Paper presented at AAAI Fall Symposium, “Psychological models of communication in collaborative systems,” North Falmouth, Masachusetts, November 5-7, 1999

How do Real People Communicate with Virtual Partners?

Herbert H. Clark

Department of Psychology
Stanford University
Stanford CA 94305-2130
herb @ psych.stanford.edu

Abstract

Disembodied language is language that is not being produced by an actual speaker at the moment it is being interpreted. That type of language is all around us—as written language and mechanized speech—and yet it is poorly understood. My proposal is that we interpret disembodied language in two layers of coordinated activity. In the first layer, we join the producer of the disembodied language in creating a pretense. In the second layer, which represents that joint pretense, we communicate with a virtual partner. I argue that layering like this is recruited whenever we interpret forms of disembodied communication in computers.

Introduction

When my father went to the office fifty years ago, he worked with both tools and people. The tools he worked with—adding machines, typewriters, desks, chairs—had publicly recognized designs, and he *used* or *applied* them according to their design. The people he worked with had publicly recognized roles—as customers, inspectors, secretaries, and plumbers—and he worked with them in these roles. But he didn’t use or apply these agents. He *participated* with them *in joint activities* that required them to coordinate with each other. And they achieved this coordination by communicating with each other. My father had a clear understanding of the difference between tools and agents.

I, too, work in an office with tools and people, but I also work on computers. Are computers tools or agents? Well, neither one exactly. Along with many others, I have come to view them—at least for many purposes—as imaginary spaces with imaginary objects in them. Some of these objects—magnifying glasses, paint brushes, crayons—are *virtual tools* that we use or apply as if they were real tools. Others are *virtual agents* with whom we carry out joint activities as if they were real agents. The trouble with most virtual agents is that they are hidden, their personalities are inconsistent, their identities are unclear, and their capacities are vague. Often, we don’t even recognize their existence. But if we don’t, are they really there? I will argue that they are. Whether we realize it or not, we create virtual agents every time we use or interpret words, sentences, and other forms of communication on computers. We create them because we have to.

The question I take up here is simple: What does it mean for a *real* person to participate with a *virtual* agent in a joint activity? At first, the question seems absurd. How can real people engage in work with fictions—with people who don’t really exist? But people have been working with fictional partners for millennia—long before the advent of computers. It is only their incarnation on computers that is new. Once we understand how virtual joint activities work in general, we will be better equipped to understand virtual agents on computers.

Disembodied Language

The basic setting for using language is face-to-face conversation (see Clark, 1996; Fillmore, 1981), but there are other settings as well. Many of these rely on what I will call *disembodied language*. This is language that is not being produced by an actual speaker at the moment it is being interpreted. Disembodied language takes two main forms. One is written language—newspaper articles, novels, cook books, street signs, food labels. The other is mechanized speech—pre-recorded television shows, recorded telephone messages, books on tape, pre-recorded fire alarms (“There is a fire in the building: Leave immediately by the nearest exit”). Both forms are exploited in computers, so it is important to understand how they work.

Creating a Virtual Partner

Several years ago I got a postcard from my sister that read as follows:

> Paris, June 1
> Dear Herb: Paris is wonderful. We spent yesterday in the Louvre. Today we visit Notre Dame and go shopping on the Boule Miche. Having a great time. Wish you were here. Love, Margaret

Here is an ordinary piece of disembodied language, but how did I manage to interpret it. Traditional speech act theories (e.g., Bach & Harnish, 1979; Searle, 1969) are designed to account for what happens when a speaker utters a sentence to an addressee in person. Can they be extended to account for Margaret's disembodied writing? Well, Margaret scribbled her words on the postcard, and that might seem to be the analogue to uttering the words to me in person. But she also addressed the postcard, placed a stamp on it, and dropped it into a mailbox, which were also essential to communicating with me. Were these actions part of her communicative act, or were they auxiliary to it? Speech act theories are simply not equipped to say.[1]

My proposal is this: *Disembodied language is to be taken as a representation of embodied language, and to interpret it, we are intended to imagine the embodied language it represents.* To understand the words on Margaret's postcard, I had to imagine her uttering them in the right order, with her voice, accent, and intonation, and against her and my common ground. Knowing Margaret as I do, I saw her in my mind's eye, heard her in my mind's ear, and recognized her allusions. She, of course, realized I had to do this, so she engineered her part of our joint undertaking, her postcard, to allow me to simulate her speech as well as possible. I, in turn, knew that she had done this and proceeded accordingly. In brief, she and I coordinated on this communication, even though we took our actions at different times and in different places.

Disembodied speech is rarely as well defined as for Margaret's postcard. A few weeks ago, I got out *The Fannie Farmer Cookbook*, found the recipe for banana nut bread, and baked a loaf. Here is the recipe:

> 3 ripe bananas, well mashed
>
> 2 eggs, well beaten
>
> 2 cups flour
>
> 3/4 cup sugar
>
> 1 teaspoon salt
>
> 1 teaspoon baking soda
>
> 1/2 cup coarsely chopped walnuts
>
> Preheat the oven to 350˚F (180˚C). Grease a loaf pan. Mix the bananas and eggs together in a large bowl. Stir in the flour, sugar, salt, and baking soda. Add the walnuts and blend. Put the batter in the pan and bake for 1 hour. Remove from the pan to a rack. Serve still warm or cooled, as you like it.

Like the postcard, the recipe is disembodied language, so I had to imagine the embodied language it represented. But how I went about that was very different.

(1) *Virtual speaker*. For the postcard, I imagined the virtual speaker to be Margaret. But for the recipe, who I should imagine? A male or female, tall or short, with a Boston accent or with some other accent? The virtual speaker I was intended to imagine was, instead, an anonymous chef. With disembodied language, virtual speakers can range from identifiable and vividly imaginable to anonymous and indistinct.

(2) *Producer*. Behind every piece of disembodied language is a *producer*—the person or institution ultimately responsible for it. The producer of the postcard was Margaret. As for the cookbook, it was originally written in 1896 by Fannie Farmer, but the twelfth edition was "revised by Marion Cunningham with Jeri Laber" and who knows how many minor writers and editors. The producer was a collection of people I will call the Fannie Farmer folks. The point is that the producer and the virtual speaker are distinct roles. Sometimes they are the same person, as with the postcard, but often they are not, as with the recipe.

(3) *Virtual time*. When Margaret wrote "Today we visit Notre Dame," she intended me to imagine her saying these words on June 1, the day she wrote them. Otherwise, I would misidentify the referent for *today*. Here, virtual time = creation time. But when I read "Add the walnuts and blend," I imagined the virtual speaker giving me the instruction at the moment I read it. Here, virtual time = interpretation time. Every time I reread Margaret's postcard, I imagine the same event—Margaret speaking to me on June 1. But each time I use the recipe, I imagine a new instance of the chef giving me instructions.

(4) *Pacing*. When I read Margaret's postcard, I imagined her uttering the entire message in one stretch. But when I used the recipe, I proceeded one sentence at a time. I first read "3 ripe bananas mashed," got out some bananas, and mashed them. Then I read "2 eggs, well beaten," got out two eggs, and beat them. Once I had assembled the ingredients, I read the first instruction, "preheat the oven," and turned on the oven. In effect, the virtual speaker and I proceeded interactively, even though the pacing was under my control.

The postcard and the recipe, then, aren't communicative acts in the traditional sense. They are props that the producers designed to get me to *imagine* certain communicative acts—Margaret telling me things on June 1 and the chef telling me how to bake bread. Intuitively, we need two domains of interaction. In one, the producers (Margaret and the Fannie Farmer folks) design props for their recipients to use. In the other, a virtual speaker is speaking to an addressee. The concept we need is what I have called *layering*.

Layers of Joint Actions

Many joint activities divide into *layers* of joint actions (Clark, 1996).[2] Suppose Beth and Calvin, two six-year-olds, get into the family car and pretend to be Mother and Father driving around town. What they are literally doing

[1] What if Margaret had dictated the words to Duncan, who scribbled the message and the address on the postcard, which she later stamped and posted? I would have interpreted the postcard in the same way.

[2] The notion of layering is derived from work by Bruce (1981), Goffman (1974), and Walton (1973, 1978, 1990) and is closely related to Laurel's (1991) conception of computers as theater.

is sitting in the car, making noises, turning the steering wheel, creating a joint pretense. That is one layer of joint activity. But in the pretense itself, Mother and Father are driving around town. That is a second layer, which is created on top of the first:

Layer 1: Beth and Calvin jointly pretend that the events in layer 2 are taking place.

Layer 2: Mother and Father are driving around town.

I picture layer 1 at ground level and layer 2 on a raised stage above layer 1. And there can be further layers above that.

Layers are needed for interpreting Beth's and Calvin's actions correctly. When Beth turns the steering wheel, her action has two interpretations. In layer 1, Beth is turning the steering wheel of a motionless car, but in layer 2, Mother is turning the corner into Lincoln Avenue. Layers are just as essential for interpreting the language that Beth and Calvin used. When Beth says, "Why don't we pretend to drive the car," she is speaking as part of layer 1, and *we* refers to Beth and Calvin. But when she says, "Now we're driving down Lincoln Avenue," she is speaking as part of layer 2, and *we* refers to Mother and Father. Layering arises in many types of communication, from fiction, jokes, plays, and operas to such everyday tropes as irony, sarcasm, hyperbole, rhetorical questions, and bantering (see Clark, 1996).

Layering is just what we need to account for disembodied language. When Margaret sent me the postcard, she expected me to join her in the pretense that she was uttering these very words to me on June 1. She intended me to create these two layers of joint actions:

Layer 1: Margaret (the producer) intends me to use her postcard as a prop for the joint pretense that the events in layer 2 are taking place.

Layer 2: Margaret (the virtual speaker) is speaking to me on June 1.

Similarly, the Fannie Farmer folks intended recipients like me to pretend with them that an actual chef was telling us step by step how to bake the bread. The two layers are these:

Layer 1: The Fannie Farmer folks (the producers) intend the recipient (me) to use the printed recipe as a prop for the joint pretense that the events in layer 2 are taking place.

Layer 2: A chef (the virtual speaker) is instructing the addressee (me) now on how to bake a loaf of banana nut bread.

Layering captures the intuition that disembodied language reflects two domains of action. In the first layer, the participants are the producers and the recipient, and in the second, the participants are the virtual speakers and their addressees. Also, actions in the first layer take place entirely in the actual world. My sister uses an actual pen to scribble on a tangible postcard, which is physically transported by the post office and winds up in my actual hands. Actions in the second layer make reference to a mix of real and imaginary objects. There are virtual speakers and imagined voices, but actual addressees, bananas, and eggs.

Successful Layering

Layers take skill to create. When Beth and Calvin pretend to be Mother and Father driving around town, they have to arrange who is to be Mother and who Father, where they are driving, and so on. The same goes for disembodied language, as the producer arranges with the recipient to get the pretense just right. Here I focus on two parts of this process—*characterization* and *props* (see also Laurel, 1991).

For layering to be successful, there must be good characterization. For Beth and Calvin's game to come off well, the Mother and Father must be credible characters. The same goes for the virtual speakers in the postcard and recipe. A character seems to be successful if it satisfies at least these criteria.

(1) The character must be *consistent*. My sister, as a virtual speaker, shouldn't do anything out of character, nor should the chef. It would be inconsistent for the chef suddenly to become abusive ("Now don't be an idiot and beat the eggs too long").

(2) The character must be *appropriate for the pretense*. The virtual speaker for the recipe should be a chef, not a doctor, plumber, or rock star, and should display the knowledge appropriate to a competent chef. The virtual speaker in the postcard should be my sister as traveler. By this criterion, the virtual speaker will sometimes be clearly identifiable (as Margaret is in the postcard), but other times anonymous and indistinct (as the chef is in the recipe).

(3) The character must be *easy to imagine*. It is no good if a virtual speaker is consistent and appropriate, but cannot be mentally simulated by the recipient.

Successful layering also depends on the props—the materials the producers designed for their recipients to work from. Here are several criteria for good design:

(1) The props should be *sufficient for the purposes*. Margaret's postcard should make it clear that it is from her, that she is in Paris, that the date is June 1, that she is addressing me, etc. I need all this to recreate her communicative acts appropriately.

(2) The props must *adhere to the right conventions*. Margaret appealed to many conventions in constructing her postcard—English writing, American script, the form of salutations on postcards ("Dear Herb"), the way to date a postcard (the date of writing, not receipt), etc. Many of the conventions in the cookbook are quite different—and they would be out of place on the postcard. Recipients take these conventions for granted, so the conventions should be appropriate.

(3) The props must *help recipients imagine* the appropriate characters and objects. Consider what Van

Der Wege and I have called *mimetic props* (Clark & Van Der Wege, 2000; see also Laurel, 1991). These include pictures, gestures, sounds, and other non-symbolic aids to imaging a scene, its inhabitants, and its furniture. The picture of the Louvre on Margaret's postcard helped me imagine what she had done the day before, and the cookbook uses illustrations in many of its instructions. The recipe also refers to the real objects I was handling—the bananas, eggs, oven, etc. These, too, become part of the scene the chef and I are creating together.

Successful layering is an art. It depends not only on the skill of the producers, but on the interests, abilities, and cooperation of the recipients.

Agents and Tools on Computers

Computer systems are inhabited by armies of agents—whether we recognize them or not. Every operating system and every computer application relies heavily on disembodied language. To understand that language, we users have to collude with its producer in the pretense that we are engaged with a virtual agent in a joint activity, and that we are communicating with that agent in order to carry it out. So whenever we use an operating system or application, we create virtual agents.

Word Processors: An Example

When I use MS Word, I do not work alone. When I want to delete a word, I select it by pointing at it with the cursor and double clicking on it, and then I delete the word by pressing "delete." How am I to understand the actions of pointing, double clicking, and pressing "delete"? Note that these actions are analogous to signals in face-to-face conversation between, say, Bill and me. Like the first action, I can point at an object for Bill—just as long as I can bring it into our joint attention. Like the second action, I can say "That book" to signal to Bill that I am indicating a book and not an entire bookshelf. And like the third action, I can say "Fetch," by which I mean that Bill is to bring me that book. So in taking these three actions, I am, in effect, asking a virtual editor to delete the word, and the editor complies by deleting it.

Do we really need the virtual editor? The answer is yes, if only to handle the disembodied word "delete." To understand that word, I have to make the joint pretense with the producer of MS Word that a virtual editor is offering to delete the selected bit of text. Indeed, the editor is offering me other options as well. If I clicked on "boldface," "italics," or "copy," I would be asking the editor to turn the word into boldface or italics or to copy it. The notion of virtual editor is needed to help me understand other aspects of MS Word as well, including these:

1. *Communicative acts*. Many of my actions can be readily understood only if they are construed as communicative acts to the virtual editor. Why do I move the cursor? Because I am pointing at something for someone—the virtual editor. Why do I double click on the word to be deleted? Because I am signaling that that something is a word and not a letter, sentence, or line. Signaling whom? The virtual editor.
2. *Communicative exchanges*. The virtual editor's actions make good sense as the next step in an interaction. The moment I say "delete," the editor complies and makes the word disappear—just as a real editor would have done. Many of these exchanges resemble adjacency pairs in language use (e.g., question and answer), and they would fail if they didn't.
3. *Real and virtual artifacts*. The virtual editor and I are not merely communicating. We are working on an artifact, a document, just as a real editor and I might do in creating a physical document. At first, the artifact is merely virtual, a representation of something concrete, but in the end, it is turned into a physical document. That is, the virtual editor and I work with artifacts, real and virtual, much as the virtual chef and I worked with real bananas, real eggs, and a real oven.
4. *Real and virtual tools*. The virtual editor and I use tools. Some of these are virtual tools, such as a magnifying glass or drawing tool. Others are physical tools, such as the printer. Ambiguities, however, abound. Is the virtual editor checking the spelling in consultation with a virtual dictionary, or is the virtual editor applying a virtual tool called the spell-checker? It isn't clear whether spelling is checked by a virtual agent or a virtual tool.

It is my interacting with the virtual editor that ties all these procedures together in a single activity.

Successful Agents

Virtual agents should be effective in computers on the same grounds that they are effective for other forms of disembodied language. The virtual agent must be (1) consistent in character, (2) appropriate for the pretense, and (3) easy to imagine. And the props behind the virtual agents should be well engineered. They should be (1) sufficient for the purposes, (2) consistent with the conventions of the joint activity at hand, and (3) an aid to users imagining the characters. There is good evidence that characterization and props are in fact important for successful agents.

The most impressive body of evidence for this proposition comes from the work of Reeves and Nass (1996) on *social responses* to computers, television, and other media. What they have shown in study after study is that people respond to computers and other media in much the way they would respond to other humans. In one study, people played a game of twenty questions on the computer and, at one point in the game, were praised or criticized by the computer for their responses. People who were praised did better at the game and liked the computer better than the people who were criticized. In other studies, people reacted to politeness, male and female voices, dominant and submissive personalities, arousing messages, and many other features of Reeves and Nass's clever computer designs much as they would have reacted to the same features in humans.

These findings are well accounted for by the layering model of disembodied language. It is well known that the more engrossed people get in a pretense—a horror film, adventure novel, or stage play—the more they experience emotions and thoughts much like those they would experience in the real horror scene, adventure, or social predicament (Gerrig, 1993; Walton, 1978). Furthermore, people are good at compartmentalizing these imagined experiences from actual experiences—at foregrounding the imagined scene and backgrounding the author or producers of the props (Gerrig, 1989). The people who played twenty questions in Reeves and Nass's study engaged in just such a pretense. They used the props created by the producers of the game to help them pretend that a virtual speaker was asking them questions. Engrossed in that imagined world, they should have felt flattered by the virtual speaker's praise, and stung by the criticism, and they did.

In discussing their work, Reeves and Nass take up the classic notion of "willing suspension of disbelief" and reject it:

> The claim is that people can intentionally forget about the fact that media are artificial and produced elsewhere and can pretend that what's in front of them is real… [But] the next time you see a play, try to suspend belief. Try to ignore the characters and sets on the stage and simply think about the author who enabled it all… The traditional notion of "suspension of disbelief," however, suggests that it is work to accept the reality of what is present (pp. 186-7).

Reeves and Nass appear to take these and other observations as evidence against all accounts based on pretense and imagination.

Layering, however, is fully consistent with Reeves and Nass's observations. Joint pretense does *not* involve the willing suspension of disbelief (Walton, 1978). Just the opposite because it requires *two* layers of action. People often get deeply engrossed in the imagined world of the pretense (layer 2), but they always come back to reality (layer 1) and recognize it when they do. And when people get engrossed in the joint pretense (layer 2), they push the producer and the props of layer 1 into the background. I can get deeply engrossed in Margaret speaking to me without knowing precisely how she produced the postcard. And I can get deeply engrossed with the chef in baking bread without knowing how the Fannie Farmer folks produced the cookbook. All I need to be able to do is imagine the virtual speakers and their communicative acts toward me.

Conclusions

Disembodied language is all around us—in newspapers, on street signs, on television and radio, on computers—and yet it is poorly understood. My proposal is that it requires two layers of joint activity. In layer 1, the producers of the disembodied language and I coordinate in creating a joint pretense—namely, the world of layer 2. In layer 2, I communicate with a virtual speaker as we do things together. In this view, many features of computer design lose their mystery. They are simply the props that support the type of joint pretense in which virtual agents and I work together, often with virtual tools, in order to accomplish real tasks.

References

Bach, K., and Harnish, R. M. 1979. *Linguistic Communication and Speech Acts*. Cambridge: MIT Press.

Bruce, B. 1981. A Social Interaction Model Of Reading. *Discourse Processes*, 4, 273-311.

Clark, H. H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., and Van Der Wege, M. 2000. Imagination in Discourse. In D. Schiffrin and D. Tannen Eds., *Handbook of Discourse*. Oxford: Basil Blackwell. Forthcoming

Fillmore, C. 1981. Pragmatics and the Description of Discourse. In P. Cole Ed., *Radical Pragmatics*, Pp. 143-166. New York: Academic Press.

Gerrig, R. J. 1989. Suspense in the Absence of Uncertainty. *Journal Of Memory and Language, 28*, 633-648.

Gerrig, R. J. 1993. *Experiencing Narrative Worlds: On the Psychological Activities of Reading*. New Haven CN: Yale University Press.

Goffman, E. 1974. *Frame Analysis*. New York: Harper and Row.

Laurel, B. 1991. *Computers as Theatre*. Reading, Massachusetts: Addison-Wesley.

Reeves, B., and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge:: Cambridge University Press.

Searle, J. R. 1969. *Speech Acts*. Cambridge: Cambridge University Press.

Walton, K. L. 1973. Pictures and Make-Believe. *Philosophical Review, 82*, 283-319.

Walton, K. L. 1978. Fearing Fictions. *Journal Of Philosophy*, 75, 5-27.

Walton, K. L. 1990. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.