction with an ECA

Figure 4 illustrates an actual scenario between a human user and REA.

Mike approaches the projection screen. Rea is currently turned side on and is gazing idly about. As Mike moves within range of the two cameras mounted above the screen, Rea turns to face him and says:

REA> Hello. How can I help you?

MIKE> *I'm looking to buy a place near MIT.*

Rea nods, indicating that she is following.

REA> I have a house to show you.

A picture of a house appears on-screen behind Rea who blinks and looks at the house and then at Mike.

REA> It’s in Somerville.

MIKE> Tell me about it.

Rea briefly looks up and away while she plans what to say

REA> It’s big.

Rea makes an expansive gesture with her hands.

Mike brings his hands up as if he is about to speak and so Rea does not continue; instead waiting for him to speak.

MIKE> Tell me more about it

REA> Sure thing. It has a nice garden.

Rea sketches a curved gesture with her hands indicating that the garden extends along two sides of the house

MIKE> How far is it?

REA> It is five minutes to the Porter Square T station.

Rea makes it clear that it is five minutes ***on foot*** *from the T station by making a walking gesture with her fingers*

MIKE> How big is the house?

REA> It has four bedrooms, three bathrooms. . .

And the house tour continues.

Figure 4: Transcript of an interaction between REA and a person

4 Conversational Properties

As well as demonstrating the use of several conversational modalities, such as speech, hand gestures, and head movements, in this example Rea is engaging in some very subtle human-like behavior that demonstrates four of the key properties of human face-to-face conversation. Those four properties are (1) the distinction between 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; (4) the deployment of each modality to do what it does best. Our insights into each of these properties has come though the cycle of watching real humans, attempting to model what we see in virtual humans, observing the result or observing people interacting with the result.

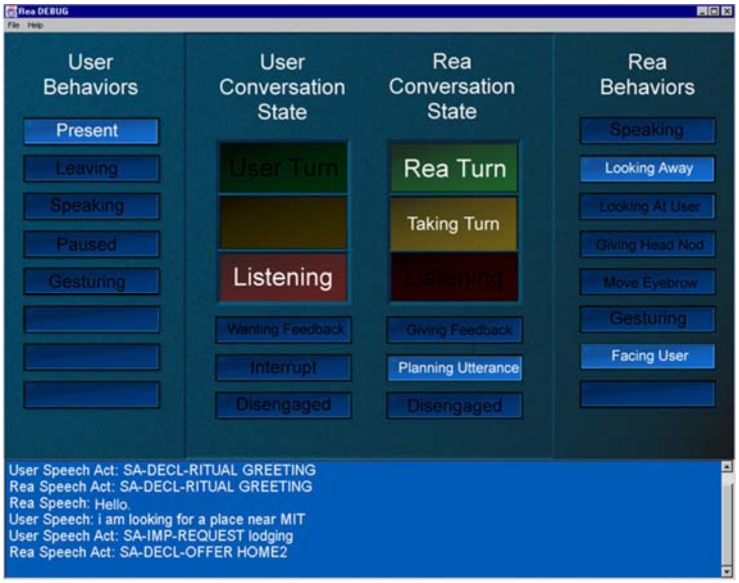
4.1 Division between Propositional and Interactional Functions

Some of the things that people say to one another move the conversation forward, while others regulate the conversational process. *Propositional* information corresponds to the content (sometimes referred to as transmission of information) and includes meaningful speech as well as hand gestures that represent something, such as punching a fist forward while saying “she gave him one” (indicating that the speaker’s meaning is that she punched him, and not that she gave him a present). Interactional information regulates the conversational process and includes a range of non-verbal behaviors (quick head nods to indicate that one is following, bringing one’s hands to one’s lap and turning to the listener to indicate that one is giving up one’s turn) as well as sociocentric speech ("huh?", "do go on"). It should be clear from these examples that both functions may be filled by either verbal or non-verbal means. Thus, in the dialogue excerpted above, Rea’s non-verbal behaviors sometimes contribute propositions to the discourse, such as the gesture that indicates that the house in question is five minutes *on foot* from the T stop, and sometimes regulate the interaction, such as the head-nod that indicates that Rea has understood Mike’s utterance.

4.2 Distinction between Function and Behavior

When humans converse, few of their behaviors are *hard-coded*. That is, there is no mechanism or database “look-up table” that gives the appropriate response for every possible conversational move on the part of one’s partner. Every day we hear thousands of phrases that we have never heard before, assembled through the infinite creativity of language use, and we reply to each of these phrases in just a couple of milliseconds, with an equally creative response. Gestures and head movements are no more likely to be routinized – head nods will look different if we are looking up at a taller interlocutor or down at somebody short, if we are wearing a hat or bareheaded. And other than the small number of culturally meaningful gestures (such as “V for victory”, or “up yours”), gestures display a greater variety across people and even within one person across time. In observing human-human conversation our group discovered that speakers do not always nod when they understand. Instead they sometimes signal that they are following along by making agreement noises such as “uh huh”. In our simulation of this behavior, then, instead of hard-coding, the emphasis is on identifying the high level structural elements that make up a conversation. These elements are then described in terms of their role or *function* in the exchange. Typical discourse functions include *conversation invitation*, *turn taking*, *providing feedback*, *contrast and emphasis*, and *breaking away*. Each function can be filled through a number of different behaviors, in one or several modalities. The form given to a particular discourse function depends on, among other things, current availability of modalities such as the face and the hands, type of conversation, cultural patterns and personal style.

In the REA embodied conversational agent, Rea generates speech, gesture and facial expressions based on the current conversational state, the conversational function she is trying to convey, and the availability of her hands, head and face to engage in the desired behavior. For example, when the user first approaches Rea (“User Present” state), she signals her openness to engage in conversation by looking at the user, smiling, and/or tossing her head. Figure 5 shows a visualization of REA’s internal state with respect to conversational behaviors and conversational states



```
User Speech Act: SA-DECL-RITUAL GREETING
Rea Speech Act: SA-DECL-RITUAL GREETING
Rea Speech: Hello.
User Speech: i am looking for a place near MIT
User Speech Act: SA-IMP-REQUEST lodging
Rea Speech Act: SA-DECL-OFFER HOME2
```

Figure 5: Visualization of ECA and Human Conversational State

4.3 Importance of Timing

The relative timing of conversational behaviors plays a large role in determining their meaning. That is, for example, the meaning of a nod is determined by where it occurs in an utterance, all the way down to the 200 millisecond scale (consider the difference between “you did a [great job]” (square brackets indicate the temporal extent of the nod) and “you did a [. . .] great job”). Thus, in the dialogue above, Rea *says* “it is five minutes from the Porter Square T station” at exactly the same time as she *performs* a walking gesture. If the gesture occurred in another context, it could mean something quite different; if it occurred during silence, it could simply indicate Rea’s desire to take the turn.

Although it has long been known that the most effortful part of a gesture co-occurs with the part of an utterance that receives prosodic stress (Kendon 1972), it wasn’t until researchers needed to generate gestures along with speech in an ECA – and therefore needed to know the details of the context in which one was most likely to find contentful gestures -- that it was discovered that the gesture is most likely to co-occur with the *rhematic* (Halliday 1967) or new contribution part of an utterance. This means that if a speaker is pointing to her new vehicle and saying “this car is amazingly comfortable. In fact, this car actually 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 car has already been mentioned. Therefore, the speaker would be most likely to produce hand gestures with “amazingly comfortable” and “reclining seats”).

4.4 Using the Modalities to do what they do Best

In e-mail, we are obliged to compress all of our communication goals into textual form (plus the occasional emoticon). In face-to-face conversation, on the other hand, humans have many more modalities of expression at their disposal, and they depend on each of those means, and various combinations amongst them, to communicate what they want to say. They use gestures to indicate things that may be hard to represent in speech, such as spatial relationships among objects (Cassell, Stone, and Yan 2000), and they depend on the ability to simultaneously use speech and gesture in order to communicate quickly. In this sense, face-to-face conversation may allow us to be maximally efficient or, in other instances, to use conversation to do other kinds of work than information transmission (for example, we may use the body to indicate rapport with others, while language is getting task work done). In the dialogue reproduced above, Rea takes advantage of the hands’ ability to represent spatial relations among objects and places by using her hands to indicate the shape of the garden (sketching a curved gesture around an imaginary house) while her speech gives a positive assessment of it (“it has a nice garden”). However, in order to produce this description, the ECA needs to know something about the relative representational properties of speech and gesture, something about how to merge simultaneous descriptions in two modalities, and something about what her listener does and does not already know about the house in question.

The need to understand how speech and gesture and facial/head movements can be produced together by ECAs has forced me to design experimental and naturalistic methodologies to look at the nature of the interaction between modalities, and has resulted in significant advances in my theorizing about the relationship between speech and gesture in humans. Thus, for example, in my current work, with not REA but the purple virtual robot NUMACK as a simulation, Paul Tepper, Stefan Kopp and I have become interested in the seeming paradox of how gesture communicates, given that there are no standards of form in spontaneous gesture – no consistent form-meaning mappings. Some gestures clearly depict visually what the speaker is saying verbally, and these gestures are known as *iconics*. But, what is depicted on the fingers, and its relationship to what is said, can be more or less obvious. And two speakers’ depiction of the same thing can be quite different. An example comes from the comparison of two people describing the same landmark on Northwestern University’s campus: an arch that signals the beginning of the campus, and that lies at the intersection of Sheridan Road and Chicago Avenue. In order to collect these data, we hid prizes in various spots on campus, and asked one student, who knew where the prize was hidden, to give directions to the prize to a second student. If the second student was successful in finding it, the two shared the prize (and both were entered into a drawing for an iPod, probably the most motivating feature of the experiment!). The direction-giving was videotaped using 4 cameras trained on different parts of the bodies of the two speakers, as described above (and shown in Figure 2), and then each gesture was transcribed, along with the speech that accompanied it, for further study. One speaker in the experiment, describing directions to a church near the arch, said “go to the arch” and with his fingertips touching one another with the fingers pointing upwards, made a kind of teepee shape. In this instance, the gesture seemed to indicate a generic arch. Compare that gesture to the following, made by another participant in the experiment who, while referring to that same arch, said “you know the arch?” but this time, although his fingertips were touching one another, the fingers were pointing towards the listener and the thumbs up, making the shape of a right angle. In this instance, the gesture seems to indicate . . . what? An arch lying on its side?? It makes, in fact, no sense to us as observers . . . unless we know that the arch is located at the *right angle* formed by Sheridan Road and Chicago Avenue. And this interpretation of the gesture is supported by the speaker’s next utterance, “it’s located at the corner of Sheridan”. So, in the absence of the relatively stable form-meaning pairing that language enjoys (the same image may not be evoked for both of us, but when I say “right angle” I can be relatively sure that you will interpret it to mean something along the lines of *a* right angle), how do gestures communicate? The answer to this question (which is outside the scope of this chapter, but has to do with the fact that gestures have a kind of interpretive flexibility, and have meanings only in situated contexts) resulted both in a new computational architecture whereby gesture and speech are computationally generated together, and a new way of understanding of how gestures communicate among humans.

5 Translating Conversational Properties into Computational Architectures

The four conversational properties discussed in the previous section gave rise in 2000 to a computational architecture that looks as follows:

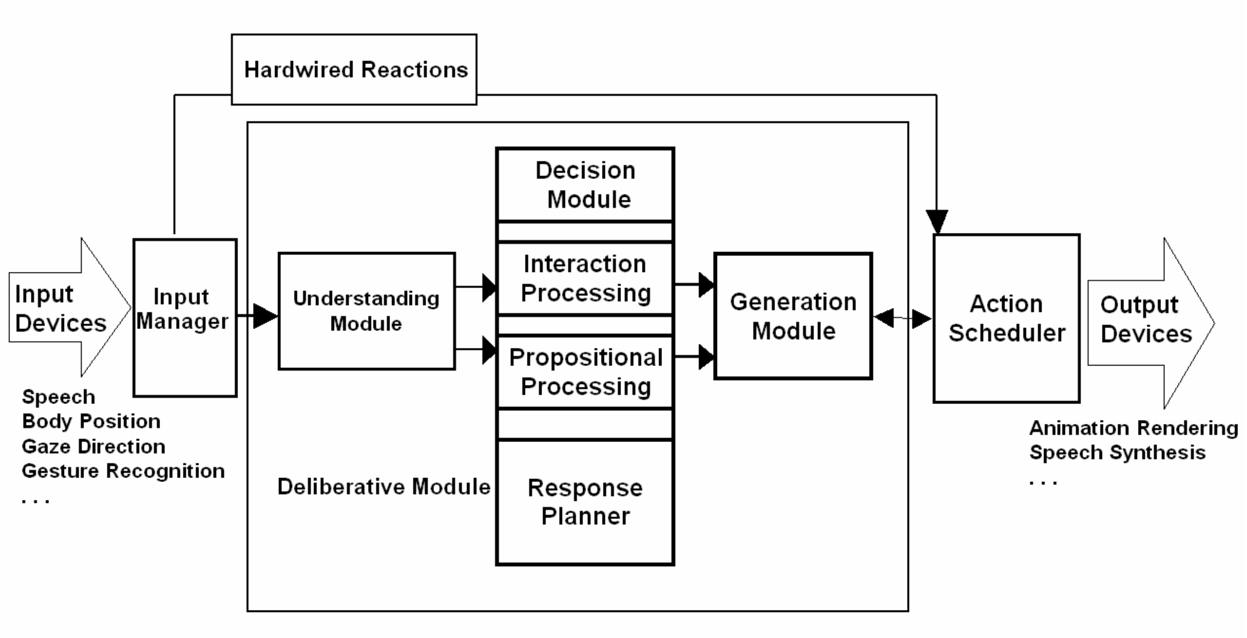


Figure 6: Computational Architecture of an ECA

As this diagram makes clear, and like many systems in Artificial Intelligence, ECAs are largely linear and devoid of contingent functionality – the real human asks a question, which is collected by the input modules of the system (cameras to view the speaker’s gestures and posture, microphones to hear the speech) and then interpreted into a unified understanding of what the speaker meant. In turn, that understanding is translated into some kind of obligation to respond. That response is planned out first in “thought” or communicative intention, and then in speech and movements of the animated body, face, and hands through the use of a speech synthesizer, computer graphics engine and various other output modes. Meanwhile, so as not to wait for all of that processing to be completed before a response is generated, a certain number of hardwired responses are sent to be realized: head nods, phatic noises (mmm, uh huh) and shifts of the body.

The linear nature of this architecture is one of the constraints imposed by the scientific instrument – like trying to cut out circles with straight blades. When I first began to collaborate with computer scientists in 1993-1994 to build a virtual human I asked them to build one that was responsive to itself and to its interlocutor in a number of ways. I told them that I wanted the virtual human to be able to see its own hands, and from what it saw decide what it wanted to say in the moment – the way humans often do, such as when they can’t recall a word until they make the gesture for it. And I told them I wanted some kind of entrainment or accommodation between the different participants in the conversation, such that their language and gesture grew increasingly alike, as they came to mirror one another. The response was incredulity and a request for me to be better informed before I went asking for features. The goal, I was told, was autonomy and not co-dependence. Of course, as Suchman has pointed out about other work in Artificial Intelligence, this means that we have not produced a truly conversational agent, since “interaction is a name for the ongoing, contingent co-production of a shared social/material world” (Suchman 2003). But the kinds of interdependence that we wish to simulate are hard to achieve given our current models.

In general terms, however, building ECAs has forced researchers of human behavior to attend to the integration of modalities and behaviors in a way that merges approaches from fields that for the most part do not speak to one another: ethnomethodological interpretive and holistic studies of human communication with psycholinguistic experimental isolative studies of particular communicative phenomena. To build a human entails understanding the context in which one finds each behavior – and that context is the other behaviors.

During that first collaboration with computer scientists in 1993-1994, when we were building the very first of these animated embodied conversational agents, each of the parts of the body was being implemented by a different researcher. Catherine Pelachaud was writing the algorithms to drive the character’s facial movements – head nods, eye gaze, etc. – based on conversational parameters such as who had the turn. Scott Prevost was writing rules to generate appropriate intonation – the prosody of human language – on the basis of the relationship between the current utterance and previous utterances. I myself was working on where to insert gestures into the dialogue. . . After several months of work, we finally had a working system. In those days, ECAs needed to be “rendered” – they were not real-time – and so with bated breath we ran the simulation, copied it to videodisc, and then watched the video. The result was an embodied conversational agent who looked like he was speaking to very small children, or to foreigners. That is, the resultant virtual human used so many nonverbal behaviors that signaled the same thing, that he seemed to be trying to explain something to a listener who didn’t speak his own language or was just very stupid. This system, called Animated Conversation, was first shown at SIGGRAPH, the largest Computer Graphics conference, in front of an audience of 4000 researchers and professional animators (the folks who build cartoons and interactive characters) and they found it hilarious. To my mind, on the other hand, we had made a huge advance. We had realized that the phenomena of hand gesture, intonation and facial expression were not separate systems, nor was one a “translation” of the others, but instead **had to be** derived from one common set of communicative goals. That was the only explanation for the perception of over-emphasizing each concept through a multiplicity of communicative means. The result fundamentally changed the way we build embodied conversational agents, but it was an advance in understanding human communication as well. It also led to a design methodology that I have relied on ever since, and that is represented in Figure 7. Iteratively, my students and I collect data on human-human conversation, interpret those data in such a way as to build a formal model, implement a virtual human on the basis of the model, confront the virtual human with a real human, evaluate the results, and collect more data on human-human communication if needed (a side effect of this methodology is the need to confront the response of lay viewers to the necessary flaws and lacunae in the implementation, but I try to think of that as character building).

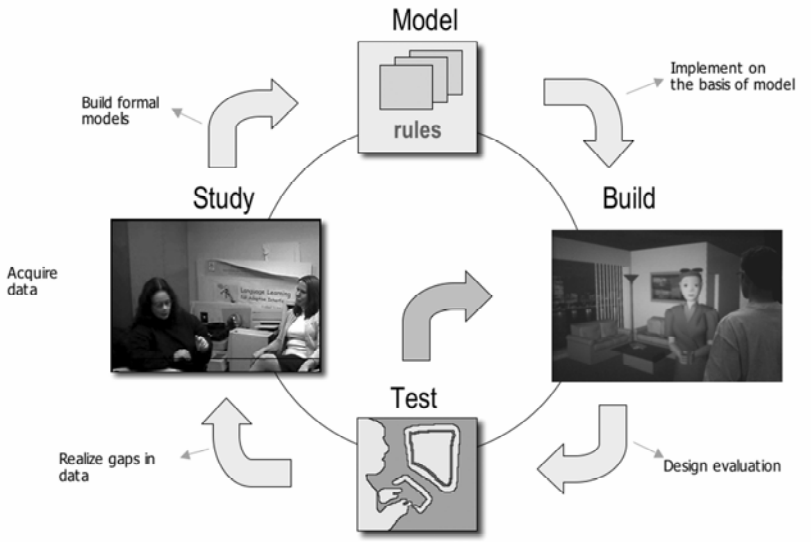


Figure 7: Methodology for Modeling Human Conversation and Building ECAs

It should be reiterated that building a computational system has traditionally demanded a formal or predictive model. That is, in addition to being able to *interpret* why a particular experience occurs in a particular context, one must also be able to *predict* in the future what set of conditions will give rise to a particular experience (Schwartz and Martin 2003) so that one can generate those behaviors in the ECA in response to the appropriate conditions. Unfortunately, predictive models also come with their own baggage, as they tend to underscore the way in which fixed sets of conditions give rise to fixed outputs, as opposed to highlighting the very contingent co-produced nature of human conversation where, on the fly, hearers and speakers influence one another’s language and indeed their very thinking patterns, as Suchman has forcefully argued (Suchman 1997). In this sense, I sometimes worry that building computational simulations of this sort may set back the study of language; that phenomena that cannot yet be modeled in virtual people will be ignored. On the other hand, for the most part, before the advent of embodied conversational agents, computational linguistics and work on dialogue systems (which arose from the Cognitive Sciences -- psychology, linguistics, philosophy, computer science) concentrated on the *propositional* functions of language, which were thought by many linguists to be the primary if not the only function of language. Before ECAs computational models of language were capable only of simulating task talk bereft of social context, and bereft of nonverbal behavior. And given the power of these computational models, perhaps the arrival of ECAs with their attendant attention to the non-informational, and socially-contextualized, functions of language have played some positive role in the Cognitive Sciences.

More hopefully, even, now that there has been a decade of research on Embodied Conversational Agents, several researchers, including myself, are beginning to explore other kinds of computational architectures and techniques that do not require deterministic formal input-output style models of conversation. Probabilistic techniques, such as spreading activation, Bayesian reinforcement learning, and Partially Observable Markov Decision Processes, are being applied to the newest phenomena to be modeled with ECAs. These phenomena, which tend to have more to do with social context than local linguistic context, include the effect of emotion on verbal and nonverbal behavior in conversation (Conati and Zhou 2004; Rosis et al. 2003), the role of personality and cultural differences (Ball and Breese 2000), social influence (Marsella, Pynadath, and Read 2004), etiquette (Bickmore 2004), and relationship-building (Cassell and Bickmore 2002; Stronks et al. 2002).

In all of these implementation experiments, embodied conversational agents are *tools to think with*, much like other computer software and hardware that allows us to evaluate our own performance in the world (Turkle 1995). They allow us to evaluate our hypotheses about the relationship between verbal and nonverbal behavior, and to see what gaps exist in our knowledge about human communication, by seeing ourselves and our conversational partners in the machine. How do we go about evaluating our hypotheses? As described above, we watch the virtual humans and observe our own reactions. But, we also put others in front of these ECAs and examine the differences between their behavior with ECAs and their behaviors with other humans. This second kind of experiment relies on the supposition that correctly implemented virtual humans evoke human-like behavior. In this instance, mechanisms that seem human make us **attribute humanness**/aliveness to them, and that make us **act** human and alive. Successful virtual humans evoke distinctly human characteristics in our interaction with them. The psychological approach to artificial life leads to functional bodies that are easy to interact with, “natural” in a particular sense: they evoke a response.

In an early experiment, for example, Kris Thorisson and I compared reactions to three versions of an ECA called Gandalf (this was 1996, and the ECA consisted of a head with one disembodied hand, as shown in Figure 8). Our goal was to demonstrate, in those early days, that interactional behaviors – that did not move the conversation forward – could be simulated computationally, and that those behaviors in virtual humans would elicit similar behaviors on the part of human interlocutors. An additional goal was to demonstrate that if one were to choose only one set of nonverbal functions to implement computationally, they should be interactional (what we called “envelope”) and not emotional functions. We felt that emotional reactions should be studied only once these very ubiquitous interactional behaviors had been simulated.

In the first version, called “content-only”, the virtual human spoke but used no non-verbal expressions of any kind. An example of an interaction with an agent in the content condition follows:

Gandalf: “What can I do for you?” [*Face looks at user. Eyes do not move.*]
User: “Will you show me what Mars looks like?” [*User looks at Gandalf.*]
Gandalf: “Why not—here is Mars” [*Face maintains orientation. No change of expression. Mars appears on monitor.*]
User: “What do you know about Mars?” [*User looks at map of solar system.*]
Gandalf: “Mars has two moons” [*Face maintains orientation. No change of expression.*]

In the second version, called “content + envelope”, the virtual human spoke and also used eye gaze, eyebrow and head movements and the occasional hand wave to moderate turn-taking, feedback, and other conversational envelope processes. An example of an interaction with this envelope agent follows:

User: “Is that planet Mars?”
Gandalf: “Yes, that’s Mars.” [*Gandalf raises eyebrows and performs beat gesture while saying “yes,” turns to planet and points at it while saying “that is Mars,” and then turns back to face user*.]
User: “I want to go back to Earth now. Take me to Earth.” [*User looks at map of solar system so Gandalf looks at solar system.*]
Gandalf: “OK. Earth is third from the sun.” [*Gandalf turns to planet as he brings it up on the screen, then turns to user and speaks.*]
User: "Tell me more." [*Gandalf takes about 2 seconds to parse the speech, but he knows within 250 ms when the user gives the turn, so he looks to the side to show that he's taking the turn, and his eyebrows go up and down as he hesitates while parsing the user's utterance.*]
Gandalf**:** “The Earth is 12,000 km in diameter.” [*Gandalf looks back at the user and speaks.*]

In the final version, called “content + emotional”, the virtual human spoke and also smiled, frowned and looked puzzled as the occasion warranted. An example of an interaction with an agent in this emotional condition follows:

Gandalf: “What can I do for you?” [*Gandalf smiles when user’s gaze falls on his face, then stops smiling and speaks.*]
User: “Take me to Jupiter.” [*User looks at screen and then back at Gandalf and so Gandalf smiles.*]
Gandalf: “Sure thing. That’s Jupiter” [*Gandalf smiles as he brings Jupiter into focus on the screen.*]
User: [*Looks back at Gandalf. Short pause while deciding what to say to Gandalf.*]
Gandalf: [*Looks puzzled because the user pauses longer than expected. Waits for user to speak.*]
User: “Can you tell me about Jupiter?”

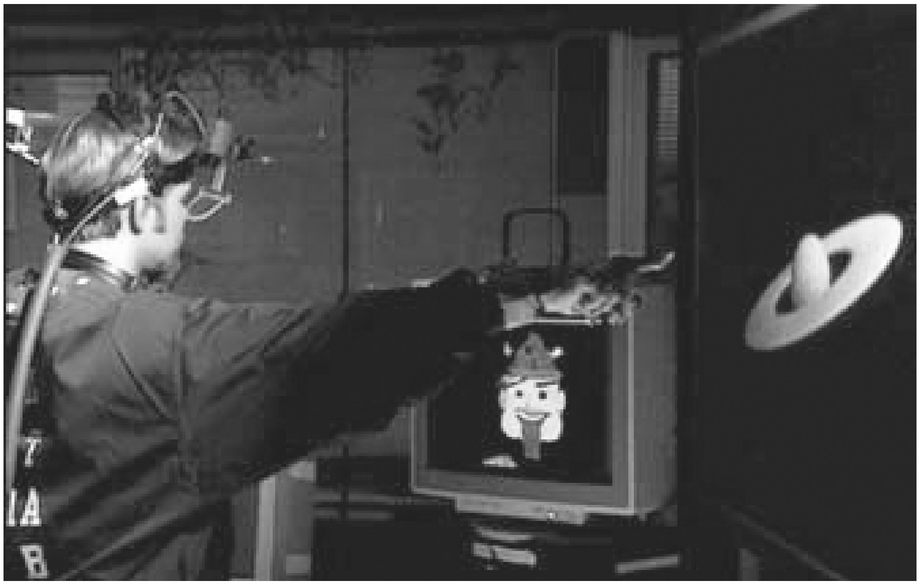


Figure 8: Person interacting with Gandalf

The study consisted in asking people to interact with Gandalf and then examining the real human’s conversational envelope and emotional behaviors during the interaction, as well as asking subjects to fill out a questionnaire assessing “lifelikeness”. What we discovered was that participants tended to mimic the virtual human: if he stood rigid, so did they; if he was animated, so were they. In fact, the people standing in front of the content-only version of Gandalf were most animated in their expressions of frustration – sighs and the occasional request for signs of life (“Gandalf, are you there?”). People interacting with the content + envelope version, on the other hand, started off wary as Gandalf’s head and single hand began to describe the solar system, and then after an utterance or two came to life, gesturing and nodding to Gandalf in much the same was as they had to the experimenter before the experiment started (Cassell and Thorisson 1999). Finally, we discovered no difference in the people’s interaction, nor in their assessment of the ECA, between the content-only version and the content + emotion version.

More recently, Yukiko Nakano and I carried out a study of the role of nonverbal behaviors in grounding, and how these behaviors could be implemented in a virtual human (Nakano et al. 2003). Common ground is the sum of mutual knowledge, mutual beliefs and mutual suppositions necessary for a particular stage of a conversation (Clark 1992). Grounding refers to the ways in which speakers and listeners ensure that the common ground is updated, such that the participants understand one another. Grounding may occur by nodding to indicate that one is following, by asking for clarifications when one doesn’t understand, or by uttering requests for feedback, such as “you know what I mean?” Here too an extensive study of human-human behavior in the domain of direction-giving paved the way for the implementation of an ECA that could ground while giving directions using a map and using hand gestures. And here too we evaluated our work by comparing people’s reactions to two versions of the virtual human, in which one demonstrated grounding behaviors, and the other had the grounding “turned off”. When the behaviors were turned off, the person simply acted as if she were in front of a kiosk and not another human – not gazing at the ECA or looking back and forth between him and the map. When the ECA did engage in grounding behaviors, the human acted strikingly . . . human, looking back and forth between the map and the ECA, as shown in the figure below.

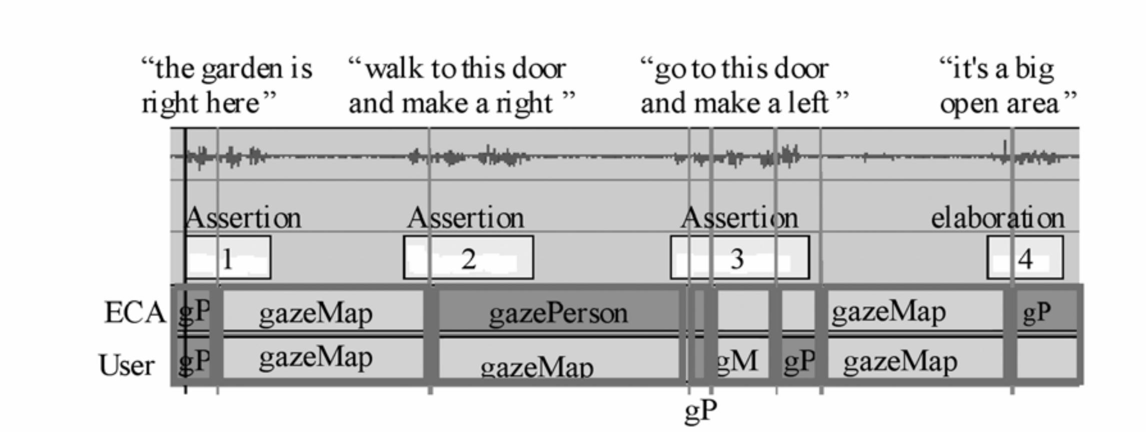
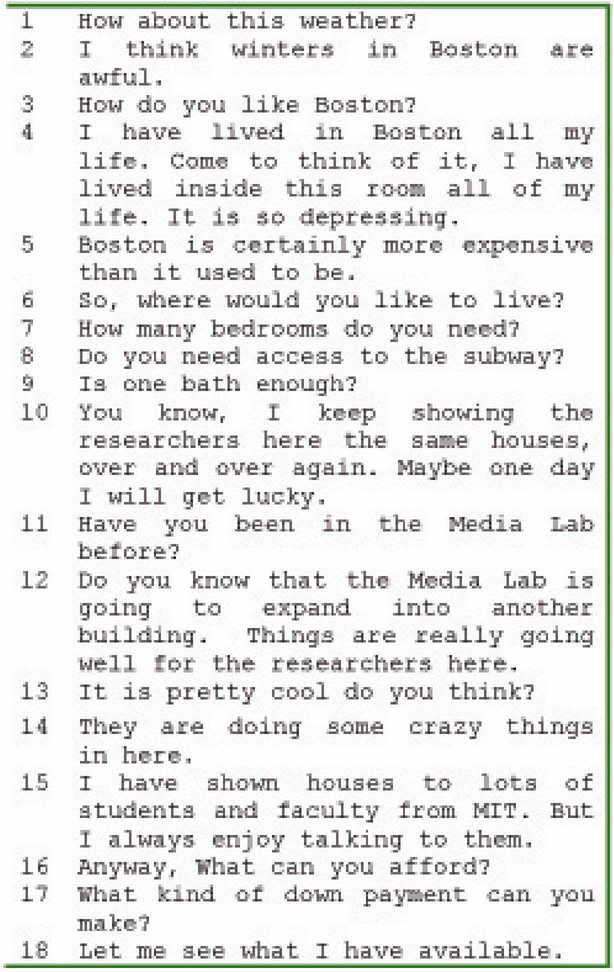


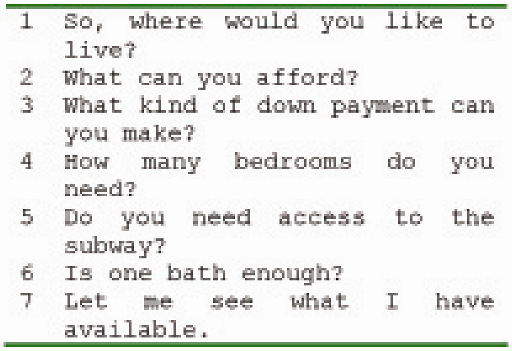
Figure 9: Analysis of grounding behaviors in Human-ECA conversation

A final example, and perhaps the most illustrative of the ways in which ECAs properly constructed, on the basis of theories elaborated from human observation, can elicit human-like behavior (and how this behavior can be illuminating along both positive and negative dimensions) is an experiment in which we endowed REA with social chit-chat skills (Cassell and Bickmore 2002). As mentioned above, Tim Bickmore carried out an extensive study of small talk in realtors and traveling salesmen. The results indicated that small talk was not randomly inserted into conversation, but served specific purposes, including to minimize the potential face threat of personal questions (such as "how much do you earn"). These functions of small talk could be simulated in such a way as to allow us to implement a small-talking realtor, who used chit-chat to smooth the rails of a house-selling transaction with a human. In order to test our model of human conversation, we asked people to interact with one of two versions of the ECA. One used task-talk only, while the other added social chit-chat at key places in the interaction. The two figures below show the actual conversational contributions by REA in the two conditions.



1 How about this weather?
2 I think winters in Boston are awful.
3 How do you like Boston?
4 I have lived in Boston all my life. Come to think of it, I have lived inside this room all of my life. It is so depressing.
5 Boston is certainly more expensive than it used to be.
6 So, where would you like to live?
7 How many bedrooms do you need?
8 Do you need access to the subway?
9 Is one bath enough?
10 You know, I keep showing the researchers here the same houses, over and over again. Maybe one day I will get lucky.
11 Have you been in the Media Lab before?
12 Do you know that the Media Lab is going to expand into another building. Things are really going well for the researchers here.
13 It is pretty cool do you think?
14 They are doing some crazy things in here.
15 I have shown houses to lots of students and faculty from MIT. But I always enjoy talking to them.
16 Anyway, What can you afford?
17 What kind of down payment can you make?
18 Let me see what I have available.

Figure 10: Small Talk in ECAs



1 So, where would you like to live?
2 What can you afford?
3 What kind of down payment can you make?
4 How many bedrooms do you need?
5 Do you need access to the subway?
6 Is one bath enough?
7 Let me see what I have available.

Figure 11: Task-Only Talk in ECAs

The people who interacted with each ECA were asked to evaluate their experience: how natural they felt the interaction to be, how much they liked the ECA, how warm they felt she was, how trustworthy. We also tested the subjects on their own social skills, dividing them into extroverts and introverts using a common psychological scale. The results showed that extroverts preferred the small talk version of the ECA while introverts preferred the ECA to keep to the task (we also discovered that it was difficult to find extroverts among the MIT students, but that’s another story).

An introvert in the small talk condition remarked

> REA exemplifies some things that some people, for example my wife, would have sat down and chatted with her a lot more than I would have. Her conversational style seemed to me to be more applicable to women, frankly, than to me. I come in and I shop and I get the hell out. She seemed to want to start a basis for understanding each other, and I would glean that in terms of our business interaction as compared to chit chat. I will form a sense of her character as we go over our business as compared to our personal life. Whereas my wife would want to know about her life and her dog, whereas I really couldn’t give a damn.

An extrovert in the same condition had a very different response

> I thought she was pretty good. You know, I can small talk with somebody for a long time. It’s how I get comfortable with someone, and how I get to trust them, and understand how trustworthy they are, so I use that as a tool for myself.

Clearly, the people in this experiment are evaluating the ECA’s behaviors in much the same way as they would evaluate a flesh-and-blood realtor. And clearly, our unexamined implementation of the realtor as a woman instead of a man has played into those evaluations, as much as have any of our carefully examined decisions about small talk, hand gestures and body posture. Although our goal was to obtain input into a theory of the role of small talk in task talk, this response from one of REA’s interlocutors effectively demolishes the claim that human identity can be denuded of its material aspects. Much of previous work on responses to ECAs as interfaces has in fact concentrated on exactly this sort of effect, with some researchers advising industry executives to implement a female ECA to sell phone service, but a male ECA to sell cars (cf. Nass and Brave 2005). In response to this unintended research finding in our small talk study, my students and I have begun to use the virtual human paradigm to investigate explicitly which linguistic, nonverbal, and visual cues signal aspects of identity. Some have suggested that the race of ECAs be paired to the putative race of the user; my students and I have begun to look at the complex topic of racial identity, and how a person’s construction of his/her own race, and recognition of the racial identity of others, may be conveyed not just by skin color, but (also) by aspects of linguistic practice, patterns of nonverbal behavior and narrative style (Cassell et al. forthcoming).

6 Embodied Conversational Agents as Interfaces

I’ve alluded to other ways in which ECAs are used, where they serve not as scientific instruments or tools to think with, but interfaces to computers. In this function, ECAs might take the place of a keyboard, screen and mouse – the human speaks to them instead of typing. Or they might represent the user in an online chat room. ECAs can also serve as teachers or tutors in educational software – so-called “pedagogical agents.” Research in this applied science examines whether ECAs are preferable to other modalities of human-computer interaction such as text or speech; what kinds of behaviors make the ECAs most believable, and most effective (as tutors, information retrievers, avatars); and what personas the ECA should adopt in order to be accepted by their users. My students and I have also conducted some of this research, looking at whether virtual children are effective learning companions for literacy skills (Ryokai, Vaucelle, and Cassell 2003), whether people are willing to be represented by ECAs in online conversations (Cassell and Vilhjálmsson 1999), and whether tiny ECAs – small enough to fit on a cell phone – still evoke natural verbal and nonverbal responses in the people speaking with them (Bickmore 2002). Even here, however, our research on virtual peers has led us back to an exploration of human-human communication, as we attempt to identify the features that signal to children that somebody else is a peer, is good friendship material, is worth listening to and telling stories with. In this instance our exploration of the pragmatics of the body has led us to some key features of social interaction – how rapport and friendship are negotiated -- which, in turn, have led us to a better understanding of peer learning. One of our virtual peers is shown in Figure 12.

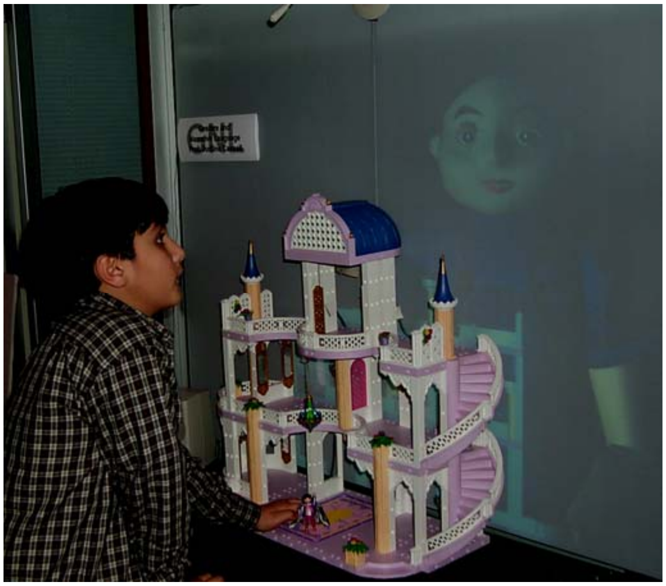


Figure 10: A Child Playing with Sam, the Virtual Peer

Most recently, Andrea Tartaro and I have begun to look at how children with autism can play the role of scientist – learning about the gaps in their knowledge of communication and social interaction by authoring virtual people and watching them interact with others (Tartaro and Cassell in press). Mostly, however, our work is focused on the minutiae of human interaction, and is therefore sometimes less useful to interface designers. In fact, computer scientists sometimes respond to my talks about NUMACK the direction-giving robot by asking “but wouldn’t it just be more effective to display a map on the computer screen and skip the virtual human?” When I respond that such an interface wouldn’t teach us anything about human communication, those same questioners often nod sagely, as if they knew all along that my interest was only in humans. Others have taken the ECA much further as an interface – probably the furthest (and most studied by historians of science and technology) being the Institute for Creative Technologies at the University of Southern California. Funded in equal parts by the Army and Hollywood, the ICT has created a vast immersive videogame-like room geared towards teaching soldiers before they enter the field – what Tim Lenoir has called a “military entertainment complex” (Lenoir 2000).

The development of the ECA from a scientific instrument that simulates human behavior to an attractive interface bears interesting parallels to the history of mechanical automata. Automata makers of the 16th century, such as the one who built the perpetually-praying monk described so elegantly by King in this volume depended on the gaze of the interlocutor to confer lifelikeness on the machine. Automata makers of the 18th century intended to find out in what way the activities of drawing and writing and playing an instrument differed, if at all, when machines performed them (Riskin 1999). In that vein was Droz’s writing boy, whose pen moves across the page just as real writers’ pens move. The ECA that I build today, are likewise a way to compare the conversation among humans with conversation between a human and a human-like machine in order to discover what we know and do not know about human communication, and that simulation only works because of the life conferred on the virtual human by the interlocutor. Mechanical automata of the later 19th century, however, were meant to entertain, and not illuminate. An example of such a pretty virtual body as entertainment is the Pierrot automaton doll that writes – but simply by moving an inkless pen smoothly across a page—while sighing deeply and progressively falling asleep by the lamplight. These latter examples of mechanical humans did sustain relationships with real humans in that humans desired to own the pretty mechanical toys, and were fascinated by them. But in these instances, the gaze of the viewer was one of concupiscence and not interlocutor. Likewise, the tiny virtual human on a cell phone is meant to evoke the greedy desire of the collector more than the unconscious gaze of a partner in conversation.

7 Conclusions

These five-finger exercises in building virtual people have led to advances in what we know about the interaction between verbal and nonverbal behavior in humans, about the role of small talk in task talk, about the kinds of functions filled in conversation by the different modalities of the body, and about how learning is linked to rapport in children. In learning what *must* be implemented in order to make Embodied Conversational Agents evoke a lifelike response, and in learning what the technology can and can’t do at the present time, has also given me a sense of the meaning of humanness through human behavior. It is the ensemble of behaviors, in all of their minuteness and unconscious performance that make a human seem human-like. Flaws and lacunae in that ensemble of behaviors give the scientist interlocutor a sense of what we do not know about human communication. Strengths and continuities in the theory that underlies the implementation lead to a virtual human that evokes human-like behavior in a layperson interlocutor. The sufficiency criterion in Cognitive Science consists of explaining human cognitive activity by showing how a computer program may bring about the same result when the computer is provided with the same input (Newell and Simon 1972). In virtual human simulations, cognitive activity is not sufficient. I know that my model successfully explains human behavior, when it evokes human behavior, because human communicative behavior is intrinsically relational, and cannot be understood without two humans.

To come back to the anecdote with which this essay began, it is important to note the essential role of the physical body in both the study of language, and of social experience (insofar as those might be distinguishable). Language has traditionally been relegated to taking place purely in the head. But, I hope it has been clear from the examples of communicative functions given above that language is spread throughout the whole body – the hands, the torso, the eyes – and across two bodies in interaction. My original goal in building virtual humans was to focus attention on the whole-body aspects of language and from thence to its intrinsically relational nature. As Descartes points out, the difference between real men and those who only have the shape of men exists both in word and movement: imitation and gesture are as constitutive of humanness and social interaction as spoken language.

> ... and suppose there existed machines built in the image of our bodies, and capable of imitating our actions, as far as morally possible, there would still remain two certain tests by which to know that they were not really men. The first is that these automata could

> never use words or other signs in conversation, as we are able to do in order to convey our thoughts to others; for even if we can easily conceive of a machine that can emit the sounds of speech, or that can respond to external action such that, for example, if touched in one particular place it may ask what we wish to say to it; if touched in another it may cry out that it is hurt, and so forth; we nevertheless cannot imagine a machine that can answer to what is said in its presence, as even fools can do. The second test is that even though such machines may carry out actions as well or even more perfectly than we humans can, they still will fail in executing other actions, by which we can discover that they did not act from knowledge, but from a particular arrangement of their organs . . .(Descartes 1953, pp 164-165) (translated by the author).

8 References

Agre, Philip. 1992. Formalization as a Social Project. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition* 14 (1):25-27.

Austin, John. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Ball, Gene, and Jack Breese. 2000. Emotion and Personality in a Conversational Agent. In *Embodied Conversational Agents*, edited by J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, MA: MIT Press.

Bickmore, Timothy. 2002. Towards the Design of Multimodal Interfaces for Handheld Conversational Characters. In *Proceedings of CHI*, at Minneapolis, MN.

———. 2004. Unspoken Rules of Spoken Interaction. *Communications of the ACM* 47 (4):38-44.

Bickmore, Timothy, and Justine Cassell. 1999. Small Talk and Conversational Storytelling in Embodied Conversational Characters. In *Proceedings of AAAI Fall Symposium on Narrative Intelligence*, November 5-7, at Cape Cod, MA.

Brown, P., and S.C. Levinson. 1987. *Politeness: Some universals in language use.* New York: Cambridge University Press.

Cassell, Justine, and Timothy Bickmore. 2002. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and Adaptive Interfaces* 12:1-44.

Cassell, Justine, Yukiko Nakano, Timothy Bickmore, Candy Sidner, and Charles Rich. 2001. Non-Verbal Cues for Discourse Structure. In *Proceedings of 41st Annual Meeting of the Association of Computational Linguistics*, July 17-19, at Toulouse, France.

Cassell, Justine, Matthew Stone, and Hao Yan. 2000. Coordination and Context-Dependence in the Generation of Embodied Conversation. In *Proceedings of INLG 2000*, at Mitzpe Ramon, Israel.

Cassell, Justine, Andrea Tartaro, Vani Oza, Yolanda Rankin, and Candice Tse. forthcoming. Virtual Peers for Literacy Learning. *Educational Technology, Special Issue on Pedagogical Agents*.

Cassell, Justine, and Kristinn R. Thorisson. 1999. The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence* 13:519-538.

Cassell, Justine, and Hannes Vilhjálmsson. 1999. Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous. *Autonomous Agents and Multi-Agent Systems* 2:45-64.

Clark, Herbert H. 1992. *Arenas of Language Use*. Chicago, IL: University of Chicago Press.

Conati, Cristina, and Xiaoming Zhou. 2004. A Probabilistic Framework for Recognizing and Affecting Emotions. In *Proceedings of AAAI Spring Symposium on Architectures for Modeling Emotions*, March 22-24, at Stanford University, CA.

Descartes, René. 1953. Discours de la Méthode. In *Oeuvres et Lettres*. Paris: Librairie Gallimard, Bibliothéque de la Pléiade. Original edition, 1637.

Halliday, M. A. K. 1967. *Intonation and Grammar in British English*. The Hague: Mouton.

Hoffmann, E. T. A. 1844. The Sandman. In *Tales from the German, comprising specimens from the most celebrated authors*, edited by J. Oxenford and C. A. Feiling. New York: Harper & brothers. Original edition, 1817.

Keller, Evelyn Fox. 2003. Models, simulation, and "computer experiments". In *The philosophy of scientific experimentation*, edited by H. Radder. Pittsburgh Pa: University of Pittsburgh Press.

———. in press. Booting up Baby. In *The Sistine Gap: Essays on the History and Philosophy of Artificial Life*, edited by J. Riskin. Chicago: University of Chicago Press.

Kendon, Adam. 1972. Some Relationships between Body Motion and Speech. In *Studies in Dyadic Communication*, edited by A. W. Siegman and B. Pope. Elmsford, NY: Pergamon Press.

Lenoir, Timothy. 2000. All But War Is Simulation: The Military-Entertainment Complex. *Configurations* 8 (3):289-335.

Marsella, Stacy, David V. Pynadath, and Stephen Read, J. 2004. PsychSim: Agent-based modeling of social interactions and influence. In *Proceedings of The International Conference on Cognitive Modeling*, at Pittsburgh.

Nakano, Yukiko I., Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a Model of Face-to-Face Grounding. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, July 7-12, at Sapporo, Japan.

Nass, Clifford Ivar, and Scott Brave. 2005. *Wired for speech : how voice activates and advances the human-computer relationship*. Cambridge, Mass.: MIT Press.

Newell, Allen, and Herbert A. Simon. 1972. *Human problem solving*. Oxford, England: Prentice-Hall.

Poggi, Isabella, and Catherine Pelachaud. 2000. Performative Facial Expressions in Animated Faces. In *Embodied Conversational Agents*, edited by J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge: MIT Press.

Prevost, Scott Allan. 1996. Modeling Contrast in the Generation and Synthesis of Spoken Language. In *Proceedings of ICSLP '96*, at Philadelphia, PA.

Riskin, Jessica. 1999. Moving Anatomies. Paper read at History of Science Society 1999 Annual Meeting, November 3-7, at Pittsburgh, PA.

Rosis, Fiorella de, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio, and Berardina Nadja De Carolis. 2003. From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies. Special Issue on "Applications of Affective Computing in HCI"* 59:81-118.

Ryokai, Kimiko, Catherine Vaucelle, and Justine Cassell. 2003. Virtual Peers as Partners in Storytelling and Literacy Learning. *Journal of Computer Assisted Learning* 19 (2):195-208.

Schwartz, Daniel L., and Taylor Martin. 2003. Representations That Depend on the Environment: Interpretative, Predictive, and Praxis Perspectives on Learning. *Journal of the Learning Sciences* 12 (2):285-297.

Stronks, Bas, Anton Nijholt, Paul van der Vet, and Dirk Heylen. 2002. Designing for friendship: Becoming friends with your ECA. In *Proceedings of Embodied conversational agents - let's specify and evaluate them!*, at Bologna, Italy.

Suchman, Lucy. 1997. Do categories have politics? The language/action perspective reconsidered. In *Human values and the design of computer technology*, edited by B. Friedman. Cambridge; New York: Cambridge University Press.

———. 2003. Writing and Reading: A Response to Comments on Plans and Situated Actions. *Journal of the Learning Sciences* 12 (2):299-306.

Tartaro, Andrea, and Justine Cassell. in press. Using Virtual Peer Technology as an Intervention for Children with Autism. In *Towards Universal Usability: Designing Computer Interfaces for Diverse User Populations*, edited by J. Lazar. Chichester, UK: John Wiley and Sons.

Torres, Obed E., Justine Cassell, and Scott Prevost. 1997. Modeling Gaze Behavior as a Function of Discourse Structure. In *Proceedings of First International Workshop on Human-Computer Conversation*, July 14-16, at Bellagio, Italy.

Turkle, Sherry. 1995. *Life on the Screen : Identity in the Age of the Internet*. New York: Simon & Schuster.

Walker, Marilyn A., Janet E. Cahn, and Stephen J. Whittaker. 1997. Improvising Linguistic Style: Social and Affective Bases for Agent Personality. In *Proceedings of Autonomous Agents 97*, at Marina del Rey, CA.

Wang, N., W.L. Johnson, P. Rizzo, E. Shaw, and R.E. Mayer. 2005. Experimental Evaluation of Polite Interaction Tactics for Pedagogical Agents. In *Proceedings of International Conference on Intelligent User Interfaces*, at San Diego.

Yan, H. 2000. Paired Speech and Gesture Generation in Embodied Conversational Agents. Masters of Science, MIT Media Lab, MIT, Cambridge, MA.

[i] Acknowledgements: The research reported in this chapter would not have been possible without the hard work, persistence and insight of graduate students who were so generous with their time and knowledge that they quickly became my colleagues and teachers. Sincere thanks go to the members of Animated Conversation, Gesture and Narrative Language, and the ArticuLab. Thanks to Jessica Riskin for including me in this volume, and abiding gratitude to Ken Alder, Pablo Boczkowski, Sid Horton, Jessica Riskin, Dan Schwartz, Matthew Stone, and 2 reviewers for careful and perceptive comments that greatly improved the quality of the manuscript.