41. [Introduction] From *Plans and Situated Actions*

Lucy Suchman made a fundamental critique of practices within artificial intelligence and presented a different concept of how people seek to accomplish goals. This led to significant changes within the language of artificial intelligence, although the use of similar terminologies by AI researchers is often founded on a misunderstanding of Suchman's argument.

Suchman held that the elaborate plans and symbolic manipulations that characterized artificial intelligence's early attempts to create interactive devices were fundamentally misguided. Artificial intelligence practitioners had assumed that these logical manipulations were much like the human planning process, and that once they were sufficiently refined they would, of course, succeed. Suchman argues that such elaborate abstract plans are never actually the primary basis for human action. They are better seen, she states, as stories that some of us, in some cultures, use to organize our actions. Many ponderous artificial intelligence projects underway when Suchman's book was published would have had little justification if this point had been conceded.

Some viewed Suchman's observations as a prescription for the development of new AI strategies that took situated action into account. In this interpretation, AI could still make progress within current institutional parameters; such progress would come by building systems that improvised toward a goal based on the current situation, rather than following a monolithic plan. Some, like Philip Agre and David Chapman, as explained in "What Are Plans For?," explored this possibility in a manner that was commensurate with Suchman's critique. In other cases new language was adopted to represent the same techniques that Suchman argued against, perhaps based on the increased military concern with plan-based AI's inability to address the rapidly-changing, difficult-to-predict situation of the battlefield.

Agre's essay on a different topic is ◊51.

Two selections from Suchman's book appear below. The first, the preface "Navigation," makes the distinction between the planning and situated action perspectives. The second selection, the chapter "Interactive Artifacts," outlines a view of what interactivity means, and how the artificial intelligence version of it can be seen in a historical context. She writes of AI's project, "Interaction between people and machines implies mutual intelligibility, or shared understanding," and goes on to describe two common scenarios for this: first, the self-explanatory tool; second, the computer as an artifact having purposes. Suchman argues that both represent unsolved, perhaps irreducible, problems—as long as these words are used in the sense they have been by traditional AI. From Suchman's perspective, "intelligibility" and "understanding"—and therefore "interaction"—between people and machines must be seen as profoundly different from that between persons.

Suchman does not believe plans to be non-existent, nor does she state that AI researchers must necessarily reject plans in favor of situated actions. Rather, her argument is that it is necessary to re-conceptualize the status of plans as products of, and resources for (rather than controllers of), situated actions. See for example her "Response to Vera and Simon's 'Situated Action: A Symbolic Interpretation.'"

Much of the work in this volume, much of the best recent work in new media, recognizes rather than attempts to erase this difference. A larger classification of such work, made up of four categories, would place traditional AI's two scenarios in one of these larger categories. In these four cases, the primary intelligence that is the concern in discussing system development may be: (1) the user's own, (2) the designer's, (3) the system's (as in the traditional AI scenarios), or (4) those of communicating users. Examples of essays in this volume which focus on each of these are (1) Seymour Papert (◊28), (2) Ben Shneiderman (◊33), (3) Alan Turing (◊03), and (4) Chip Morningstar and R. Randall Farmer (◊46). Such a classification may prove to be an interesting way of considering one layer of interaction, but it can be limiting in that it considers only that layer. It does not reveal much about the levels at which interaction is seen to occur in the writings of Augusto Boal (◊22) and Jean Baudrillard (◊19), nor does it provide insight into the larger context in which this interaction takes place, as might be found by considering the arguments of Phil Agre (◊51) or Langdon Winner (◊40). Yet potential categories of interaction can be devised in other ways, e.g., based on interactive

◊003 49
◊019 277
◊022 339
◊028 413
◊033 485
◊040 587
◊046 663
◊051 737

technologies (CRT & mouse / handheld / movement tracking / voice) or on the human purpose in interacting (as discussed in the Aristotelian theory of Brenda Laurel (◊38)). Selecting the salient features from which to construct a typology is an always difficult, but potentially revealing, enterprise. Becoming too enamored of such abstractions, or identifying one type as somehow fundamental to intelligence, is always dangerous.
—NWF

Further Reading

Agre, Philip E., and David Chapman. "What Are Plans For?" *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*, ed. Pattie Maes. Cambridge: MIT Press, 1990, 17–34.

Suchman, Lucy. "Response to Vera and Simon's 'Situated Action: A Symbolic Interpretation.'" *Cognitive Science* 17 (1993): 71–75.

Original Publication

"Navigation," vii–x and "Interactive Artifacts," 5–26, *Plans and Situated Actions: The Problem of Human-Machine Communication.* Cambridge: Cambridge University Press, 1987. 2nd ed. in preparation.

From *Plans and Situated Actions*

Lucy A. Suchman

Preface: Navigation

> Thomas Gladwin (1964) has written a brilliant article contrasting the method by which the Trukese navigate the open sea, with that by which Europeans navigate. He points out that the European navigator begins with a plan—a course—which he has charted according to certain universal principles, and he carries out his voyage by relating his every move to that plan. His effort throughout his voyage is directed to remaining "on course." If unexpected events occur, he must first alter the plan, then respond accordingly. The Trukese navigator begins with an objective rather than a plan. He sets off toward the objective and responds to conditions as they arise in an *ad hoc* fashion. He utilizes information provided by the wind, the waves, the tide and current, the fauna, the stars, the clouds, the sound of the water on the side of the boat, and he steers accordingly. His effort is directed to doing whatever is necessary to reach the objective. If asked, he can point to his objective at any moment, but he cannot describe his course.
> (Berreman 1966, p. 347)

◊38
123

The subject of this book [*Plans and Situated Actions*] is the two alternative views of human intelligence and directed action represented here by the Trukese and the European navigator. The European navigator exemplifies the prevailing cognitive science model of purposeful action, for reasons that are implicit in the final sentence of the quote above. That is to say, while the Trukese navigator is hard pressed to tell us how he actually steers his course, the comparable account for the European seems to be already in hand, in the form of the very plan that is assumed to guide his actions. While the objective of the Trukese navigator is clear from the outset, his actual course is contingent on unique circumstances that he cannot anticipate in advance. The plan of the European, in contrast, is derived from universal principles of navigation, and is essentially independent of the exigencies of his particular situation.

Given these contrasting exemplars, there are at least three, quite different implications that we might draw for the study of purposeful action:

First, we might infer that there actually are different ways of acting, favored differently across cultures. How to act purposefully is learned, and subject to cultural variation. European culture favors abstract, analytic thinking, the ideal being to reason from general principles to particular instances. The Trukese, in contrast, having no such ideological commitments, learn a cumulative range of concrete, embodied responses, guided by the wisdom of memory and experience over years of actual voyages. In the pages that follow, however, I will argue that all activity, even the most analytic, is fundamentally concrete and embodied. So while there must certainly be an important relationship between ideas about action and ways of acting, this first interpretation of the navigation example stands in danger of confusing theory with practice.

Alternatively, we might posit that whether our actions are *ad hoc* or planned depends upon the nature of the activity, or our degree of expertise. So we might contrast instrumental, goal-directed activities with creative or expressive activities, or contrast novice with expert behavior. Dividing things up along these lines, however, seems in some important ways to violate our navigation example. Clearly the Truk is involved with instrumental action in getting from one island to another, and just as clearly the European navigator relies upon his chart regardless of his degree of expertise.

Finally, the position to be taken—and the one that I will adopt here—could be that, however planned, purposeful actions are inevitably *situated actions*. By situated actions I mean simply actions taken in the context of particular, concrete circumstances. In this sense one could argue that we all act like the Trukese, however much some of us may talk like Europeans. We must act like the Trukese because the circumstances of our actions are never fully anticipated and are continuously changing around us. As a consequence our actions, while systematic, are never planned in the strong sense that cognitive science would have it. Rather, plans are best viewed as a weak resource for what is primarily *ad hoc* activity. It is only when we are pressed to account for the rationality of our actions, given the biases of European culture, that we invoke the guidance of a plan. Stated in advance, plans are necessarily vague, insofar as they must accommodate the unforeseeable contingencies of particular situations. Reconstructed in retrospect, plans systematically filter out precisely the particularity of detail that characterizes situated actions, in favor of those aspects of the actions that can be seen to accord with the plan.

This third implication, it seems, is not just a symmetric alternative to the other two, but is different in kind, and somewhat more serious. That is, it calls into question not just the adequacy of our distinctions along the dimensions of culture, kinds of activity, or degrees of expertise, but the very productivity of our starting premises—that representations of action such as plans could be the basis for an account of actions in particular situations. Because the third implication has to do with foundations, and not because there is no truth in the other two, I take the idea that actions are primarily situated, and that situated actions are essentially *ad hoc*, as the starting point for my investigations.

The view of action exemplified by the European navigator is now being reified in the design of intelligent machines. I examine one such machine, as a way of uncovering the strengths and limitations of the general view that its design embodies. The view, that purposeful action is determined by plans, is deeply rooted in the Western human sciences as *the* correct model of the rational actor. The logical form of plans makes them attractive for the purpose of constructing a computational model of action, to the extent that for those fields devoted to what is now called cognitive science, the analysis and synthesis of plans effectively constitute the study of action. My own contention, however, is that as students of human action we ignore the Trukese navigator at our peril. While an account of how the European navigates may be in hand, the essential nature of action, however planned or unplanned, is situated. It behooves us, therefore, to study and to begin to find ways to describe the Trukese system.

There is an injunction in social studies of science to eschew interest in the validity of the products of science, in favor of an interest in their production. While I generally agree with this injunction, my investigation of one of the prevailing models of human action in cognitive science is admittedly and unabashedly interested. That is to say, I take it that there is a reality of human action, beyond either the cognitive scientist's models or my own accounts, to which both are trying to do justice. In that sense, I am not just examining the cognitive science model with the dispassion of the uncommitted anthropologist of science, I am examining it in light of an alternative account of human action to which I am committed, and which I attempt to clarify in the process.

Interactive Artifacts

> Marginal objects, objects with no clear place, play important roles. On the lines between categories, they draw attention to how we have drawn the lines. Sometimes in doing so they incite us to reaffirm the lines, sometimes to call them into question, stimulating different distinctions.
> (Turkle 1984, p. 31)

In *The Second Self* (1984), Sherry Turkle describes the computer as an "evocative object," one that raises new questions regarding our common sense of the distinction between artifacts and intelligent others. Her studies include an examination of the impact of computer-based artifacts on children's conceptions of the difference between categories such as "alive" versus "not alive," and "machine" versus "person." In dealing with the questions that computer-based objects evoke, children make clear that the differentiation of physical from psychological entities, which as adults we largely take for granted, is the end product of a process of establishing the relationship between the observable behavior of a thing and its underlying nature.[1] Children have a tendency, for example, to attribute life to physical objects on the basis of behavior such as autonomous motion, or reactivity, though they reserve humanity for entities evidencing such things as emotion, speech, and apparent thought or purposefulness. Turkle's observation with respect to computational artifacts is that children ascribe to them an "almost aliveness," and a psychology, while maintaining their distinctness from human beings: a view that, as Turkle points out, is remarkable among other things for its correspondence to the views held by those who are the artifacts' designers.[2]

I take as a point of departure a particular aspect of the phenomenon that Turkle identifies: namely, the apparent challenge that computational artifacts pose to the longstanding distinction between the physical and the social, in the special sense of those things that one designs, builds, and uses, on the one hand, and those things with which one communicates, on the other. While this distinction has been relatively non-problematic to date, now for the first time the term "interaction"—in a sense previously reserved for describing a uniquely interpersonal activity—seems appropriately to characterize what goes on between people and certain machines as well.[3] Interaction between people and machines implies mutual intelligibility, or shared understanding. What motivates my inquiry, therefore, is not only the recent question of how there could be mutual intelligibility between people and machines, but the prior question of how we account for the shared understanding, or mutual intelligibility, that we experience as people in our interactions with others whose essential sameness is not in question. An answer to the more recent question, theor-etically at least, presupposes an answer to the earlier one.

In this chapter I relate the idea of human-machine communication to some distinctive properties of computational artifacts, and to the emergence of disciplines dedicated to making those artifacts intelligent. I begin with a brief discussion of cognitive science, the interdisciplinary field devoted to modeling cognitive processes, and its role in the project of creating intelligent artifacts.[4] Along with a theoretical interest in intelligent artifacts, the computer's properties have inspired a practical effort at engineering interaction between people and machines. I argue that the description of computational artifacts as interactive is supported by their *reactive, linguistic,* and internally *opaque* properties. With those properties in mind, I consider the double sense in which researchers are interested in artifacts that explain themselves: on the one hand, as a solution to the longstanding problem of conveying the artifact's intended purpose to the user, through its design and attendant instructions and, on the other hand, as a means of establishing the intelligence, or rational accountability, of the artifact itself.

1 Automata and Cognitive Science

Historically the idea of *automata*—the possibility of constructing physical devices that are self-regulating in ways that we commonly associate with living, animate beings—has been closely tied to the simulation of animal forms. McCorduck (1979) points out that human-like automata have been constructed since Hellenic times: statues that moved, gestured, spoke, and generally were imbued by observers—even those well aware of the internal mechanisms that powered them—with everything from minds to souls.[5] In the fourteenth century in Western Europe, learned men were commonly believed to construct talking heads made of brass, considered as both the source of their creator's wisdom and its manifestation. More prosaically, Jacques de Vaucanson in the eighteenth century designed a series of renowned mechanical statues, the most famous being a duck, the inner workings of which produced a variety of simple outward behaviors.

At the same time, Julien de la Mettrie published *Man, A Machine,* in which he argued that the vitality characteristic of human beings was the result of their physical *structure,* rather than either of something immanent in their material substance or of some immaterial force. Cognitive scientists today maintain the basic premise of de la Mettrie with respect to mind, contending that mind is best viewed as neither substantial nor insubstantial, but as an abstractable structure implementable in any number of possible physical substrates. Intelligence, in other words, is only incidentally embodied in the neurophysiology of the human brain, and what is essential about intelligence can be abstracted from that particular, albeit highly successful, substrate and embodied in an unknown range of alternative forms. This view decouples reasoning and intelligence from things uniquely human, and opens the way for the construction of intelligent artifacts.[6]

The preoccupation of cognitive science with mind in this abstract sense is in part a concern to restore meaning to psychological explanation (see Stich 1983, ch. 1). At the turn of this century, the recognized method for studying human mental life was introspection and, insofar as introspection was not amenable to the emerging canons of scientific method, the study of cognition seemed doomed to be irremediably unscientific. In reaction to that prospect, the behaviorists posited that all human action should be understandable in terms of publicly observable, mechanistically describable relations between the organism and its environment. In the name of turning cognitive studies into a science, in other words, the study of cognition as the study of something apart from overt behavior was effectively abandoned in mainstream psychology.

Cognitive science, in this respect, was a project to bring thought back into the study of human action, while preserving the commitment to scientism. Cognitive science reclaims mentalist constructs such as beliefs, desires, intentions, symbols, ideas, schemata, planning, and problem-solving. Once again human purposes are the basis for cognitive psychology, but this time without the unconstrained speculation of the introspectionists. The study of cognition is to be empiricized not by a strict adherence to behaviorism, but by the use of a new technology: namely, the computer.

The sub-field of cognitive science most dedicated to the computer is artificial intelligence. Artificial intelligence arose as advances in computing technology were tied to developments in neurophysiological and mathematical theories of information. The requirement of computer modeling, of an "information processing psychology," seemed both to make theoretical sense and to provide the accountability that would make it possible to pursue a science of otherwise inaccessible mental phenomena. If a theory of underlying mental processes could be modeled on the computer so as to produce the right outward behavior, the theory could be viewed as having passed at least a sufficiency test of its psychological validity.

The cognitivist strategy is to interject a mental operation between environmental stimulus and behavioral response: in essence, to relocate the causes of action from the environment that impinges upon the actor to processes, abstractable as computation, in the actor's head. The first premise of cognitive science, therefore, is that people—or "cognizers" of any sort—act on the basis of symbolic representations: a kind of cognitive code, instantiated physically in the brain, on which operations are performed to produce mental states such as "the belief that *p*," which in turn produce behavior consistent with those states. The relation of environmental stimuli to those mental states, on the one hand, and of mental states to behavior, on the other, remains deeply problematic and widely debated within the field (see, for example, Fodor 1983; Pylyshyn 1974, 1984; Stich 1983). The agreement among all participants in cognitive science and its affiliated disciplines, however, is that cognition is not just potentially *like* computation, it literally *is* computational. There is no reason, in principle, why there should not be a computational account of mind, therefore, and there is no a priori reason to draw a principled boundary between people, taken as "information-processors" or "symbol manipulators" or, in George Miller's phrase, "informavores" (Pylyshyn 1984, p. xi), and certain computing machines.

The view that intelligence is the manipulation of symbols finds practical implementation both in so-called expert systems, which structure and process large amounts of well-formulated data, and industrial robots that perform routine, repetitive assembly and control tasks. Expert systems—essentially sophisticated programs that manipulate data structures to accord with rules of inference that experts are understood to use—have minimal sensory-motor, or "peripheral," access to the world in which they are embedded, input being most commonly through a keyboard, by a human operator. Industrial robots—highly specialized, computer-controlled devices designed to perform autonomously a single repetitive physical task—have relatively more developed sensory–motor apparatus than do expert systems, but the success of robotics is still confined to specialized activities, under controlled conditions. In both cases, the systems can handle large amounts of encoded information, and syntactic relationships of great sophistication and complexity, in highly circumscribed domains. But when it comes either to direct interaction with the environment, or to the exercise of practical, everyday reasoning about the significance of events in the world, there is general agreement that the state-of-the-art in intelligent machines has yet to attain the basic cognitive abilities of the normal five-year-old child.

2 The Idea of Human-Computer Interaction

In spite of the current limits on machine intelligence, the use of an intentional vocabulary is already well established in both technical and popular discussion of computers. In part, the attribution of purpose to computer-based artifacts derives from the simple fact that each action by the user effects an immediate machine *reaction* (see Turkle 1984, ch. 8). The technical definition of "interactive computing" (see, for example, Oberquelle, Kupka, and Maass 1983, p. 313) is simply that real-time control over the computing process is placed in the hands of the user, through immediate processing and through the availability of interrupt facilities whereby the user can override and modify the operations in progress. This definition contrasts current capabilities with earlier forms of computing, specifically batch processing, where user commands were queued and executed without any intermediate feedback. The greater reactivity of current computers, combined with the fact that, like any machine, the computer's reactions are not random but by design, suggest the character of the computer as a purposeful, and, by association, as a social object.

A more profound basis for the relative sociability of computer-based artifacts, however, is the fact that the means for controlling computing machines and the behavior that results are increasingly *linguistic*, rather than mechanistic. That is to say, machine operation becomes less a matter of pushing buttons or pulling levers with some physical result, and more a matter of specifying operations and assessing their effects through the use of a common language.[7] With or without machine intelligence, this fact has contributed to the tendency of designers, in describing what goes on between people and machines, to employ terms borrowed from the description of human interaction—dialogue, conversation, and so forth: terms that carry a largely unarticulated collection of intuitions about properties common to human communication and the use of computer-based machines.

While for the most part the vocabulary of human interaction has been taken over by researchers in human-machine communication with little deliberation, several researchers have attempted to clarify similarities and differences between computer use and human conversation. Perhaps the most thoughtful and comprehensive of these is Hayes and Reddy (1983). They identify the central difference between existing interactive computer systems and human communication as a question of "robustness," or the ability on the part of conversational participants to respond to unanticipated circumstances, and to detect and remedy troubles in communication:

> The ability to interact gracefully depends on a number of relatively independent skills: skills involved in parsing elliptical, fragmented, and otherwise ungrammatical input; in ensuring that communication is robust (ensuring that the intended meaning has been conveyed); in explaining abilities and limitations, actions and the motives behind them; in keeping track of the focus of attention of a dialogue; in identifying things from descriptions, even if ambiguous or unsatisfiable; and in describing things in terms appropriate for the context. While none of these components of graceful interaction has been entirely neglected in the literature, no single current system comes close to having most of the abilities and behaviours we describe, and many are not possessed by any current systems. (p. 232)

Hayes and Reddy believe, however, that:

> Even though there are currently no truly gracefully interacting systems, none of our proposed components of graceful interaction appears individually to be much beyond the current state of the art, at least for suitably restricted domains of discourse. (p. 232)

They then review the state of the art, including systems like LIFER (Hendrix 1977) and SCHOLAR (Carbonell 1971), which display sensitivity to the user's expectations regarding acknowledgement of input; systems that resolve ambiguity in English input from the user through questions (Hayes 1981); systems like the GUS system (Bobrow *et al* 1977) which represent limited knowledge of the domain that the interaction is about; work on the maintenance of a common focus over the course of the interaction (Grosz 1977; Sidner 1979); and Hayes and Reddy's own work on an automated explanation facility in a simple service domain (1983).

Two caveats on Hayes and Reddy's prescription for a gracefully interacting system (both of which, to their credit, they freely admit) are worth noting. First, they view the abilities cited as necessary but not sufficient for human interaction, their claim for the list being simply that "it provides a good working basis from which to build gracefully interacting systems" (1983, p. 233). And, not surprisingly, the abilities that they cite constitute a list of precisely those problems currently under consideration in research on human-machine communication. There is, in other words, no independent assessment of how the problems on which researchers work relate to the nature and organization of human communication as such. Secondly, research on those problems that have been identified is confined to highly circumscribed domains. The consequence of working from an admittedly partial and *ad hoc* list of abilities, in limited domains, is that practical inroads in human-computer communication can be furthered, while the basic question of what human interaction comprises is deferred. Deferred as well is the question of why it is, beyond methodological convenience, that research in human-machine interaction has proceeded only in those limited domains that it has.

Moreover, while Hayes and Reddy take the position that "it is very important for a gracefully interacting system to conduct a dialogue in as human-like a way as possible" (ibid., p. 233), this assertion is a point of controversy in the research community. On the one side, there is an argument to the effect that one should acknowledge, and even exploit, the fact that people bring to computer use a tremendous range of skills and expectations from human interaction. Within research on human-computer interaction, for example, some progress has been made toward allowing people to enter commands into computers using natural language (i.e. languages like English, in contrast to programming languages). On the other side, even Hayes and Reddy admit that:

> the aim of being as human-like as possible must be tempered by the limited potential for comprehension of any foreseeable computer system. Until a solution is found to the problems of organizing and using the range of world knowledge possessed by a human, practical systems will only be able to comprehend a small amount of input, typically within a specific domain of expertise. Graceful interaction must, therefore, supplement its simulation of human conversational ability with strategies to deal naturally and gracefully with input that is not fully understood, and, if possible, to steer a conversation back to the system's home ground. (ibid., p. 233)

While Hayes and Reddy would make these recovery strategies invisible to the user, they also acknowledge the "habitability" problem identified by Watt (1968) with respect to language: that is, the tendency of human users to assume that a computer system has sophisticated linguistic abilities after it has displayed elementary ones. This tendency is not surprising, given the fact that our only precedent for language-using entities to date has been other human beings. As soon as computational artifacts demonstrate *some* evidence of recognizably human abilities, we are inclined to endow them with the rest. The misconceptions that ensue, however, lead some like Fitter (1979) to argue that English or other "natural" languages are in fact not natural for purposes of human-computer interaction:

> for the purpose of man–computer communication, *a natural language is one that makes explicit the knowledge and processes for which the man and computer share a common understanding* . . . it becomes the responsibility of the systems designer to provide a language structure which will make apparent to the user the procedures on which it is based and will not lead him to expect from the computer unrealistic powers of inference. (ibid., p. 340, original emphasis)

In view of our tendency to ascribe full intelligence on the basis of partial evidence, the recommendation is that designers might do best to make available to the user the ways in which the system is *not* like a participant in interaction.[8] In this spirit, Nickerson (1976) argues that:

> The model that seems appropriate for this view of person–computer interaction is that of an individual making use of a sophisticated tool and not that of one person conversing with another. The term "user" is, of course, often used to denote the human component in a person–computer interaction, as it has been in this paper. It is, to my taste, preferable to the term "partner," not only because it seems more descriptive of the nature of the relationships that existing systems permit, and that future systems are likely to, but because it implies an asymmetry with respect to goals and objectives that "partner" does not. "User" is not a term that one would normally apply to a participant in a conversation. (p. 111)

The argument that computational processes should be revealed to the user, however, is potentially counter to the promotion of an intentional vocabulary in speaking about computer-based devices. As Dennett (1978) points out, it is in part our inability to see inside each other's heads, or our mutual *opacity*, that makes intentional explanations so powerful in the interpretation of human action. So it is in part the internal complexity and opacity of the computer that invites an intentional stance. This is the case not only because users lack technical knowledge of the computer's internal workings but because, even for those who possess such knowledge, there is an "irreducibility" to the computer as an object that is unique among human artifacts (Turkle 1984, p. 272). The overall behavior of the computer is not describable, that is to say, with reference to any of the simple local events that it comprises; it is precisely the behavior of a myriad of those events in combination that constitutes the overall machine. To refer to the behavior of the machine, then, one must speak of "its" functionality. And once reified as an entity, the inclination to ascribe actions to the entity rather than to the parts is irresistible.

Intentional explanations relieve us of the burden of understanding mechanism, insofar as one need assume only that the design is rational in order to call upon the full power of common-sense psychology and have, ready at hand, a basis for anticipating and construing an artifact's behavior. At the same time, precisely because the mechanism is in fact unknown, and, insofar as underspecification is taken to be characteristic of human beings (as evidenced by the fact that we are inclined to view something that is fully specified as less than human), the personification of the machine is reinforced by the ways in which its inner workings are a mystery, and its behavior at times surprises us. Insofar as the machine is somewhat predictable, in sum, and yet is also both internally opaque and liable to unanticipated behavior, we are more likely to view ourselves as engaged in interaction with it than as just performing operations upon it, or using it as a tool to perform operations upon the world (see MacKay 1962).

3 Self-Explanatory Artifacts

In the preceding pages I have proposed that the reactive, linguistic, and opaque properties of the computer lead us to view it as interactive, and to apply intentional explanations to its behavior. This tie to intentionality has both theoretical and practical implications. Practically, it suggests that, like a human actor, the computer should be able to explain itself, or the intent behind its actions, to the user. Theoretically, it suggests that the computer actually has intent, as demonstrated precisely in this ability to behave in an accountably rational, intelligible way.

For practical purposes, "user interface" designers[9] have long held the view that machines ideally should be self-explanatory, in the broad sense that their operation should be discoverable without extensive training, from information provided on or through the machine itself. On this view, the degree to which an artifact is self-explanatory is just the extent to which someone examining the artifact is able to reconstruct the *designer's intentions* regarding its use. This basic idea, that a self-explanatory artifact is one whose intended purpose is discoverable by the user, is presumably as old as the design and use of tools. With respect to computer-based artifacts, however, the notion of a self-explanatory artifact has taken on a second sense: namely, the idea that the artifact might actually *explain itself* in something more like the sense that a human being does. In this second sense the goal is that the artifact should not only be intelligible to the user as a tool, but that it should be *intelligent*—that is, able to understand the actions of the user, and to provide for the rationality of its own.

In the remainder of this chapter, I look at these two senses of a self-explanatory machine and at the relation between them. The first sense—that a tool should be decipherable by its user—reflects the fact that artifacts are constructed by designers, for a purpose, and that the user of a tool needs to know something of that design intent. Given their interactional properties, computational tools seem to offer unique capabilities for the provision of instruction to their users. The idea that instructions could be presented more effectively using the power of computation is not far from the idea that computer-based artifacts could actually instruct: that is, could interact with people in a way that approximates the behavior of an intelligent human expert or coach. And this second idea, that the artifact could actually interact instructively with the user, ties the practical problem of instruction to the theoretical problem of building an intelligent, interactive machine.

3.1 The Computer as an Artifact Designed for a Purpose

At the same time that computational artifacts introduce new complexity and opacity into our encounters with machines, our reliance on computer-based technology and its proliferation throughout the society increases. One result is the somewhat paradoxical objective that increasingly complex technology should be usable with decreasing amounts of training. Rather than relying upon the teachings of an experienced user, the use of computers is to be conveyed directly through the technology itself.

The inherent difficulty of conveying the use of a technology directly through its design is well known to archaeologists, who have learned that while the attribution of design intent is a requirement for an artifact's intelligibility, the artifact's design as such does not convey unequivocally either its actual or its intended use. While this problem in construing the purpose of artifacts can be alleviated, it can never fully be resolved, and it defines the essential problem that the novice user of the tool confronts. Insofar as the goal of a tool's design is that use of the tool should be self-evident, therefore, the problem of deciphering an artifact defines the problem of the designer as well.

As with any communication, instructions for the use of a tool are constrained by the general maxim that utterances should be designed for their recipients. The extent to which the maxim is observed is limited in the first instance by the resources that the medium of communication affords. Face-to-face human interaction is the paradigm case of a system for communication that, because it is organized for maximum context-sensitivity, supports a response designed for just these recipients, on just this occasion. Face-to-face instruction brings that context-sensitivity to bear on problems of skill acquisition. The gifted coach, for example, draws on powers of language and observation, and uses the situation of instruction, in order to specialize instruction for the individual student. Where written instruction relies upon generalizations about its recipient and the occasion of its use, the coach draws pedagogical strength from exploitation of the unique details of particular situations.[10]

A consequence of the human coach's method is that his or her skills must be deployed anew each time. An instruction manual, in contrast, has the advantage of being durable, re-usable, and replicable. In part, the strength of written text is that, in direct contrast to the pointed commentary of the coach, text allows the *disassociation* of the occasion of an instruction's production from the occasion of its use. For the same reason, however, text affords relatively poor resources for recipient design. The promise of interactive computer systems, in these terms, is a technology that can move instructional design away from the written manual in the direction of the human coach, and the resources afforded by face-to-face interaction.

Efforts at building self-explicating machines in their more sophisticated forms now adopt the metaphor of the machine as an expert, and the user as a novice, or student. Among the most interesting attempts to design such a computer-based "coach" is a system called WEST (Burton and Brown 1982). The design strategy adopted in WEST is based on the observation that the skill of a human coach lies as much in what isn't said as what is. Specifically, the human coach does not disrupt the student's engagement in an activity in order to ask questions, but instead diagnoses a student's strengths and weaknesses through observation. And once the diagnosis is made, the coach interjects advice and instruction selectively, in ways designed to maximize learning through discovery and experience. In that spirit, the WEST system attempts to infer the student's knowledge of the domain—in this case a computer game called "How the West Was Won," designed to teach the use of basic arithmetic expressions—by observing the student's behavior.[11]

While the project of identifying a student's problems directly from his or her behavior proved considerably more difficult than expected, the objectives for the WEST coach were accomplished in the prototype system to an impressive degree. Because in the case of learning to play WEST the student's actions take the form of input to the computer (entries on a keyboard) and therefore leave an accessible trace, and because a context for those actions (the current state of, and history of consecutive moves across, the "board") is defined by the system, each student turn can be compared against calculations of the move that a hypothetical expert player would make given the same conditions. Each expert move, in turn, requires a stipulated set of associated skills. Evidence that a particular skill is lacking, accumulated across some number of moves, identifies that skill as a candidate for coaching. The coach then interjects offers of advice to the student at opportune moments in the course of the play, where what constitutes an opportune moment for interjection is determined according to a set of rules of thumb regarding good tutorial strategy (for example, always coach by offering the student an alternative move that both demonstrates the relevant skill and accomplishes obviously superior results; never coach on two turns in a row, no matter what, and so forth).

3.2 The Computer as an Artifact Having Purposes

While the computer-based coach can be understood as a logical development in the longstanding problem of instruction, the requirement that it be interactive introduces a second sense of self-explanatory machine which is more recent, and is uniquely tied to the advent of computing. The new idea is that the intelligibility of artifacts is not just a matter of the availability to the user of the *designer's* intentions for the artifact, but of the intentions of the *artifact* itself. That is to say, the designer's objective now is to imbue the machine with the grounds for behaving in ways that are accountably rational: that is, reasonable or intelligible to others, including, in the case of interaction, ways that are responsive to the other's actions.

In 1950, A. M. Turing proposed a now-famous, and still controversial, test for machine intelligence based on a view of intelligence as accountable rationality. Turing argued that if a machine could be made to respond to questions in such a way that a person asking the questions could not distinguish between the machine and another human being, the machine would have to be described as intelligent. To implement his test, Turing chose a game called the "imitation game." The game was initially conceived as a test of the ability of an interrogator to distinguish which of two respondents was a man and which a woman. To eliminate the evidence of physical embodiment, the interaction was to be conducted remotely, via a teleprinter. Thus Turing's notion that the game could easily be adapted to a test of machine intelligence, by substituting the machine for one of the two human respondents.

Turing expressly dismissed as a possible objection to his proposed test the contention that, although the machine might succeed in the game, it could succeed through means that bear no resemblance to human thought. Turing's contention was precisely that success at performing the game, regardless of mechanism, is sufficient evidence for intelligence (Turing 1950, p. 435). The Turing test thereby became the canonical form of the argument that if two information-processors, subject to the same input stimuli, produce indistinguishable output behavior, then, regardless of the identity of their internal operations, one processor is essentially equivalent to the other.

The lines of the controversy raised by the Turing test were drawn over a family of programs developed by Joseph Weizenbaum in the 1960s under the name ELIZA, designed to support "natural language conversation" with a computer (Weizenbaum 1983, p. 23). Of the name ELIZA, Wiezenbaum writes:

> Its name was chosen to emphasize that it may be incrementally improved by its users, since its language abilities may be continually improved by a "teacher." Like the Eliza of *Pygmalion* fame, it can be made to appear even more civilized, the relation of appearance to reality, however, remaining in the domain of the playwright. (p. 23)

Anecdotal reports of occasions on which people approached the teletype to one of the ELIZA programs and, believing it to be connected to a colleague, engaged in some amount of "interaction" without detecting the true nature of their respondent, led many to believe that Weizenbaum's program had passed a simple form of the Turing test. Notwithstanding its apparent interactional success, however, Weizenbaum himself denied the intelligence of the program, on the basis of the underlying mechanism which he described as "a mere collection of procedures" (p. 23):

> The gross procedure of the program is quite simple; the text [written by the human participant] is read and inspected for the presence of a *keyword* If such a word is found, the sentence is transformed according to a *rule* associated with the keyword, if not a content-free remark or, under certain conditions, an earlier transformation is retrieved. The text so computed or retrieved is then printed out. (p. 24, original emphasis)

In spite of Weizenbaum's disclaimers with respect to their intelligence, the ELIZA programs are still cited as instances of successful interaction between human and machine. The grounds for their success are clearest in DOCTOR, one of the ELIZA programs whose script equipped it to respond to the human user as if the computer were a Rogerian therapist and the user a patient. The DOCTOR program exploited the maxim that shared premises can remain unspoken: that the less we say in conversation, the more what is said is assumed to be self-evident in its meaning and implications (see Coulter 1979, ch. 5). Conversely, the very fact that a comment is made without elaboration implies that such shared background assumptions exist. The more elaboration or justification is provided, the less the appearance of transparence or self-evidence. The less elaboration there is, the more the recipient will take it that the meaning of what is provided should be obvious.

The design of the DOCTOR program, in other words, exploited the natural inclination of people to deploy what Karl Mannheim first termed the *documentary method of interpretation* to find the sense of actions that are assumed to be purposeful or meaningful (Garfinkel 1967, p. 78). Very simply, the documentary method refers to the observation that people take appearances as evidence for, or the document of, an ascribed underlying reality, while taking the reality so ascribed as a resource for the interpretation of the appearance. In the case of DOCTOR, computer-generated responses that might otherwise seem odd were rationalized by users on the grounds that there must be some psychiatric intent behind them, not immediately obvious to the user as "patient," but sensible nonetheless:

> If, for example, one were to tell a psychiatrist "I went for a long boat ride" and he responded "Tell me about boats," one would not assume that he knew nothing about boats, but that he had some purpose in so directing the subsequent conversation. It is important to note that this assumption is one made by the speaker. Whether it is realistic or not is an altogether different question. In any case, it has a crucial psychological utility in that it serves the speaker to maintain his sense of being heard and understood. The speaker further defends his impression (which even in real life may be illusory) by attributing to his conversational partner all sorts of background knowledge, insights and reasoning ability. But again, these are the speaker's contribution to the conversation. They manifest themselves inferentially in the *interpretations* he makes of the offered response. (Weizenbaum 1983, p. 26, original emphasis)

In explicating the ELIZA programs, Weizenbaum was primarily concerned with the inclination of human users to find sense in the computer's output, and to ascribe to it an understanding, and therefore an authority, unwarranted by the actual mechanism.[12] While unmasking the intelligence of his program, however, Weizenbaum continued to describe it as "a program which makes natural language conversation with a computer possible" (1983, p. 23). Nevertheless, as part of his disclaimer regarding its intelligence, Weizenbaum points to a crucial shortcoming in the ELIZA strategy with respect to conversation:

> ELIZA in its use so far has had as one of its principal objectives the concealment of its lack of understanding. But to encourage its conversational partner to offer inputs from which it can select remedial information, it must *reveal* its misunderstanding. A switch of objectives from the concealment to the revelation of misunderstanding is seen as a precondition to making an ELIZA-like program the basis for an effective natural language man-machine communication system. (p. 27, original emphasis)

More recently, the inevitability of troubles in communication, and the importance of their remedy to the accomplishment of "graceful interaction," has been reintroduced into the human-machine communication effort by Hayes and Reddy (1983). They observe that:

> During the course of a conversation, it is not uncommon for people to misunderstand or fail to understand each other. Such failures in communication do not usually cause the conversation to break down; rather, the participants are able to resolve the difficulty, usually by a short clarifying sub-dialogue, and continue with the conversation from where they left off. Current computer systems are unable to take part in such clarifying dialogues, or resolve communication difficulties in any other way. As a result, when such difficulties occur, a computer dialogue system is unable to keep up its end of the conversation, and a complete breakdown is likely to result; this fragility lies in stark and unfavourable contrast to the robustness of human dialogue. (p. 234)

Hayes and Reddy go on to recommend steps toward a remedy for the fragility of human-computer interaction, based on the incorporation, from human communication, of conventions for the detection and repair of misunderstanding. They acknowledge, however, that their recommendations are unlikely to be sufficient for successful communication in other than the simplest encounters, e.g., automated directory assistance, or reservation systems. The question of why this should be so—of the nature of the limits on human-machine communication, and the nature and extent of robustness in human interaction—is the subject of the following chapters [of *Plans and Situated Actions*].

Notes

1. Though see Carey 1985, chapter 1 for a critique of the Piagetian notion that children at first have no concept for mechanical causation apart from intentional causation.

2. See especially pp. 62–3; Turkle finds some cause for alarm in the fact that for children the distinction of machine and person seems to turn centrally on a separation of thought from feeling; that is, computers exhibit the former, but lack the latter. This view, she argues, includes a kind of dissociation of intellect and emotion, and consequent trivialization of both, that characterizes the attitudes of many in the field of Artificial Intelligence.

3. Actually, the term "interaction" has its origins in the physical sciences, to describe a reciprocal action or influence. I use it here in the common sense assigned to it by social science: namely, to mean communication between persons. The migration of the term from the physical sciences to the social, and now back to some ground that stands between them, relates in intriguing ways to a general blurring of the distinction between physical and social in modern science, and to the general question of whether machines are actually becoming more like people or whether, in fact, people are coming to define themselves more as machines. There is clearly a mutual influence at work. For more on this last point, see Dreyfus 1979, ch. 9.

4. For an extensive treatment, see Gardner 1985.

5. See McCorduck 1979, ch. 1; Churchland 1984, ch. 6. For a further history of automata, see Cohen 1966.

6. See Turkle 1984, ch. 7; and McCorduck 1979, ch. 5. Turkle's description of the present academic AI culture at MIT is particularly insightful.

7. Notwithstanding the popular fantasy of the talking machine, the crucial element that invites a view of computers as interactive is language, not speech. While strictly speaking buttons and keys remain the principal input devices in computing, this is relatively trivial. The synthesis of speech by computers may well add to our inclination to ascribe understanding to them, but will not, in itself, contribute substantively to their sensibility. On the other hand, simulation of natural language understanding, even when the language is written rather than spoken, is proving to be a profoundly difficult problem that is inseparable from the problem of simulating intelligence as such.

8. In fact, Nickerson (1976) points out that there are some ways in which a computer is not like another person which lend a certain advantage to the user, e.g. interruptions can be made without concern about giving offense, responses can be delayed as long as is necessary.

9. In design parlance, the term "user interface" refers both to the physical place at which the user issues commands to a device, finds reports of its state, or obtains the products of its operation, and the procedures by which those interactions occur.

10. Face-to-face interaction is in most cases a necessary, but of course never a sufficient, condition for successful human coaching. Coombs and Alty (1984) provide an interesting discussion of the failings of interactions between human advisors and new computer users. At the same time, they point out that the characteristics of the advisory sessions that new users found unsatisfactory show marked similarities to human interactions with most rule-based computer help systems, e.g. that the advisors provide only the recommended solutions to reported problems, while failing either to elicit the view of the user, or to articulate any of their own rationale. Satisfactory sessions, in contrast, were characterized by what initially appeared to be less structure and less economy, but which on further investigation was revealed as "well-motivated despite surface appearances, the objective not being strict problem-solving as we had assumed, but problem-solving through mutual understanding. This required sensitivity to different structural factors" (pp. 24–5).

11. The student is presented with a graphic display of a game board made up of 70 squares (representing the Western frontier), a pair of icons (representing the two players—user and computer), and three spinners. A player's task in each turn is to combine the three numbers that the spinners provide, using the basic operations, to produce a value that becomes the number of spaces the icon is moved along the board. To add an element of strategy, squares on the board are more and less desirable—for example, "towns" occur every ten spaces, and landing on one advances you to the next. The object is to be the first player to land on 70.

Early observation of students playing the game revealed that they were not gaining the full benefit of the arithmetic practice, in that they tended to settle on a method for combining numbers (for example, multiply the first two numbers and add the third), and to repeat that same methods at each turn. Recognizing that this might reflect either a weakness in the student's proficiency at constructing expressions, a failure to grasp the strategy of the game, or both, Brown and Burton saw the potential usefulness of a "coach" that could guide the student to an expanded repertoire of skills and a better understanding of the domain. For a description of a similarly motivated "advisory" system for the programming language PROLOG, see Coombs and Alty 1984.

12. In this regard it is interesting to note that a great debate ensued surrounding the status of the DOCTOR program as a psychotherapeutic tool. That debate took on a humorous tone when Weizenbaum submitted a letter to the Forum of the Association for Computing Machinery, an excerpt from which follows:

Below is a listing of a pl/1 program that causes a typewriter console to imitate the verbal behavior of an autistic patient. The "doctor" types his interrogatories on the console. It responds exactly as does an autistic patient—that is, not at all. I have validated this model following the procedure first used in commercial advertising by Carter's Little Liver Pills ("Seven New York doctors say . . .") and later used so brilliantly by Dr K. M. Colby in his simulation of paranoia [a reference to Colby. K. M. *et al.* 1972]; I gave N psychiatrists access to my program and asked each to say from what mental disorder it suffered. M psychiatrists ($M < N$) said the (expletive deleted) program was autistic. (The methodological assumption here is that if two processes have identical input/output behaviors, then one constitutes an explanation of the other.)

The program has the advantage that it can be implemented on a plain typewriter not connected to a computer at all. (Weizenbaum 1983, p. 28)

References

Berreman, G. 1966. Anemic and emetic analyses in social anthropology. *American Anthropologist* 68(2)1:346–54.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. 1977. GUS: a frame-driven dialogue system. *Artificial Intelligence* 8:155–73.

Burton, R. and Brown, J. S. 1982. An investigation of computer coaching for informal learning activities. In *Intelligent Tutoring Systems,* D. Sleeman and J. S. Brown, eds. London: Academic Press.

Carbonell, J. R. 1971. *Mixed-Initiative Man–Computer Dialogues.* Technical Report 1970, Bolt Beranek and Newman, Inc., Cambridge, MA.

Carey, S. 1985. *Conceptual Change in Childhood.* Cambridge, MA: MIT Press.

Churchland, P. 1984. *Matter and Consciousness.* Cambridge, MA: MIT Press.

Cohen, J. 1966. *Human Robots in Myth and Science.* London: Allen and Unwin.

Colby, K. M. et al. 1972. Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence* 3:199–221.

Coombs, M. and Alty, J. 1984. Expert systems: an alternative paradigm. *International Journal of Man–Machine Studies* 20:21–43.

Coulter, J. 1979. *The Social Construction of Mind.* Totowa, NJ: Rowman and Littlefield.

———. 1983. *Rethinking Cognitive Theory.* New York, NY: St. Martin's Press.

Dennett, D. 1978. *Brainstorms.* Cambridge, MA: MIT Press.

Dreyfus, H. 1979. *What Computers Can't Do: The Limits of Artificial Intelligence,* revised edition. New York, NY: Harper and Row.

———, ed. 1982. *Husserl Intentionality and Cognitive Science.* Cambridge, MA: MIT Press.

Fitter, M. 1979. Towards more "natural" interactive systems. *International Journal of Man–Machine Studies* 11:339–49.

Fodor, J. 1983. *The Modularity of Mind.* Cambridge, MA: MIT Press.

Gardner, H. 1985. *The Mind's New Science.* New York: Basic Books.

Garfinkel, H. 1967. *Studies in Ethnomethodology.* Englewood Cliffs, NJ: Prentice-Hall.

Gladwin, T. 1964. Culture and logical process. In *Explorations in Cultural Anthropology: Essays Presented to George Peter Murdock,* W. Goodenough, ed. New York, NY: McGraw-Hill.

Grosz, B. 1981. Focusing and description in natural language dialogues. In *Elements of Discourse Understanding,* Joshi, A., Webber, B., and Sag, I., eds. Cambridge University Press.

Hayes, P. 1981. A construction-specific approach to focused interaction in flexible parsing. *Proceedings of Nineteenth Annual Meeting of the Association for Computational Linguistics,* pp. 149–52. Stanford, CA: Stanford University.

Hayes, P. and Reddy, D. R. 1983. Steps toward graceful interaction in spoken and written man–machine communication. *International Journal of Man–Machine Studies* 19:231–84.

Hendrix, G. G. 1977. Human engineering for applied natural language processing. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence,* pp. 183–91. Cambridge MA: MIT.

McCorduck, P. 1979. *Machines Who Think.* San Francisco, CA: W. H. Freeman.

MacKay, D. M. 1962. The use of behavioral language to refer to mechanical processes. *British Journal of Philosophical Science,* 13:89–103.

Nickerson, R. 1976. On conversational interaction with computers. In *Proceedings of ACM/SIGGRAPH workshop,* October 14–15, pp. 101–13. Pittsburgh, PA.

Oberquelle, H., Kupka, I., and Maass, S. 1983. A view of human-machine communication and cooperation. *International Journal of Man–Machine Studies* 19:309–33.

Pylyshyn, Z. 1974. Minds, machines and phenomenology: some reflections on Dreyfus' What Computers Can't Do. *Cognition* 3:57–77.

———. 1984. *Computation and Cognition.* Cambridge, MA: MIT Press.

Sidner, C. L. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse.* Technical Report TR-537, MIT AI Laboratory. Cambridge, MA.

Stich, S. 1983. *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236): 433–61.

Turkle, S. 1984. *The Second Self.* New York, NY: Simon and Schuster.

Turner, R. 1962. Words, utterances and activities. In *Ethnomethodology: Selected readings,* ed. Turner. Harmondsworth, Middlesex: Penguin.

Watt, W. C. 1968. Habitability. *American Documentation* 19(3):338–51.

Weizenbaum, J. 1983. ELIZA: a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 25th Anniversary Issue,* 26(1):23–7. (Reprinted from Communications of the ACM, 29(1):36–45, January 1966.)