Face-to-Face Interface

2.

In this chapter we will discuss the general advantages and disadvantages of face-to-face interaction and how this relates to human-computer interaction, and look at some of the early history of humanoid agents. We will also take a non-traditional look at the issue of anthropomorphization—the act of attributing human qualities to non-human things.

....at every screen are two powerful information-processing capabilities, human and computer. Yet all communication between the two must pass through the low-resolution, narrow-band video display terminal, which chokes off fast, precise, and complex communication.

—Edward R. Tufte (1990, p. 89)

2.1 Humanoid Agents: Early History

The fascination with humanoid, artificial agents can be traced at least to the beginning of this century—not in research but in fiction. The first multimodal, interactive agents were probably Karel Capêk's mechanoids, in his play *R.U.R.* ("Rossum's Universal Robots") [1920]. This piece is the origin of the word "robot", the Czech word for "worker". Another landmark in robot fiction was Fritz Lange's *Metropolis* [1925], sporting a robot that was so believable it was virtually indistinguishable from humans. An all-time favorite multimodal agent in fiction was the artistically designed Robbie the Robot, first making its appearance in the movie *Forbidden Planet* [1956] and in many others after. Toward the latter half of the century we witnessed the appearance of an awe-inspiring HAL-9000 computer in Kubrick's *2001: A Space Odyssey* [1968] (communicating through multimodal input but only speech output), C3PO of *Star Wars* [1977] (multimodal I/O), and Holly—the ever cynical computer on-board the spaceship *Red Dwarf* [1988] from the BBC series with the same name. Holly is identical to HAL-9000 except for the very important aspect of having an embodiment as an on-screen face, entering the world of the user and capable of multimodal output. It seems that in fiction through the ages multimodal interaction has always been assumed; perhaps because it comes so naturally to us it has never seemed an issue. And, perhaps because robot researchers have been

FIGURE 2-1. Robbie the Robot saves its master.

busy working on vision, smarts and action control in separate corners of their laboratories, the issue hasn't really come up there either, until very recently.

In various recent "visions of the future" promotional videos, companies like Hewlett-Packard and Apple Computer [Laurel 1992] have presented the idea of agents that inhabit the world of the computer, but seem to have at least limited perception for outside things like the user's presence. These agents communicate mostly via speech and visual appearance as output, and simply speech as input. The visual channel as input is highly de-emphasized. However, recent progress in computer vision leads us to believe that recognizing people—where they are looking and what they are doing—may well be within a decade of being commercially viable [Essa 1995, Maes at al. 1995]. The added richness of a visual input channel could well make all the difference in interacting with artificial agents, determining whether people will actually "buy"—pun intended—this kind of interaction style with machines.

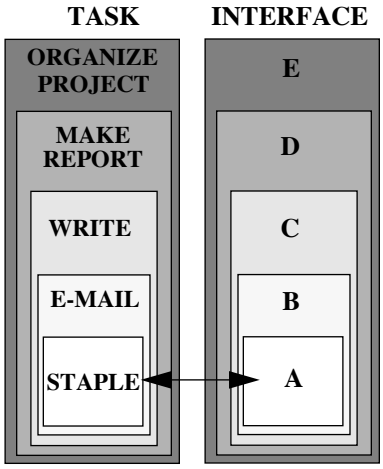

FIGURE 2-2. Darker colors indicate increasing underlying complexity, letters indicate a choice of interface. The task of stapling calls for a simple physical-tool interface such as a stapler (A); a high-level task, such as organizing a project, requires a communicative interface (E) and may require interfaces from the lower levels also. A mismatch between task and interface, for example replacing face-to-face interaction (E), with the interface for writing—a word processor (C) —is likely to compromise efficiency.

Most present-day robots have little idea about a "user" and their design is generally not "user-centered" in the usual sense of the term, although new research seems to be focusing more on this issue. For instance, Cannon's [1992] system employs a camera that the user can point at objects, give simple commands like "put that...and that...there," accompanied by a camera pointing in the directions, and the robot will automatically plan the execution of action for its mobile platform and arm. Brooks' [Brooks & Stein 1993] proposal for a humanoid robot includes a full upper body humanoid with stereo cameras for vision, stereo microphones for hearing, duplication of the human upper body degrees of freedom, and a massively parallel computer for brains. (Who needs fiction?)

While robots have changed relatively little in fiction since Câpek, research on various fronts is filling in missing knowledge and moving us closer to realizing well rounded artificial humans [Pelachaud et al. 1996, Prevost & Steedman 1994, Cassell et al. 1994, Thórisson 1994, Badler et al. 1993, Brooks & Stein 1993]. Although the main focus here is taking another step toward a new kind of interaction—not toward replacing or "bettering" any of the existing human-computer interfaces in existence—for completeness sake we will now quickly review the most obvious benefits and limiting factors of face-to-face interaction.

2.2 *Face-to-Face: When & Why*

In answering the question of when and why we would want to use a face-to-face[1] interface, two different perspectives can be taken:

1. *What kinds of tasks and systems are amenable to a face-to-face interface?*, and
2. *what are the necessary qualifications a system has to have to justify sporting a face-to-face interface?*

These issues are both really part of the same problem: how to fit an interface to a system (Figure 2-2). The issue boils down to two simple arguments: {1} Certain kinds of real-world tasks, namely supervision, need different interaction methods, namely communication, and {2} better systems can better support the complexity of natural interaction methods such as language, gesture and facial expression.

When determining the kind of interface for a system, we need to ask ourselves *What is the system capable of doing?* In other words *What is the nature of the task?* It makes little sense to install advanced speech recognition and in a normal, dumb, toaster when all it can do is turn on and off—the interface needed for such a dumb device is simply a switch labeled *"off-on"*. By the same token, if the toaster is extremely intelligent and can do many different things besides toast bread, crumpets or bagels[2], it is equally inappropriate to provide a user with a single on-off switch to interact with it. Because the relationship between the user action (turning the toaster on) and the outcome (the toaster heating up) is always the same—it is a completely reflex-based system. Such systems don't make any decisions of their own; they follow blindly the user's input. Even in systems such as nuclear power plants, which are orders of magnitude more complex than toasters, the interface is based on the same principle. The main difference is that the number of variables is exponentially higher, and the operators of a nuclear power plant have trained for months in how to interact with the system (plant) through the interface (control room, Figure 2-3).

Laurel [1992] lists some of the chores that intelligent interface agents might help with (Figure 2-4). A number of these tasks can be communicated about with less-than-human multimodal capabilities. Sheth and Maes [1993] for example describe agents that retrieve, filter, sort and organize a person's electronic news that simply use a "point-and-click" interface. So why would we even want to discuss face-to-face interaction?

On the other side of the coin is the intelligence level of the system: is the system really capable of supporting a face-to-face interaction? Can it support natural language without constant misunderstandings, break-

FIGURE 2-3. A nuclear power plant is really just a giant toaster. (Control Room One, D.C. Cook Nuclear Power plant, Ann Arbor, Michigan.)

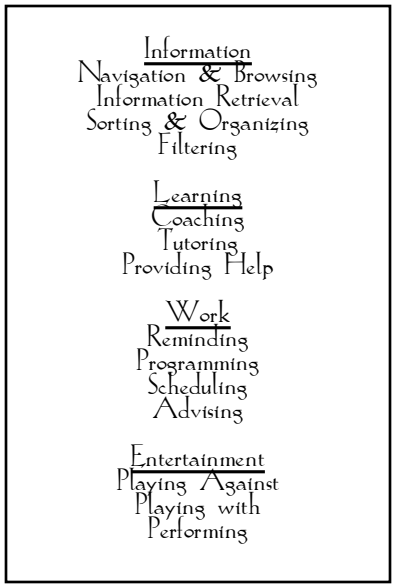

FIGURE 2-4. Laurel [1992] suggests these kinds of tasks as being ideal to delegate to a semi-autonomous computer agent.

1. I use the term "face-to-face" not only to refer to the presence of faces, but in general to the issue of co-presence and non-mediated communication.
2. My hat goes off to Grant and Naylor, the writers of Red Dwarf [1988], for the AI toaster example. You have to see it.

downs in communication or breaches in trust? Is it capable of feedback at multiple levels of granularity [Thórisson 1994, Clark & Brennan 1990]? If it is missing any of these, then it may be a mistake to try to force the link. However, even a dog-level intelligence justifies a multimodal interface, as evidenced by all the dog owners who happily use prosody, gesture and keywords to interact meaningfully with the canine friends. It may well be that dogs are an example of the lowest-level, non-verbal intelligence worthy of a multimodal interface. As the system gets better at understanding language, making its own decisions and executing actions, the need for a richer interface arises. The more capable a system is, generally speaking, the more complex the tasks that can be delegated to it, the more it makes sense to use high-level interaction methods like natural language. If the task to be accomplished with the system is formulated at a high level, including concepts such as *goals*, *intentions*, *plans*, etc., then it is very likely that natural language and a face-to-face metaphor will be a useful interface to the system. Natural language is also unique in that it allows for a kind of "downward compatibility": it has great flexibility in the level of detail it can address ("I want to make a pizza"; "is the oven on?"), and in that way does not paint a user in a corner.

Before discussing when *not* to use face-to-face interaction, let's look at some more reasons *for* it.

2.2.1 Some Compelling Reasons for Interacting Face-to-Face

A common knee-jerk reaction to the face-to-face interface might be the following: Interacting multimodally is really our only choice in the case of humans and animals, but *Why should we even consider communicating with a computer program in the form of a human when we can interact with it in a million other ways?* In this section I will review the most general arguments for using agent-based interfaces and in the next discuss some of the limitations of these as well as the relationship between the choice of interface with relation to a task.

The arguments in favor of the face-to-face metaphor can be divided into two categories: implementation dependent—related to particular realizations of multimodal systems—and implementation independent—related to the psychological and physical makeup of human beings. The latter is based on 4 key points:

1. *Synergy,*
2. *naturalness,*
3. *flexibility and*
4. *limits of metaphor*.

We will first take a look at these, and then discuss the implementation dependent arguments.

Synergy

The issue of benefits involves several interconnected questions: Why multiple modes? Why dialogue? Why face-to-face interaction? I say interconnected because in face-to-face interaction, sensory organs, bodily constraints, attentional and other mental limitations are linked together in a way that is intimately integrated and intertwined with the dialogue process. If we want to interact with an intelligent machine, it is therefore a big win if we model its interface in our own image, i.e. with a head, face, gaze, arms, hands, and a body—organs that have to do with communication.

Bolt [1987] discusses some of the strongest arguments for such multimodal interfaces. He points out the clear benefit of increased redundancy in the input, potentially reducing errors in the communication. Signals that occur in verbal communication are tightly linked with nonverbal cues. Recognizing both of these can increase the reliability of the interaction. He also points out the added richness of a multimodal interface: different modes have different ways of communicating. A face is an incredibly rich information display [Tufte 1983], and, more importantly, a natural part of the human communication mechanism. It is important to recognize that this argument serves on both sides of the equation, not only for input, but for computer output as well. These points relate to the *synergy* of multimodal communication, i.e. they argue in favor of an interface that integrates many features of face-to-face interaction rather than one that selects or singles out one or two features in isolation.

Related to the point of synergy is the following argument: It is the year 2010. I walk up to a speech recognizer in a train station I have never been to before to buy tickets. What kinds of words am I allowed to use? What kind of sentences are acceptable? Just speaking into a microphone, it is hard enough to pace the interaction, not to mention selecting the right things to say. When interacting with beasts of unknown intelligence, with vaguely known skills and unknown linguistic capabilities, we humans need all the help we can get to make it easier to predict what kinds of things we can talk about with it, what kinds of words we may use and what kind of performance we may expect from it. This information can be given by the interactive intelligence's appearance, body language, facial expressions, gaze behavior and turn taking skills. The stilted way I am asked if I can be helped, the fact that the face on the screen looks non-human, the hesitating manner of answering, the jerkiness of the smile all tell me that I should use simple language and get straight to the point as I ask if for the schedule of the D-train.

Naturalness

Why dialogue? Why not just write a letter or send the computer an e-mail to tell it what we want? Dialogue is structured around the turn. People cannot, as it were, talk and listen at the same time. Turn taking makes using language, as well as the various multimodal communicative devices, very efficient and effortless [Sacks et al. 1974]. It also makes it easier to integrate interaction between collaborating agents into an ongoing, common task by giving interlocators greater process control [Clark & Brennan 1990]. Thus, dialogue and turn taking are both an integral part of any language-based multimodal system. This relates to the *naturalness* of face-to-face interaction.

Flexibility

Why use a metaphor of human face-to-face communication instead of simply designing each system to accept exactly the kinds of modes needed for the task it is to perform? Pen and speech here, gesture and gaze there, etc. This question relates to the *flexibility* of the interface. It has a two-part answer. To a certain extent, of course, people do this when interacting with each other: we grab a pen and scribble on a napkin, they gesture at certain times and not others, etc. But notice that these options are all available in an instant, once we decide to use them. It is flexibility that makes multimodal dialogue so attractive. And although speech has been shown to be sufficient to successful human communication in many cases [Ochsman & Chapanis 1974], in its "high-bandwidth" instantiation it is accompanied by feedback mechanisms on multiple levels [Goodwin 1981, Yngve 1970]. A primary thrust for using social communication as a metaphor in human-computer-interaction stems from thus the presumed increase in "bandwidth" as when compared, for example, to command-line or graphical user interfaces [Brennan 1990, Clark & Brennan 1990], and the flexibility of being able to switch reliably between—and freely combine—gesture, language, glances and facial expressions to convey one's wishes and requests [Whittaker & Walker 1991, Bolt 1987].

The second part of the argument centers on *coordination:* pacing dialogue is difficult in the absence of feedback [Nespolous & Lecours 1986]. Thus, if we want to communicate complex commands to the computer that involve multiple steps, the best method is doing it face to face in the presence of clear, socially compliant feedback mechanisms that indicate comprehensibly to us that our commands have been understood.

Limits of Metaphor

My personal favorite support for multimodal human-computer interaction comes from a simple observation: computers are becoming more and more capable; speech recognition, gesture recognition, face recognition, object recognition ... the list goes on. If we continue to interact with computers in the old style that is modeled after the way we manipulate inanimate objects in the world, then computers will continue to appear more and more complex and confusing to their operators, until they become so cumbersome that new additions are not worth the trouble. We can try to imagine a typical error message in this hypothetical future:

> WARNING! FILE ERASE PREVENTED BY COLLABORATIVE AGREEMENT. You cannot erase those files because the trash folder recognized your voice and asked the files to verify your identity; they in turn have identified your face and warned the trash folder that they are 87% certain (average certainty for all files in question; standard deviation = 28.23%) that you don't have the right privileges to delete them.

This future hell of files with perceptual abilities, thinking folders and decision-making icons can and should be avoided. What is needed is a new interaction metaphor that takes us to the next level of human-machine relationship. Fortunately this metaphor exists: we already use such an interaction style with each other. Its called social interaction and is based on the notion of localized agency (a person is a localized agent capable of action). Since we interact with intelligent beings (agents) by communication, it only makes sense to start looking at communication as the next logical step in the evolution of the computer. And the most basic method of such interaction—the one that all others are and will continue to be compared to[3]—is face-to-face dialogue.

Implementation-Dependent Arguments

So far we have reviewed implementation independent arguments for the face-to-face metaphor. However, other more practical concerns related to technology also come into play. One relatively new line of argument for focusing on robustness in this kind of communication is the promise that future machines will be equipped with cameras that can sense their users [cf. Essa 1995, Maes et al. 1994]. By introducing cameras the user is freed from having to "dress up" into body-tracking gear such as

3. Some may object to this claim on the grounds that we could simply re-engineer ourselves to allow us wireless transmission of thought or perception of multiple places and times simultaneously. When this becomes a viable option, I am willing to reconsider my stance. In the mean time this argument will belong in the science fiction domain.

gloves and suits [see Bers 1995a]. However, because of various confounding factors such as variation in lighting, occlusion, etc., reliability in the analysis of input may be expected to drop.[4] Capturing information from multiple sources and modes will enable the computer to make more reliable inferences about the state of the dialogue and the user's input, and make it possible for the user to adapt to the situation by dynamically choosing the most appropriate mode combinations depending on the computer's multimodal responses.

A similar case can be made for speech recognition: by collecting information such as a speaker's direction of gaze, direction of head-turning, etc., a speech recognition system can know when an utterance is meant for it and when it is meant for a by-stander. Variations on this theme could allow a system to switch dynamically between vocabularies during interaction and thus increase the reliability of the recognition process.

Numerous other arguments have been put forth about the benefits of multimodal, socially-oriented interaction with machines [Brennan 1990, Laurel 1990, Bolt 1987]. However, the strongest argument for interacting socially with computers comes from the simple observation that most people in the world interact frequently with other people, and are thus constantly practicing this kind of communication.

2.2.2 Face-to-Face: When *NOT?*

We have already mentioned that if a system cannot support the most important features of face-to-face interaction, we shouldn't try to attach that kind of an interface to it. However, given that we want a communicative-style interface, when would face-to-face provide the right features? The following discussion in this section is mainly based on Clark and Brennan's and paper *Grounding in Communication* [1990] and Whittaker and Walker's *Toward a Theory of Multimodal Interaction* [1991].

Clark and Brennan's work is directed toward the process of grounding, the process in which two interacting agents come to share mutual knowledge, mutual beliefs and mutual assumptions. This process is considered to be inherent in any communication task. Whittaker and Walker [1991] show how these concepts are generalizable to the analysis of the cost of different media for various tasks in the computer domain. The key concepts the authors identify are:

4. It may be argued that a certain level of uncertainty will always be present, save perhaps for highly artificial environments, because the world is far too complex to be completely predicted, hence the increased need to ensure reliability.

1. EXPRESSIVITY — what kind of information can be conveyed in the medium?
2. PERMANENCE — how permanent is the medium; does it allow for review and revision?
3. INCREMENTALITY — what is the granularity of feedback the system can give to user actions?

They reach the conclusion that for tasks with strict requirements in permanence, speech is not a good medium—if we have the choice of a single medium only. Examples of tasks that rely heavily on permanent media are writing, drawing or construction in general. However, in almost all tasks is there a use or preference for a separate, less-permanent channel. The inverse is also true: for tasks such as brainstorming, planning or coordination that rely heavily on speech, the use of permanent media (e.g. a pencil and paper) enriches the interface tremendously, while leaving information transmission mainly to the speech channel.

If a system is highly restricted to a single level of granularity, a face-to-face metaphor is unlikely to provide the most efficient interface. Examples of such tasks would be manipulations of single, unique objects, where the need to repeat the same action on multiple objects does not exist, and that requires few or no abstract relations between objects and actions (e.g. "Find *all* Ys that are *part of* X *and* have attribute Z, *excluding* Ys that are also Ts"). The same can be said for tasks with limited need for temporal specification ("Do X and Y simultaneously, then Z") and tasks with a minimal real-time component [Walker 1989].

Rather than restricting conditions of the task, as in the above examples, restricting the transmission channel has more obvious effects on our choice of interaction method. When there is a high latency in the information transmission channel, face-to-face interaction is generally a bad choice of interaction, because to be effective it requires rapid, full duplex feedback on multiple levels in multiple modes. If the transmission medium allows for only limited bandwidth, face-to-face interaction is not feasible, since out-of-sync sound and pictures tend to disorient rather than enhance [cf. Whittaker 1994, Whittaker & O'Connaill 1993, O'Connaill et al. 1993]. This condition, however, may still be perfectly suitable for speech-only communication. Asynchronous delays in message transmission will further diminish our reasons to choose face-to-face or speech-only interaction over for example, e-mail, fill-forms, or any other method where the permanence of the transmission medium allows error-free communication to take place.

Any good theory of communication should be able to allow us to constrain at will the initial conditions of the system for which we want to design an interface, and this is precisely what the Whittaker & Walker

[1991] research tries to do. The interested reader should be able to follow this thread in more detail through their work.

an.thro.po.mor.phism *n* (1753): an interpretation of what is not human or personal in terms of human or personal characteristics: humanization — **an.thro.po.mor.phist** *n*

— Merriam-Webster's Collegiate Dictionary, Tenth Edition

2.2.3 Anthropomorphization: A Non-Traditional Perspective

Orthogonal to the task of choosing the right interface are the issues of agency, autonomy and anthropomorphization. Is anthropomorphization of an agent-based interface necessary? Considerable fuss has been made over the perceived pitfalls of anthropomorphization—the act of attributing human-like qualities to inanimate objects or animals of low intelligence. With regard to the anthropomorphization of computers, researchers seem to have varied opinions [Lanier 1995, Maes 1993, Chin 1991, Laurel 1992, Laurel et al. 1990, Laurel 1990] and the systems to date that deliberately use anthropomorphization seem as varied as people's opinions of them.

As I have already alluded to, the kind of interface chosen should be justified by system capabilities, and be suited to the task being performed with that system. In this view, whether the interface follows a face-to-face metaphor or is less a question of personal preference and more an issue of efficiency. But what about anthropomorphization—should it be avoided—can it be avoided? I would argue that for many complex tasks, there is little choice on the designer's part whether the system is presented in anthropomorphic terms or not. The argument is based on the simple observation that as systems become "smarter", i.e. become capable of handling behaviors and concepts that are normally attributed to people, like integrating various data sources, perceiving their environment and making independent decisions, understand speech, people's willingness to anthropomorphize increases. To take an example, we would have a hard time convincing anyone that a rock is autonomous or has any amount of intelligence. So, it follows that it is difficult for us to imagine a rock has having a character or being an agent. A rock represents one extreme end of a continuum from "dead" to "alive". Moving along this continuum, our ability to anthropomorphize is made somewhat easier given systems that handle simple delegation, for example systems for fetching electronic mail at certain times of the day. Because you delegate the task of fetching mail to the system, the system embodies some level of autonomy, and hence is easier to anthropomorphize. Most people would probably agree that dogs are very easily anthropomorphized. This path from dumb to smart systems indicates a trend that implies an impasse toward high intelligence: as a system becomes increasingly smarter, the designer's ability of that system to influence a user's tendency to anthropomorphize that system decreases toward nothing. I call this "The Intelligent System Designer's Deadlock", or just *Designer's Deadlock* for short.

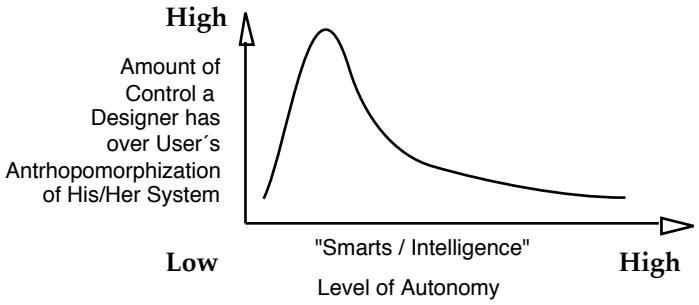

FIGURE 2-5. As the "smarts" or level of autonomy of a system increases, its designer has less and less control over the user's tendency to antrhopomorphize that system.

What we get is a curve that looks something like Figure 2-5. For any system, as we select that system's level of intelligence on the abscissa, we get a value on the mantissa that shows the freedom the system's designer has in controlling a user's anthropomorphization of that system. Somewhere toward the lower end on the "smarts" scale people are very good at imagining a system as being an agent; this is the level of our most intelligent systems today. The Designer's Deadlock effect is exacerbated as we add to the system features that borrow visual human or animal-like features like faces, hands, eyes, facial expressions, etc. Two of the strongest factors in driving home anthropomorphization of an intelligent system are probably speech (even a system capable of very limited speech may be seen as a "stupid" humanoid) and a face (even a toy with a face can be perceived as having human-like characteristics; a toy with no face has much less of a chance).

This argument is supported by recent research on users' perception of technology [Nass et al. 1994, Nass et al. 1993]. This research has shown that when computers are equipped with human-like capabilities such as speech synthesis or speech recognition—in fact, even when it communicates with simple text—users perceive them as agents with human-like capabilities. Rather than ignoring or trying to eliminate the agent-like qualities that computers are perceived to have, one can capitalize on the fact and make the interaction more stable and effective.

2.3 Summary

In this chapter we reviewed some of the early fascination in the arts with humanoid artificial agents. We discussed the advantages and disadvantages of the face-to-face interaction metaphor as applied to interaction with computers, and the nature of anthropomorphization. We con-

cluded that in spite of human-like interfaces not being problem-free, it is better to acknowledge them and try to take advantage of them in interface design, than to wish they would go away and get stuck with unsolvable problems. Whatever may be said against face-to-face interaction as a method of communication, the evidence reviewed certainly supports the argument that trying to build a computer system based on these ideas is far from being a waste of time.