In press, AI Magazine

Embodied Conversational Agents: Representation and Intelligence in User Interface

Justine Cassell
MIT Media Lab
E15-315, 20 Ames Street, Cambridge MA 02139
justine@media.mit.edu

> *Suppose that sometimes he found it impossible to tell the difference between the real men and those which had only the shape of men, and had learned by experience that there were only two ways of telling them apart: first, that these automata never answered in word or sign, except by chance, to questions put to them; and second, that though their movements were often more regular and certain than those of the wisest men, yet in many things which they would have to do to imitate us, they failed more disastrously than the greatest fools.*
>
> *Descartes*

Introduction

How do we decide how to represent an intelligent system in its interface, and how do we decide how the interface represents information about the world and about its own workings to a user? This article addresses these questions by examining the interaction between representation and intelligence in user interfaces. The rubric "representation" covers at least three topics in this context: how a computational system is represented in its user interface, how the interface conveys its representations of information and of the world to human users, and how the system's internal representation affects the human user's interaction with that system. I will argue that each of these kinds of representations (of the system, of information and the world, of the interaction) is key to how users make the kind of attributions of intelligence that facilitate their interactions with intelligent systems. I will argue for representing a system as a human in those cases where social collaborative behavior is key, and I will argue for the system representing its knowledge to humans in multiple ways on multiple modalities. I will demonstrate my claims by discussing issues of representation and intelligence in an embodied conversational agent – an interface in which the system is represented as a person, in which information is conveyed to human users via multiple modalities such as voice and hand gestures, and in which the internal representation is modality-independent, and both propositional and non-propositional.

In order to start with a convenient counter claim, let me quote from a recent call for proposals on the topic of the disappearing computer: "in this vision, the technology providing these capabilities is unobtrusively merged with real world objects and places, so that in a sense it disappears into the background, taking on a role more similar to electricity - an invisible pervasive medium." In this vision of intelligent user interfaces there is no representation of the system, and no modalities by which information is conveyed to users. One interpretation of this vision (instantiated, for example, in ubiquitous computing) has been to make interactions *transparent* by embedding the interface to intelligent systems in old and familiar objects, which are therefore easy to use. A newer approach, however, has been to really dispense with objects altogether – to suffuse spaces with computation, therefore avoiding any point of interaction.

But, how many times have you seen hapless pedestrians stuck in front of an automatic “smart door”? They are unable to proceed because they don’t know where the sensors, or the door’s eyes, are located, and therefore they can’t make the door open by making the right size of movements in the right quadrant. As Harry Potter says, "never trust anything that can think for itself, if you can't see where it keeps its brain". Confusingly, projects involving "invisible computers" describe them as ways for people to interact with computation "as they interact with another person". As useful as is embedding computation in our environment, the notion must be tempered with knowledge of how humans actually do interact. We depend on forms of embodied interaction that offer us guidance in dealing with a complex world – interacting with invisibility does not fit one of the scripts. We need to locate intelligence, and this need poses problems for the invisible computer. The best example of located intelligence, of course, is the body. I’ll talk about how the body . . . *embodies* intelligence, both the usual knowledge about a particular domain, and a less-commonly discussed social interactional intelligence about conversational process, such as how to initiate, take-turns and interrupt in a conversation. And I’ll demonstrate how intelligent user interfaces can take advantage of embodied intelligence to facilitate human-machine interaction with a series of what we refer to as *Embodied Conversational Agent* (ECA) systems.

An example of a person interacting with another person may serve to explain how humans actually do interact in their natural context, and demonstrate some of the potential problems with interacting with invisibility. Figure 1 shows a young woman describing the layout of a house to a young man. Her eyes focus diagonally up and away as she plans her first utterance, and then turn to her listener as she describes a complicated set up with her words and with her hands. When the speaker says that the house is “surrounded by a porch all around”, her hands demonstrate that the porch actually covers three sides of the house. The eye gaze towards the listener (depicted in the frozen frame in Figure 1) elicits a feedback nod from him, during which the speaker is quiet (++ indicates silence). Once the speaker receives the listener’s reaction, ensuring that speaker and listener share a common ground or understanding of what has already been described, she continues. She looks up as she plans her next utterance and repeats the gesture performance as she completes the description of the porch. The timing of the eye gaze, head movements, and hand gestures are tightly synchronized to speech, as marked by square brackets in the transcript. They are also tightly synchronized to the listener’s behavior, as demonstrated by the feedback-eliciting gaze. The basic point is that people communicate *with* and *to* other people, and not in a vacuum. Eyes gaze *at* other people, and focus other people’s attention on shared targets, hands gesture *between* people, faces express *to* other people. These behaviors are the external manifestations of social intelligence and trustworthiness (Cassell and Bickmore 2000), as well as a localization of the conversational processes of grounding information, and a representation of information in their own right. Thus, if our goal is to reproduce how people communicate in natural contexts, we must also reproduce a way to *localize the interaction* and to *represent the system’s intelligence* in space, to make the agency and intelligence of the participants visible by their actions and their reactions to communication.

Figure 1: Describing a house

The first floor ‘s [+++]
- Gaze diagonally up
- Hands make loose beat gesture

it’s like a [box and it’s surrounded by a porch]
- Gaze at listener
- Hands describe three sides of a box, twice

[++] all around + um + [so]
- Tilt head gaze diagonally up
- Hands repeat previous box gesture

———- SIDEBAR: The history of Embodied Interfaces: Automata ——————-

Wherefore are they endowed with organs so like to those of ourselves? Wherefore have they eyes, ears, nostrils, and a brain? It may be answered, that they may regulate the movements of the automata, by the different impressions which they receive from the exterior objects.
D’Alembert

Attempts to model the body, and bodily interfaces, as well as attempts to make pretty bodies as entertainment systems, have been around for a very long time. In repudiation of Descartes’ strict separation between the stuff that humans are made of, and the thoughts they think, the organicist automaton makers of the 18th century asked whether one could design a machine that could talk, write, interact, play chess, and so forth, in the way people do. They intended to find out in this way what these activities consisted of when human beings perform them; and how they differed, if at all, when machines perform them (Riskin 1999). For these reasons, designers in the organicist tradition tried to make their machines as much as possible like the organic subjects and processes they were imitating. Some organicist machines were strikingly life-like representations of human processes, and contributed significantly to knowledge about human functioning – for example, Droz’s writing boy (Figure 2) whose pen moves across the page just as real writers’ pens move.

Some organicist machines also acted as interfaces between humans and the world – for example, Von Kemplen’s Speaking Machine (1791). Von Kempelen's machine (Figure 3) was the first that allowed users to produce not only some speech sounds, but also whole words and short sentences. According to von Kempelen, it was possible to become a proficient user of the machine within three weeks, and to then be able to produce strings of Latin, French, Italian or German.

In contrast to these serious and scientific attempts to build embodiments and interfaces based on human function, were 19th century automata that were meant to entertain, regardless of how human-like their actions might be. An example of such a pretty 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.

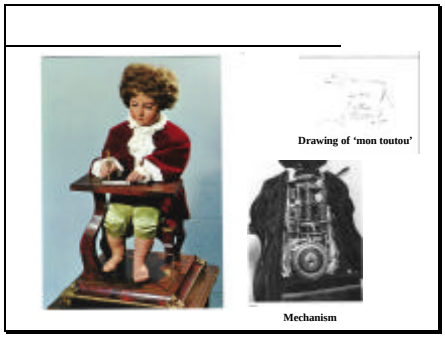

Figure 2: Droz's Automaton "the Writing Boy"

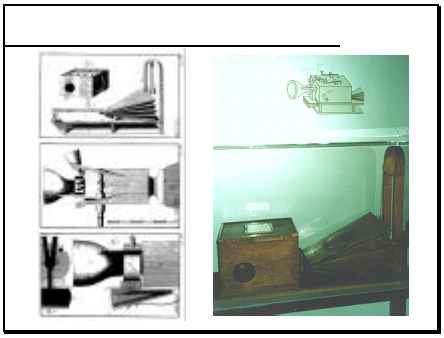

Figure 3: Von Kemplen's Speaking Machine

Automaton-makers were burned at the stake in the middle ages. And, today in the interface community, we suspect some traditional HCI researchers would be happy to do the same thing! As one of the most prominent critics has put it,

> For those who built stone idols or voodoo dolls or the golem or Frankenstein, it's long been a dream.... But no mature technology resembles [animal] form. Automobiles don't run with legs, and planes don't flap their wings… [Anthropomorphized agents] are things that think for people who don't.

Nevertheless, technology will most likely always be used to model humans, in order to better understand how humans function, and in order to leverage human understanding of the world by building that understanding into an interface that we learn to use from so early on in life.

-————— END OF SIDEBAR—————

Human Representation and Intelligence in Face-to-Face Conversation

The speaker in Figure 1 knows something about the world that she is trying to convey to her listener, and she knows something about social conventions that is influencing how she goes about her task. We call these propositional and interactional functions, or skills[1]. As described, both are carried by a number of behaviors in a number of different modalities: the voice, the hands, eye gaze, head movement.

We know that language is a representational medium (the *ur* representational medium), but are these other modalities anything other than fluff, pretty movements to occupy the body while the mouth is working? Eyes are not good representational tools (they can't *describe*) but they can certainly annotate (a discrete roll of the eyes while mentioning the election), focus the attention of one's interlocutor (as when one looks at one's own hands during a particularly complex gesture), and index appropriate social behavior (as above, when the speaker requests feedback by letting her gaze rest on her listener momentarily). Hands are excellent representational tools, better than speech even at representing the simultaneity of two events, or the respective spatial locations of two objects ("so Lucy stood **there** and Betsy stood **there**") and at disambiguating anaphoric reference ("and then **she** showed **her** how to move her feet"). In different contexts, gesture takes different forms: the more unfamiliar or surprising a speaker thinks a concept might

[1] Or even task and social *intelligences*, based on the fact that some people are more skilled than others in one or more of the two domains, and that each can be selectively knocked out by brain injury or genetic disorders.

be, the more representational the gesture accompanying mention of that concept (Cassell, Stone et al. 2000). Thus, when talking to the human-computer interaction community, a speaker might clasp her two hands together in front of her while saying the phrase "shared plans". In a computational linguistics conference, a nod of the head in the direction of Barbara Grosz and Candy Sidner sitting in the audience would suffice. Both hands and head are skilled at taking up the slack of communication: a nod to acquiesce when one's mouth is too full to say "yes", a point towards one's full mouth to explain that one cannot speak. The body is the master of alternate and multiple representations, according to the needs and style of speaker and listener. Embodiment, therefore, would seem to fit the description of the ultimate interface, which "ultimately will include the ability to both retrieve and generate alternate representations of information according to the needs and personal styles of users" (Laurel 1990:362).

So these behaviors can both convey information and regulate communication in face-to-face conversation, but do they communicate – that is, does the listener pay any attention? In fact, yes, listeners depend on such embodied behaviors in face-to-face conversation. For example, they use the hand gesture they have seen in these situations to form a mental representation of the propositional content conveyed (Cassell, McNeill et al. 1999), and they use the eye gaze to constrain when they make their own bids for the floor (Duncan 1974). We also know, however, that listeners are unable to remember what hand gestures they saw (Krauss, Morrel-Samuels et al. 1991), and when they redescribe a monologue, they are likely to transpose the modality in which the information was conveyed. And, although teachers have been shown to use their pupils' gestures to judge the accuracy of the children's underlying understanding of mathematical concepts, they are unaware that they are so doing (Goldin-Meadow, Alibali et al. 1993). So, just as with speech, the meanings underlying embodied interaction are extracted, but the behaviors themselves are not retained. However, when these embodied behaviors are omitted in face-to-face interaction between a user and an embodied system, users repeat themselves more, and judge the system's use of language, and understanding of language to be worse (Cassell and Thorisson 1999). And, when speech is ambiguous between humans (Thompson and Massaro 1986) or in a speech situation with some noise (Rogers 1978), listeners rely more on gestural cues (and, the higher the noise-to-signal ratio, the more facilitation by gesture). Thus, although the behaviors are not consciously retained, they are key to the interaction among humans, and between humans and machines. Note that the evidence presented thus far argues for different depictions on different modalities, but one underlying modality-free common conceptual source that gives rise to the different instantiations, wherein each modality is called on to do what it does best. This semantic and pragmatic sharing of work recalls the interaction of words and graphics in an early kind of intelligent user interface: automatic generation of multimodal presentations (Wahlster, Andre et al. 1991), (Feiner and McKeown 1991), and recalls the separation of description and mechanism in Rosenschein & Kaebling's classic AI paper (1986).

From an ontological perspective the importance of multiple representations in multiple modalities is not surprising – it has been argued that gestures are our first representational activities, arising from early sensorimotor schemata (Piaget 1952), and continuing to replace unknown words in children's communication (Bates, Bretherton et al. 1983), and certainly eye gaze and head movement regulate proto-conversations between caregivers and infants, before infants can even produce a semblance of language (Trevarthen 1986.). Even in adults nonverbal behaviors do not fade. About three-quarters of all clauses in narrative discourse are accompanied by gestures of one kind or another, regardless of cultural background (McNeill 1992).

The Conversational Model

So humans engage in complex representational activity involving speech and hand gesture, and they regulate that activity through social conversational protocols that include speech and eye gaze and head movement and hand gesture. In this context, we can view the human in our example above as providing structure for her interlocutor that helps him navigate a complex description of the world. Her entire embodied performance provides cues as to the shape of objects for which there is no adequate description in English, cues about who has the floor, whether the two participants have reached common knowledge, and when the speaker's internal conceptual representation is in the process of being translated into words. Such a performance is helpful to the listener in understanding what is being said, and integrating it into an ongoing discourse. It is also helpful in that it indicates that the speaker is a kind of representational device that the listener is familiar with, it allows the listener to apply a theory of mind (Astington, Harris et al. 1988) and by doing so, to map the speaker's behaviors onto richer underlying representations, functions and conventions – to attribute intelligence to the other.

In building an embodied conversational agent, we wish both to help users steer their way through complex descriptions of the world, and we wish to prod them into automatically applying such a theory of mind as will allow them to not have to spend their time constructing awkward new theories of the machine's intelligence on the fly. Thus, our model of conversational behavior must be able to predict exactly this kind of conversational behaviors and actions; that is, it must provide a way of *realizing* this set of conversational surface behaviors in a principled way.

Our model, however rich, will not be able to predict all of the behaviors displayed in human-human conversation nor describe all of the functions that give rise to these surface behaviors, nor would we wish it to. The model is not mimicking what people look like, but adopting those aspects of the human interface that provide structure for our interpretations of meaning, and for the process of interpreting meaning. Our goal is to target those behaviors that regulate and facilitate the process of interaction, and represent information efficiently and effectively, all the while evoking a sense of *another intelligence.* Of course, however poor our model is, it will give rise to attributions that we have not planned: side-effects of cultural stereotypes of gender, race, and age that are evoked by the pitch of a voice, the tilt of a head, the form of a question (Reeves and Nass 1996). Steering our way through this Scylla and Charybidis of personification is helped by frequent evaluations of the system, in which users reveal the attributions – desired and not – that the system's behavior provokes.

These principles lead to a conversational model with several key properties: the system-internal representation of the world and of information must be modality-free, but able to be conveyed via any one of several modalities; the functions of the system must be modality-free, but able to be realized in any one of a number of different surface behaviors in a number of different modalities; the representations of conversation cannot be all symbolic, as cultural and social conventions may not be able to be captured in logical form; co-occurrences of surface-level behaviors carry meaning, over that carried by each of the constituent behaviors. In sum, we might describe such a model as a 'multiplicity of representations'. We capture these properties and insights about human conversation in the **FMBT** (pronounced *fembot*) model:

F. Division between Propositional and Interactional **F**unctions

Contributions to the conversation are divided into *propositional functions* and *interactional functions*. The propositional function corresponds to the content of the conversation. This includes meaningful speech as well as hand gestures (gestures that indicate size in the utterance “it was *this* big” or that represent fingers walking in the utterance “it took me 20 minutes to get here”). The interactional function consists of cues that regulate 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 the turn) as well as regulatory speech ("huh?", "do go on"). In short, the interactional discourse functions are responsible for creating and maintaining an open channel of communication between the participants, while propositional functions shape the actual content. Both functions may be fulfilled by the use of a number of available communication modalities.

M. Modality

Both verbal and nonverbal modalities are responsible for carrying out the interactional and propositional functions. It is not the case that the body behaviors are redundant. The use of several different modalities of communication - such as hand gestures, facial displays, eye gaze, and so forth - is what allows us to pursue multiple goals in parallel, some of a propositional nature and some of an interactional nature. For example, a speaker can raise her pitch towards the end of the sentence while raising the eyebrows to elicit feedback in the form of a head nod from the listener, all without interrupting the production of propositional content. It is important to realize that even though speech is prominent in conveying content in face-to-face conversation, spontaneous gesture is also integral to conveying propositional content. In fact 50% of gestures add non-redundant information to the common ground of the conversation (Cassell, Stone et al. 2000). For interactional communicative goals, the modality chosen may be more a function of what modality is free at a given point in the conversation - for example, is the head currently engaged in attending to the task, or is it free to give a feedback nod?

B. Behaviors are not functions

The same communicative function does not always map onto the same observed behavior. For instance, the interactional function of giving feedback could either be realized as a head nod or a short "mhm". The converse is also true - the same behavior does not always serve the same function. For example, a head nod could be feedback or equally well a salutation or emphasis on a word. The particular set of surface behaviors exhibited may differ from person to person and from conversation to conversation (not to mention from culture to culture). Therefore to successfully build a model of how conversation works, one cannot refer to these behaviors, or surface features alone. Instead, the emphasis has to be on identifying the high level structural elements or functions that make up a conversation. It is the understanding of these functions and how they work together to form a successful interaction that allows us to interpret the behaviors in context.

Communicative Functions	**Communicative Behavior**

Initiation and termination:	
React to new person	Short glance at other
Break away from conversation	Glance around
Farewell	Look at other, head nod, wave
Turn-Taking	
Give Turn	Look, raise eyebrows (followed by silence)
Want Turn	Raise hands into gesture space
Take Turn	Glance away, start talking
Feedback	
Request Feedback	Look at other, raise eyebrows
Give Feedback	Look at other, nod head

Table 1. Some examples of conversational functions and their behavior realization (taken from (Cassell and Vilhjálmsson 1999)

T. Time

Timing is a key property of human conversation, both within one person's conversational contributions, and between participants. Within one person's contribution, the meaning of a nod is determined by where it occurs in an utterance, to the 200 millisecond scale. For example, consider the difference between "John [escaped]" (even though we thought it was impossible) and "[John] escaped" (but Bill did not). Between participants, a listener nod at precisely the moment a speaker requests feedback (usually by way of a rise in intonation and the flashing of eyebrows) is displaying understanding while a delayed head nod may signify confusion. The rapidity with which behaviors such as head nods achieve their goals emphasizes the range of time scales involved in conversation. While we have to be able to interpret full utterances to produce meaningful responses, we must also be sensitive to instantaneous feedback that may modify our interpretation and production as we go.

Although the FMBT model shares with Rodney Brooks and his colleagues a reliance on social interaction, and a distinction between surface level behaviors and underlying (deep structure) functions (Brooks, Brezeal et al. 1998), in our model these key properties do not displace the need for a explicit internal representation. Having a physical body, and experiencing the world directly through the influence of the world on that body, does not obviate the need for a model of the world. In Brooks' work, the autonomous systems he builds exploit features of the world and of humans in order to learn – the systems can go without a representation of their own so long as the world and humans manifest structure. In our work, humans exploit features of the interface to autonomous systems in order to achieve their goals – the interface must then present structure that the human can use. Intelligent user interfaces must provide representation[2]; intelligent creatures can rely on the representations provided by others.

[2] Cf. "A basic principle underlying multimedia systems is that the various constituents of a multimodal communication should be generated on the fly from a common representation of what is to be conveyed without

REA: Implementing a FMBT Embodied Conversational Agent

Thus far I've talked about some of the properties of embodied human-human conversation that are essential for conveying information, regulating the course of the interaction, and giving one's interlocutor the sense that one is a familiar kind of representational creature. I've captured these key properties in the FMBT intelligent interface model, and distinguished the model from 'intelligence without representation' models of autonomous creatures. In this section, I give the details of how an embodied conversational agent can be implemented based on the model. To demonstrate, I turn to Rea, an embodied conversational agent whose verbal and non-verbal behaviors are generated from underlying conversational functions and representations of the world and information. Rea is the most extensive embodied conversational agent that we have built on the basis of the FMBT model, which is why she is serving as example. However, the architecture described here is independent of the Rea implementation, and has been used for a number of other embodied conversational agents (described more briefly in the last section below). To start with an overview:

Figure 4: REA welcoming a user to her virtual realty office

- Rea has a human-like body (shown in Figure 4), and uses her body in human-like ways during the conversation. That is, she uses eye gaze, body posture, hand gestures, and facial displays to contribute to the conversation, and to organize and regulate the conversation. She also understands (some aspects of the use of) these same modalities when employed by her human interlocutor.
- The architecture allows for multiple threads of interaction to be handled, thus allowing Rea to watch for feedback and turn requests, while the human user can send these at any time through various modalities. The architecture is flexible enough to track these different threads of communication in ways appropriate to each thread. Because different threads have different response time requirements, the architecture allows different processes to concentrate on activities at different time scales.

---

using any preplanned text or images; that is, the principle is 'no generation without representation'." (Maybury and Wahlster 1998)

- Dealing with propositional information requires building a model of the user's needs and knowledge. Thus the architecture includes both a static knowledge base that deals with the domain (here, real estate) and a dynamic discourse knowledge base (dealing with what has already been said). To generate propositional information the system plans how to present multi-sentence multimodal output and manage the order of presentation of interdependent facts. To understand interactional information, on the other hand, the system builds a model of the current state of the conversation with respect to conversational process (who is the current speaker and who is the listener, has the listener understood the speaker's contribution, and so on).
- The core modules of the system operate exclusively on functions (rather than sentences or behaviors, for example), while other modules at the edges of the system translate input into functions, and functions into outputs. This also produces a symmetric architecture where the same functions and modalities are present in both input and output. Such models have been described for other conversational systems: for example, by Brennan and Hulteen (1995) We extend this previous research by developing a conversational model that relies on the function of non-verbal behaviors as well as speech, and that makes explicit the interactional and propositional contribution of these conversational behaviors.

Architecture

Figure 5 shows the modules of the Rea architecture. Three main points translate the FMBT model for Embodied Conversational Agents:

- Input is accepted from as many modalities as there are input devices. However the different modalities are integrated into a single conceptual representation that is passed from module to module.
- This conceptual representation frame has slots for interactional and propositional information so that the regulatory and content-oriented contribution of every conversational act can be maintained throughout the system.

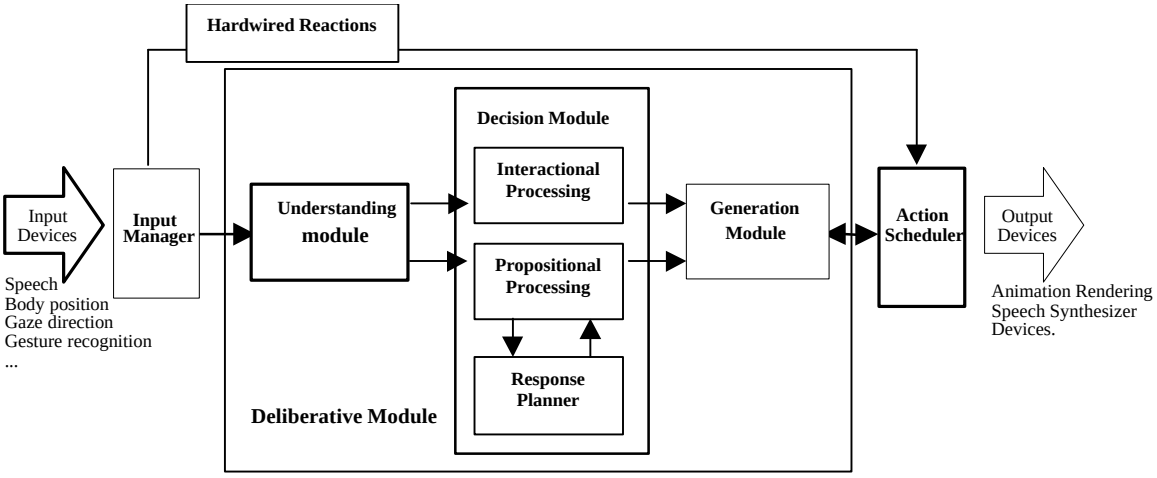

Figure 5: REA Architecture (Co-developed with the Fuji-Xerox Palo Alto Laboratory)

- The categorization of behaviors in terms of their conversational functions is mirrored by the organization of the architecture which centralizes decisions made in terms of functions (the

understanding, decision, and generation modules), and moves to the periphery decisions made in terms of behaviors (the input manager and action scheduler).

The Input Manager collects input from all modalities and decides whether the data requires instant reaction or deliberate discourse processing. Hardwired Reaction handles rapid (under 200 milliseconds) reaction to stimuli, such as the appearance of the user. These stimuli can then directly affect the agent's behavior without much delay. This means that, for example, the agent's gaze can keep up with tracking the user's movement, without first processing the meaning of the user's appearance. The Deliberative Discourse Processing module handles all input that requires a discourse model for proper interpretation. This includes many of the interactional behaviors as well as all propositional behaviors. Lastly the Action Scheduler is responsible for scheduling motor events to be sent to the animated figure representing the agent. A crucial function of the scheduler is to synchronize actions across modalities, so that, for example, gesture stroke and pitch peak in speech co-occur within milliseconds of each other. The modules communicate with each other using KQML, a speech-act based inter-agent communication protocol, which serves to make the system modular and extensible.

Implementation

The system currently consists of a large back projection screen on which Rea is displayed and in front of which the user stands. Two cameras mounted on top of the projection screen track the user's head and hand positions in space. Users wear a microphone for capturing speech input. A single SGI Octane computer runs the conversation engine (originally written in C++ and CLIPS, currently moving to Java), while several other computers manage the speech recognition (until recently IBM Via Voice; currently moving to SUMMIT) and generation (previously Microsoft Whisper; currently BT Festival), image processing (STIVE (Azarbayejani, Wren et al. 1996) and VGUS (Campbell forthcoming)), and graphics (written in OpenInventor).

In the implementation of Rea we have attended to both propositional and interactional components of the conversational model, and all of the modalities at Rea's disposal (currently speech with intonation, hand gesture, eye gaze, head movement, body posture) are available to express these functions. Rea's current repertoire of interactional functions includes:

- Acknowledgment of user's presence by turning to face the user;
- Feedback - Rea gives feedback in several modalities: she may nod her head or emit a paraverbal (e.g. "mmhmm") or a short statement such as "okay" in response to short pauses in the user's speech; she raises her eyebrows to indicate partial understanding of a phrase or sentence.
- Turn-taking– Rea tracks who has the speaking turn, and only speaks when she holds the turn. Currently Rea always allows verbal interruption, and yields the turn as soon as the user begins to speak. If the user gestures she will interpret this as an expression of a desire to speak, and therefore halt her remarks at the nearest sentence boundary. Finally, at the end of her speaking turn she turns to face the user.

These conversational functions are realized as conversational behaviors. For turn taking, for example, the specifics are as follows: Rea generates speech, gesture and facial expressions based on the current conversational state and the conversational function she is trying to convey. 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. When conversational turn-taking begins, she orients her body to face the user at a 45 degree angle.

When the user is speaking and Rea wants the turn she looks at the user. When Rea is finished speaking and ready to give the turn back to the user she looks at the user, drops her hands out of gesture space and raises her eyebrows in expectation. Table 2 summarizes Rea's current interactional output behaviors.

State	Output Function	Behaviors
User Present	Open interaction	Look at user. Smile. Toss head.
	Attend	Face user.
	End of interaction	Turn away.
	Greet	Wave. Say "hello" .
Rea Speaking	Give turn	Relax hands. Look at user. Raise eyebrows
	Signoff	Wave. Say "bye"
User Speaking	Give feedback	Nod head, paraverbal ("hmm")
	Want turn.	Look at user. Raise hands.
	Take turn.	Look at user. Raise hands to begin gesturing. Speak.

Table 2. Output Functions

In terms of the propositional component, Rea's speech and gesture output is generated in real-time, and words and gesture are treated on a par, so that a gesture may be just as likely to be chosen to convey Rea's meaning as a word. The descriptions of the houses that she shows, along with the gestures that she uses to describe those houses are generated using the SPUD natural language generation engine, modified so as to also generate natural gesture (Cassell, Stone et al. 2000). New propositional information is conveyed using *iconic* gestures (for concepts with concrete existence), *metaphoric* gestures (for concepts which do not have concrete existence and thus must make use of spatial metaphors for depiction), or *deictic* gestures (for indicating or emphasizing an object in Rea's virtual world, such as features of homes she is showing to the user). These gestures are either wholly redundant with or complementary to the speech channel,

based on semantic and pragmatic constraints. *Beats* are used to indicate points of emphasis in the speech channel without conveying additional meaning.

When Rea produces an utterance, then, she first determines several pieces of pragmatic and semantic information, including:

- semantics - speech act description of Rea's communicative intent (e.g., OFFER the user a particular property, DESCRIBE a room, etc.)
- information structure - which entities are new vs. previously-mentioned
- focus - which entity (if any) is currently in focus
- mutually-observable - which entities in the virtual world are visible to both Rea and the user

This information is then passed to the SPUD unified text and gesture generation module which generates Rea's natural language responses. This module distributes the information to be conveyed to the user across the voice and gesture channels based on the semantic and pragmatic criteria described above, and timed so that gestures coincide with the new material in the utterance. If a new entity is in focus and it is mutually-observable, then a deictic is used. Otherwise, Rea determines if the semantic content can be mapped into an iconic or metaphoric gesture (using heuristics derived from studies of the gestures humans produce in describing real estate (Yan 2000) to determine whether the gestures should be complementary or redundant). For example, Rea may make a *walking gesture* (extending her index and second finger with the tips downward, as if they are legs, and wiggling the fingers back and forth) as she says “The house is five minutes from MIT”. In this case, the gesture carries complementary information – that the house is five minutes on foot, rather than five minutes by car. Or, Rea may make a sweeping “sun-rising” gesture with both arms above her head, as she says “the living room is really luminous”. In this case, the gesture is redundant to the notion of sunniness conveyed by speech.

Rea is also able to detect these same classes of gestures made by the user and combine this information with speech input to interpret propositional content, and make decisions about appropriate responses. Here, once again, our reliance on conversational function, and distinction between function and surface behavior, allows us to bypass questions of gestural *form* (a very difficult vision problem indeed) and concentrate on how the user is employing a given gesture in the current conversational context. The gesture classification module uses a set of Hidden Markov Model (HMM) recognizers to classify gestures into the categories of rest (no gesture), beat, deictic, butterworth (searching for a word), or illustrative (iconic or metaphoric). The HMMs that classify into these categories were trained in an offline process from a set of 670 gestures obtained by tracking naive subjects with STIVE as they engaged in real-estate oriented conversations, and then hand-segmenting and classifying the subjects' conversational gestures (Campbell forthcoming).

So far, REA is only capable of integrating the beat and deictic categories into the decision module. The beat category is used to interpret user emphasis with respect to the speech channel, allowing Rea to know which part of an user utterance to respond to (if the user says "I like the blue tiles" with a beat gesture on the word 'blue,' then Rea responds by saying "Blue is my favorite color." If the user says "I like the blue tiles" with a beat gesture on the word 'tiles,' then Rea responds by saying "I love tiles."). Rea can also pick up when users point to objects in REA's world, and can associate the deictic gesture and the graphical object pointed at with the word that co-occurs with the deictic, thus enabling Rea to resolve a wider range of referring expressions ("that house", "the wall", etc.).

By modeling behavioral categories as discourse functions we have developed a natural and principled way of combining multiple modalities, in both input and output. Thus when REA decides to give feedback, for example, she can choose any of several modalities based on what is appropriate and available at the moment. When she conveys her knowledge of the world, all of her behaviors are marshaled towards giving a well-rounded description of what she knows.

An Example Interaction

The following records an actual interaction between a 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 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. . .

Mike interrupts Rea who stops speaking immediately.

MIKE> Wait. Tell me, where is the master bedroom?

REA> I'm sorry, I didn't catch that. What did you ask me?

MIKE> Where is the master bedroom?

REA> It's upstairs.

Rea points up
MIKE> Where is the master bathroom?

REA> It's next to the bedroom.

Rea brings her hands together to indicate the relationship between the bedroom and the bathroom.
And the house tour continues...

Related Work

Human-like embodied interfaces have become popular as the front end to commercial systems, so it doesn't seem like what I've described here is altogether novel. However, while these interfaces look like bodies, few of them display behaviors or manifest the types of functions of bodies in conversation that I've argued for[3]. Well-designed interfaces have *affordances* or visual clues to their operation. Bodies provide very strong visual clues about the protocols that they engage in, and these protocols must be integrated into the very heart of a system, and must give rise to appropriate surface level behaviors in the interface, for the embodied interface to be successful. In contradistinction to this methodology, many of the human-like interfaces on the market simply consist of an animated character slapped onto a system, capable of portraying a series of affective or "communicative" poses, without much attention paid to how humans actually convey their knowledge of the world and of human interaction to their interlocutors. Two examples of less-than-perfect embodied interfaces that come to mind are the Microsoft Office Assistant (the dreaded "Paper Clip") and Ananova. The Paper Clip (or the more anthropomorphic version Einstein) interrupts in an impolite and socially inappropriate manner, and when not actually itself typing, manifests its profound boredom in the user's work by engaging in conversationally irrelevant behaviors (as if one's interlocutor checked his watch while one was speaking and then snapped to attention when it was his turn to talk). Interestingly, interruption itself is not the problem – participants in conversations interrupt one another all the time. But interruption must be motivated by the demands of the conversation – requests for further information, excitement at what is being said – and must follow the protocol of conversation (for example, raise the hands into gesture space, clear the throat, extend the feedback noise for longer than usual, as ways of requesting the floor). Ananova is advertised as a way to personalize web users' interaction with information, but is not in fact capable of interaction (nor, currently, of personalization). This not uncommon confusion of "personalization" and "graphical representation of person" puts a body on the interface for looks and "personality" as opposed to arising from the functions of the body. Ananova sways slightly and winks or sneers occasionally, but does not pause or request feedback, check her viewer's response to what she is saying, or in any other way attend to her viewers. Nor does she use any modalities other than voice to convey content. The pages advertised as showing "technical drawings" tell us "Then one of [my creators] had the bright idea: why not unleash my full potential by giving me a human face and full-rounded personality so that I could better interact with people as technology develops." As far as one can tell, Ananova's behaviors are hand-scripted as annotations to text (for example, it's hard to imagine the set of underlying rules for conversation, and representation for information that would make her lip curls slightly in a sneer when she says "I've been locked in a room for 12 months with only geeks and techies for company"). And even when the day comes where she is able to deliver all and only the news a

[3] Surprisingly, not much has changed since the abstract of a CHI panel about anthropomorphism in 1992 declared "Recently there has been a discernible increase in the gratuitous use of the human figure with poorly lipsynched talking heads or systems that fool the user into thinking that the system is intelligent."(Don, Brennan et al. 1992). Actually, lipsynching has improved.

particular web viewer requests, her interaction with that web user will not resemble conversational interaction, nor will her embodiment make her appear any more intelligent an interface.

Figure 6: Ananova

Such systems represent an enormous missed opportunity. As I've argued, used appropriately, an embodied interface can provide not only something pretty and entertaining to look at, but also enables the use of certain communication protocols in face-to-face conversation that facilitate user interaction, and provide for a more rich and robust channel of communication than is afforded by any other mediated channel available today. The use of gaze, gesture, intonation, and body posture play an essential role in the proper execution of many conversational behaviors – such as conversation initiation and termination, turn-taking and interruption handling, and feedback and error correction – and these kinds of behaviors enable the exchange of multiple levels of representation of information in real time. In essence, I am saying that for the face-to-face metaphor of human-computer interaction to be successfully employed in intelligent user interfaces, the metaphor must be followed both in the surface level behaviors that the interface manifests, and the functions by which the system operates. Although *believability* is often named as the primary purpose behind embodying the interface (see, for example, (Elliott and Brzezinski 1998)), functionality would appear to be more effective.

Other researchers have created embodied interfaces that do rely on the function of conversation. Takeuchi and Nagao (Takeuchi and Nagao 1993), implemented a 'talking head' not too dissimilar from Ananova (modulo green hair) that understood human input and generated speech in conjunction with conversational facial displays of very similar types to those described in this article. Unfortunately, their system had no notion of conversational state, nor was it able to precisely time behaviors with respect to one another. In the field of Embodied Conversational Agents, some of the other major ECA systems developed to date are Steve (Rickel and Johnson 1998), which uses nonverbal signals to orient attention, give feedback, and manage turn-taking; the DFKI Persona (Andre 1997), which makes a distinction between creation of discourse content and the communication of that content (*acquisition* and *presentation* acts); and the pedagogical agents developed by Lester, et al, (Lester and Stone 1997), in which deictic gestures are more likely to occur in contexts of referential ambiguity. In all three systems, however, the association between verbal and nonverbal behaviors is additive – that is, the information conveyed by hand gestures, for example, is always redundant with the information conveyed by speech. The affordances of the body are not exploited for the kinds of tasks that it performs better than speech. There are also a growing number of commercial ECAs based on how bodies function in conversation or presentation, such as those developed by Extempo, Headpedal, and Artificial Life. These systems vary greatly in their linguistic capabilities, input modalities (most are

mouse/text/speech input only), and task domains, but all share the common feature that they attempt to engage the user in natural, full-bodied (or half-bodied) conversation.

More generally, linking intelligence to embodiment in the way done here is a position most often attributed to robotics, in particular the work of Rodney Brooks and his colleagues, though, as described above, Brooks explicitly denies the link between embodiment and representation, and intelligence and representation. I believe that one outcome of his position, however, is that his robots make better infants than parents. That is, they are better capable of learning from the structure provided by the people around them, than they are at providing structure for a user attempting to acquire information from an intelligent system.

Future Work

Our work on the Rea system continues along a number of parallel fronts. In order to explore whether verbal and nonverbal modalities are equally suitable for propositional and interactional functions, we have expanded the range of social conversational protocols Rea is able to engage in. Conversational regulation occurs not just at the level of eye gaze, but also at the level of entire utterances. For example, speakers often use small talk to set their listeners at ease. Realtors demonstrate a particularly advanced kind of social conversational protocols: they learn your name easily, remember what you've told them about your kids and your job, and they always have a story that's relevant to the interaction at hand, but that doesn't lead the conversation away from what you wanted to talk about. These skills allow them to appear trustworthy to their clients, and therefore are essential to their ultimate task-oriented goals. We have implemented a discourse planner capable of generating small talk in appropriate contexts, and carried out an evaluation comparing "small talk Rea" to "task-talk-only Rea". For high engagement users (extroverts and conversational initiators) these small talk behaviors lead to increased trust in the system's capabilities; for low engagement users, small talk decreases their perception of the system's skills (Cassell and Bickmore submitted for publication). We are currently reflecting on how to model these user characteristics such that Rea can anticipate which conversational style would be most effective with a particular user. We are also investigating the role of discourse-sensitive intonation as another modality for Rea to exploit.

In addition to REA, we have developed a number of other ECAs. In three different application domains (children's learning, seniors' reminiscence therapy, information kiosks) we have explored how to allow a user to share an actual physical space with an embodied conversational agent, so that speech and gesture can be joined by *action* as a representational strategy. In the Sam project (Cassell, Ananny et al. 2000), an embodied peer companion conversational agent

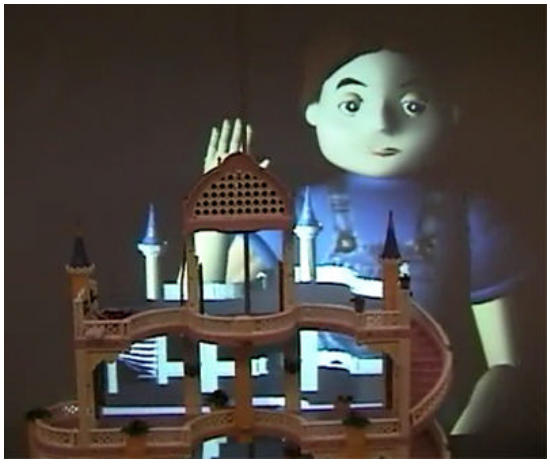

Figure 7:Sam, Embodied Conversational Peer

encourages young children to engage in storytelling by taking turns playing with a figurine in a toy castle which exists half in the child's physical reality, and half in Sam's virtual reality. Sam is designed to support the kind of oral narrative skills that are precursors to later literacy. In the GrandChair project, an ECA who appears to be a young child in a rocking chair sits across from an old person in an actual rocking chair to listen to grandparents' family stories (Smith 2000). The stories are videotaped so that they can be watched by future generations. In the MACK (Media Lab Autonomous Conversational Kiosk) project, an ECA is projected behind a table in the lobby of a research laboratory (Figure 8). Visitors can place their map of the building on the table, and MACK will point out features of the lab on the map, recognize user gestures to the map, and give descriptions of research projects, as well as directions on how to find those projects. Our ongoing work on MACK examines how users divide their focus of attention between the physical map, and MACK himself, in order to better allocate information to MACK's hands and face, and to the map.

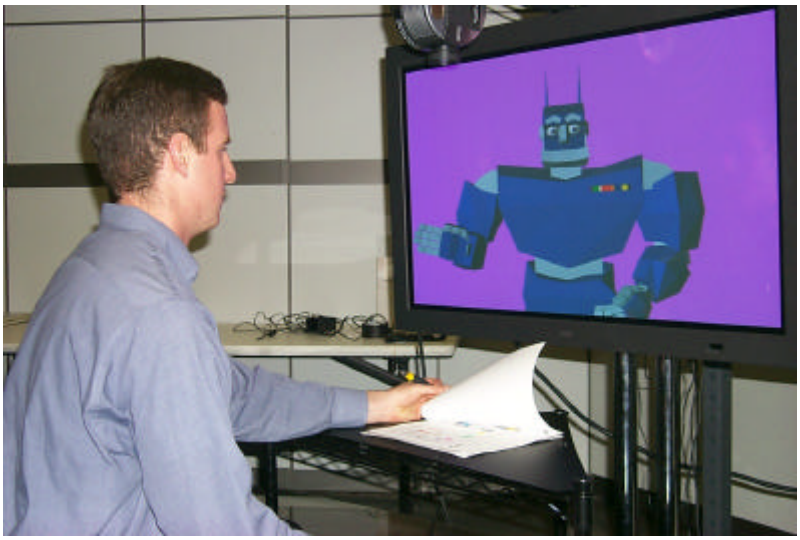

Figure 8: MACK: Media Lab Autonomous Conversational Kiosk

Finally, in the BodyChat system (Cassell and Vilhjálmsson 1999) we research how to derive the kinds of nonverbal interactional and propositional behaviors discussed here from people's typed text. In this context we have developed semi-autonomous ECA avatars to represent users in interactions with other users in graphical chat systems. In these systems users control the content of what their avatar says while much of the nonverbal conversational behavior displayed by the avatar is automatically generated based on the typed text, the conversational context, and the graphical space.

Conclusions

What's the point of discussing conversational smarts in an AI magazine article on intelligent user interfaces? In the past, the kind of intelligence that we modeled in AI was domain-oriented smarts. And the kind of intelligence with which we imbued our user interfaces was domain intelligence. But, increasingly we are realizing that other kinds of intelligence will also be helpful to machines, interesting to model, and useful in the interface with users. In this article I presented a model that reconciles two kinds of intelligence – interactional and propositional (or, social and domain-content) – and demonstrated how both kinds of intelligence, and the interaction between them, can make interfaces more intelligent in their dealings with humans,

and how their very presence can facilitate interaction with computational intelligence. The model relies on the functionality of embodiment: the conversational protocols that bodies support, and the kinds of representations that are carried by embodied behaviors.

As humans interact more with software agents, and come to rely on them more, it becomes more important (not less) that the systems rely on the same interactional rules that other humans do. If you're going to lean over and talk to the dishwasher as an aside while conversing with your daughter-in-law, the dishwasher had better obey the same conversational protocols, such as knowing when it is being addressed, what counts as an interruption, and so forth.

The point is that when it comes to interaction with an intelligent entity – human or machine – people have an enormous amount of knowledge and skill that can be leveraged, if we can understand their expectations, and build on the basis of how they understand the structure built into the world. Our goal is not to make every interface anthropomorphic, but to remember that for contexts like the smart room or ubiquitous computer, where users are expected to be fully surrounded by intelligent systems, the principles of spatialized interaction and embodied conversation may helpfully locate intelligence for users. So, our position is to base our work solidly in the theory of human conversation, and in the role of the human body in conversation. Our goal is embodied conversational agents that can engage humans in natural face-to-face conversation, including speech, nonverbal modalities such as gesture, gaze, intonation, body posture. These ECAs are not just there for prettiness (in fact they are often not particularly pretty at all), but they do engage in the proper use of conversational protocols and functions. Their purpose is to leverage users' knowledge of how to conduct face-to-face conversation, leverage users' natural tendencies to attribute humanness to the interface, and at the same time, embodied conversational agents also allow us to extend theories of cross-modality and modality-independent representation and intelligence.

Acknowledgements
Thanks to the other members of the REA team – in particular Timothy Bickmore and Hannes Vilhjálmsson - for their contribution to the work and comments on this paper. Research leading to the preparation of this article was supported by the National Science Foundation (award IIS-9618939), AT&T, Deutsche Telekom, and other generous sponsors of the MIT Media Lab.

References

Andre, E. (1997). "Animated Interface Agents, Making Them Intelligent". Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan, IJCAI.

Astington, J. W., P. L. Harris, et al., Eds. (1988). Developing Theories of Mind. Cambridge, Cambridge University Press.

Azarbayejani, A., C. Wren, et al. (1996). "Real-time 3-D tracking of the human body". Proceedings of IMAGE'COM, Bordeaux, France.

Bates, E., I. Bretherton, et al. (1983). Names, Gestures, and Objects: Symbolization in Infancy and Aphasia. Children's Language. K. E. Nelson. Hillsdale, NJ, Lawrence Erlbaum Associates. **4:** 59-123.

Brennan, S. E. and E. A. Hulteen (1995). “Interaction and Feedback in a Spoken Language System: A Theoretical Framework.” Knowledge-Based Systems **8**(2-3): 143-151.

Brooks, R., C. Brezeal, et al. (1998). "Alternative Essences of Intelligence". Proceedings of the American Association of Artificial Intelligence Annual Meeting.

Campbell, L. (forthcoming). Visual Classification of Discourse Gestures for Gesture Understanding. Media Lab. Cambridge, MA, MIT.

Cassell, J., M. Ananny, et al. (2000). "Shared Reality: Physical Collaboration with a Virtual Peer". Proceedings of CHI 2000, The Hague, The Netherlands.

Cassell, J. and T. Bickmore (2000). “External Manifestations of Trustworthiness in the Interface.” Communications of the ACM **43**(12).

Cassell, J. and T. Bickmore (submitted for publication). “Negotiated Collusion: Modeling Social Language and its Interpersonal Effects in Intelligent Agents.” User Modeling and Adaptive Interfaces.

Cassell, J., D. McNeill, et al. (1999). “Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information.” Pragmatics and Cognition **7**(1): 1-33.

Cassell, J., M. Stone, et al. (2000). "Coordination and context-dependence in the generation of embodied conversation". Proceedings of INLG 2000.

Cassell, J. and K. 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, J. and H. Vilhjálmsson (1999). “Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous.” Autonomous Agents and Multi-Agent Systems **2**: 45-64.

Don, A., S. Brennan, et al. (1992). "Anthropomorphism: From Eliza to Terminator 2". Proceedings of CHI, ACM Press**:** 67-70.

Duncan, S. (1974). “On the structure of speaker-auditor interaction during speaking turns.” Language in Society **3**: 161-180.

Elliott, C. and J. Brzezinski (1998). Autonomous Agents as Synthetic Characters. AI Magazine. **Summer 1998:** 13-30.

Feiner, S. and K. McKeown (1991). "Automating the Generation of Coordinated Multimedia Explanations." IEEE Computer **24**(10): 33-41.

Goldin-Meadow, S., M. W. Alibali, et al. (1993). "Transitions in Concept Acquisition: Using the Hands to Read the Mind." Psychological Review **100**(2): 279-297.

Krauss, R. M., P. Morrel-Samuels, et al. (1991). "Do Conversational Hand Gestures Communicate?" Journal of Personality and Social Psychology **61**(5): 743-754.

Laurel, B. (1990). The Art of Human-Computer Interface Design. New York, Addison-Wesley.

Lester, J. C. and B. A. Stone (1997). "Increasing Believability in Animated Pedagogical Agents". Proceedings of Autonomous Agents 97, Marina Del Rey, CA**:** 16-21.

Maybury, M. T. and W. Wahlster (1998). Introduction. Readings in Intelligent User Interfaces. M. T. Maybury and W. Wahlster. San Francisco, CA, Morgan Kaufmann**:** 1-38.

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. Chicago, IL/London, UK, The University of Chicago Press.

Piaget, J. (1952). The Origins of Intelligence in Children. New York, International Universities Press.

Reeves, B. and C. Nass (1996). The Media Equation: how people treat computers, televisions and new media like real people and places. Cambridge, Cambridge University Press.

Rickel, J. and W. L. Johnson (1998). "Animated Agents for Procedural Traning in Virtual Reality: Perception, Cognition and Motor Control." Applied Artificial Intelligence.

Riskin, J. (1999). "Moving Anatomies". Proceedings of History of Science Society, Pittsburgh.

Rogers, W. (1978). "The Contribution of Kinesic Illustrators towards the Comprehension of Verbal Behavior within Utterances." Human Communication Research **5**: 54-62.

Rosenschein, S. and L. Kaelbling (1986). "The Synthesis of Machines with Provable Epistemic Properties". Proceedings of Conference on Theoretical Aspects of Reasoning about Knowledge, Los Altos, California, Morgan Kaufmann Publisher**:** 83-98.

Smith, J. (2000). GrandChair: Conversational Collection of Family Stories. Media Lab. Cambridge, MA, MIT.

Takeuchi, A. and K. Nagao (1993). "Communiative facial displays as a new conversational modality". Proceedings of InterCHI '93, Amsterdam, Netherlands, ACM**:** 187-193.

Thompson, L. and D. Massaro (1986). “Evaluation and Integration of Speech and Pointing Gestures during Referential Understanding.” Journal of Experimental Child Psychology **42**: 144-168.

Trevarthen, C. (1986.). Sharing makes sense: intersubjectivity and the making of an infant's meaning. Language topics: essays in honour of M. Halliday. R. Steele and T. Threadgold. Amsterdam, John Benjamins**:** 177-200.

Wahlster, W., E. Andre, et al. (1991). "Designing illustrated texts". Proceedings of EACL'91**:** 8-14.

Yan, H. (2000). Paired Speech and Gesture Generation in Embodied Conversational Agents. MIT Media Lab. Cambridge, MA, MIT.