*Understanding Computers and Cognition* is an extremely important book, not because it provides a ready-made set of answers, but because it raises hard questions about the current premises and practices of cognitive science and computer design. It is not a textbook, but a book to argue with and talk about. If the discussions reduce to complaints about the particular characterizations offered, the book's value to the community will be minimized. But if the discussions are about not only this book, but what this book is about, then we may in fact gain new orientations toward cognitive science and computer design. And if Winograd and Flores manage to stir up those winds of change, they will have made a contribution indeed.

ACKNOWLEDGMENT

Thanks to Susan Newman, Debbie Tatar, and Randy Trigg of the Intelligent Systems Lab, Xerox PARC, for helping to clarify these reactions.

REFERENCES

1. Atkinson, J.M. and Drew, P., *Order in Court: The Organization of Verbal Interaction in Judicial Settings* (Macmillan, London, 1979).
2. Garfinkel, H., *Studies in Ethnomethodology* (Prentice-Hall, Englewood Cliffs, 1967).
3. Heritage, J., *Garfinkel and Ethnomethodology* (Polity Press, Cambridge, U.K., 1984).
4. Lynch, M., *Art and Artefacts in Laboratory Science* (Routledge & Kegan Paul, London, 1985).
5. Suchman, L., *Plans and Situated Actions: The Problem of Human-Machine Communication* (Cambridge University Press, Cambridge, U.K., 1987).
6. Zimmerman, D., The practicalities of rule use, in: J. Douglas (Ed.), *Understanding Every Life* (Routledge & Kegan Paul, London, 1971).

**T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design* (Ablex, Norwood, NJ, 1986); 207 pages, $24.95.**

Reviewed by: William J. Clancey
*Stanford Knowledge Systems Laboratory, Palo Alto, CA 94304, U.S.A.*

---

## 1. Introduction

> Every triumphant theory passes through three stages: first it is dismissed as untrue; then it is rejected as contrary to religion; finally, it is accepted as dogma and each scientist claims that he had long appreciated the truth. (Gould [6] quoting embryologist von Baer)

AI researchers and cognitive scientists commonly believe that thinking involves manipulating representations. For example, when we speak our thoughts are translated into words. We don't know what the mental operations are, but we assume that they are analogous to computer models of reasoning. There are hierarchical networks in the brain and stored associational links between concepts. There are propositions, implication rules, control processes, and so on. Thinking involves search, inference, and making choices. This is how we model reasoning, and what goes on in the brain is similar.

Winograd and Flores present a radically different view. In a nutshell, intelligence of the kind exhibited by people isn't possible by manipulating representations alone. In fact, they claim that our knowledge is not represented in the brain at all, at least not as stored facts and procedures. Many readers will reject this argument as obviously wrong, so obviously wrong they won't have to read the book to be convinced that it is just some variant of an Eastern religion or simply anti-scientific and not worth their time.

After reading the book twice and much consideration, I believe that Winograd and Flores are mostly right. We have the stuff here of Copernicus, Darwin, and Freud: At its heart, the human world is not what we thought. However, I believe that Winograd and Flores significantly understate the role of representation in mediating intelligent behavior, specifically in the process of reflection, when representations are generated prior to physical action. Furthermore, while the book convincingly describes the limitations of formal reasoning in the extreme, the practical extent of what can be accomplished is uncertain.

In understanding a book like this, it is useful to start with the problem that the authors are trying to address, that is, what they believe needs to be fixed. Winograd and Flores object to how computers are described in the popular literature, how AI researchers talk about intelligence, and the kinds of programs AI researchers, particularly in the area of natural language, are trying to develop. Winograd and Flores reject the commonly accepted beliefs that expert systems or any program could be intelligent, that representations can be used to model intelligent behavior, and that developing autonomous agents is an effective use for computers.

The book is based on the idea that understanding the nature of human cognition and what computers can do will enable us to use them more effectively. While many examples are given to illustrate limitations of current computer programs and to raise new possibilities, it is important to keep in mind that the authors have little interest in establishing what lies within practical bounds, that is, what the representational paradigm will allow. Rather they are trying to define the limit of what is possible. This makes the book of strong theoretical interest, but the practical implications, for example, what expert systems might ultimately be able to do not clear. The authors' philosophical stance places more value "in asking meaningful questions—ones

that evoke an openness to new ways of being" not "in finding the 'right answers'" (p. 13). Without a doubt, this book raises good questions.

Any attempt to summarize the arguments of this book in a few paragraphs is sure to raise many more questions than it answers. I will only present a few of the important terms and describe the general structure of the argument. In subsequent sections, I describe what I learned by reading the book and the problems that I perceive.

In general, Winograd and Flores approach cognition and computation in terms of what it means to "understand language in the way people do." Their analysis leads them to conclude that computers cannot understand natural language—not just now, never. This is because all programs—all representations, abstractions and primitives alike—are based on preselected objects and properties. The background that motivates representations, the experience behind the designer's analysis, has been cut out. Thus, when breakdowns occur, that is, when an inability to cope occurs because the demands for action placed upon the program are different from expected, there will be no basis for moving beyond the initial formalization. Yes, the designer can anticipate typical breakdowns and provide for representational change, but these will themselves be limited and prone to breakdown. The only way out is to generate new representations from outside the representational realm.

The key to this argument is that new representations spring from a shared, *unformalized background*. Coping with a breakdown involves articulating the basis of a representation. If you don't have this background, you can't speak with commitment, that is, with an implicit promise to clarify your meaning if questioned. Since you can't negotiate meaning, you can't engage in language.

According to Winograd and Flores, the view that we codify and store experiences in representations that exist is the brain is naive. Rather, representation is a post-hoc interpretation of history. What we articulate has meaning within a context, and what we say has been shaped historically by that context. But it is only formalized (represented) when we speak. We are not translating what we have already formalized.

The question naturally arises, just what are we storing in the brain? Perhaps we do not *store* anything? What is memory anyway? What are experiences? Surely we retain something. But perhaps we are not carrying around things in our heads? Consider how much we take for granted, in particular how our conception of objects and space shapes our conception of the mind, and how little we understand.

This book is anti-illusion, not anti-technology. It is not about what computers can't do (see [3]) or shouldn't do (see [12]), but what they can do and how we should use them. It is primarily a positive statement, an insider's attempt to articulate AI practice, to understand what an AI researcher is doing when he writes a program. The goal is to understand how programs relate to life, what they capture of our nature, and what they leave out.

While the stated objective of the book is "how to design computer tools suited to human use and purposes," the authors are most interested in understanding what it means to be human (p. 12). They believe that the rationalistic tradition, based on ideas such as internal representation, search, and choice among a set of alternatives, must be replaced if we are to understand human thought.

## 2. What the book is like

*Understanding Computers and Cognition* is intelligent, measured, and instructive. It deliberately avoids "philosophic scholarship" in order to focus on central points critical to developing a new understanding. In four introductory chapters, the authors describe the rationalistic tradition, hermeneutics, consensual domains, and speech act theory. The discussion is admirably crisp. In just fifty pages, the book relates subtle, unfamiliar philosophical, biological, and linguistic ideas to what AI researchers do everyday as programmers.

While a cursory scan shows the book to be full of jargon—thrownness, readiness-to-hand, shared background, blindness, breakdown, commitment—these words turn out to be useful for retaining the message. Like Freud's jargon (e.g., ego, subconscious), these terms introduce a new language for thinking about familiar things (p. 40).

The book is also remarkable for sharp, definite statements that seem so contrary to common belief: "One cannot construct machines that either exhibit or successfully model intelligent behavior." Amazingly, this comes from someone who gave us another book entitled, *Understanding Natural Language*. A "born again" conviction might lie behind the book's bold remarks.

An AI researcher who reads only the section on expert systems and does not take the time to seriously study the arguments is likely to be greatly alarmed by the inflammatory language: "Calling it intelligent might be useful for those trying to get research funding. . . ." "The designers themselves are blind to the limits. . ." (p. 93). In a few places the polemic becomes obscure and is easily dismissed: "World as the background of obviousness is manifest in our everyday dealings as the familiarity that pervades our situation, and every possible utterance presupposes this" (p. 58). But with rare exception, the jargon and original point of view combine in clear and thought-provoking observations: "In trying to understand a tradition, the first thing we must become aware of is how it is concealed by its obviousness" (p. 7). Almost every page has an idea worth underlining.

Sometimes the book has a poetic, mystical tone: ". . . [We] present the main points, listening for their relevance to our own concerns." The authors evoke reverence for their ideas, reflected by the book's final sentence: "The transformation we are concerned with is not a technical one, but a continuing evolution of how we understand our surroundings and ourselves—of how we continue

becoming the beings that we are" (p. 179). This is a book of religious philosophy, an inquiry into the origin of beliefs, values, and practices, of their nature, why they work, why they fail, and how they change. To quickly dismiss this book on technical grounds is to miss much more than the authors' conclusions about cognition.

If you are committed to understanding the book, I encourage reading the introduction (Chapter 1), then skipping around in whatever order appeals to you. Chapters 6 and 8 are excellent reviews; you could start there and then go forwards for the detailed discussion. Or if you prefer to start with the familiar, work backwards from the final chapter (which gives a good example of program design) to Chapter 10 (on current directions in AI) and the discussion in Chapters 7, 8 and 9 (on representation and language). However, I think that the book must be read completely to be accepted, and I found that reading it twice, separated by more than a year, was valuable.

### 3. Important ideas

Well into my second reading, I realized that somebody was wrong in a big way and I kept stopping in amazement at the possibilities. Could it be that Winograd and Flores are mostly wrong? What a grand embarrassment! From the other perspective, it's equally amazing how many researchers will respond with staunch certainty, "That's not right. Computers will do whatever we program them to do." Now I've accepted the main argument and have settled down to musing over details. I still think people will just shake their heads and go about their business.

On first reading, the idea that seemed most important was that the computer should be thought of as a medium for communication, rather than an autonomous agent. A computer does not understand, it is exhibiting my commitments remotely. It is not the computer that makes requests or promises, but the programmer. The computer shows my patterns, my associations, my preferences.

This view increases my sense of responsibility and gratification. It is my work after all, not somebody else printing things on the screen. This also leads to an interesting question: How should I project myself? What should I put on the screen to reflect my choices? But after a year, I don't think the "computer as a medium" idea has changed what I do, just my theoretical understanding of what I am doing. After all, I always felt embarrassed or proud about my programs. I always knew that the computer was just showing my own constructions (or that of fellow programmers).

On second reading, a completely different message hit me. I realized that I don't have any patterns, associations, or preferences stored in my mind. This is a somewhat depressing and confusing thought. As Winograd and Flores indicate, the implications are more than technical, relating to our image of what a person is, and this can influence our response to the argument. But the

new conception is useful, and after awhile it starts to make sense. To illustrate this, in the following sections I summarize the ideas that I find particularly exciting by quoting from and paraphrasing the book. I amplify these ideas by discussing other connections and research implications.

### 3.1. *All behavior proceeds from the subconscious*

> To exist historically means that knowledge of oneself can never be complete (p. 33).

Language is necessarily blind to its context because it involves a formalization based on the *historical* structure of interactions. "As carriers of a tradition, we cannot be objective observers of it. Continuing work to revealing a tradition is at the same time a source of concealment" (p. 179). Language crystallizes what we are, but it is always partial, biased, and momentary. The power of language is to articulate recurrence, to identify patterns, to claim structure, to explain. But it is always post-hoc and apart from our being.

While these ideas may sound strange, most people are familiar with the idea that some behavior (at least) proceeds subconsciously, that is, not from articulated beliefs. This is the Freudian view of the subconscious: We act without knowing our own motivations. We do not always act rationally, by choice. Winograd and Flores take this to the extreme: All behavior is direct, without intervening representation.

In the popular understanding of psychiatric problems, subconsciously-directed behavior is associated with illness. We associate the subconscious with unusual, unhealthy behavior because the subconscious only becomes important to us when a breakdown occurs: a failure of a commitment, a violated expectation, a frustrated desire. When we perceive that we are apart from the world or when our actions are confused, this is when we use language to articulate what lies behind our behavior and our discomfort. We try to spell out what is not obvious, the assumed background that is affecting our behavior or our emotions.

In the "talking cure" of psychotherapy, a person articulates the recurrent structure in his behavior, naming situations and responses to them. Thus, he may become aware of the structural coupling in his life, allowing a new interpretation and new behavior. In this way, Freudian psychology has been reinterpreted as a form of hermeneutics: "The mental self is a story whose meaning is only interpretable in my life's history" [13, p. 276].

### 3.2. *We are always already interpreting*

In one of the most powerful ideas in the book, Winograd and Flores tell us that we are always attending, always selecting. To understand the value of this conception, consider the problem of explaining how we happen to attend to something. Suppose that I walk into a museum and see something interesting

and walk over to study it. How did I know that it was interesting? What little clue made me decide to attend to it and to realize that it was interesting? What littler clue made me notice the first clue? Maybe the frame showed me where to look, but when did I decide to decide to notice that frame? In the museum I am always attending, always making interpretations. I am not matching preconceptions and recognizing value, as if they pre-existed as symbols in my head.

Following our practice in naming objects and properties around us, we place things in the mind: memories, symbols, patterns. We say: "Something gets recognized." There is an objective something in the world; there is a pattern being searched for in the brain; there is a matching process. Instead, Winograd and Flores tell us, there are just interpretations. There are no preconceived representations, no matching process. Instead, there is a "pre-orientation." "We are always already oriented to a certain direction of possibilities" (p. 147).

Similarly, in psychiatric analysis, the idea of a symbol is used as if it were something that resides in the head. But to say that "*X* symbolizes *Y* for person *P*" is only to say that *P* responds to *X as if* it were *Y*. In a historical interpretation of behavior, we note a pattern and explain it by this association. There need be no translation, no "symbol mapping rules."

The point is more stark than it might first appear. The argument leaves no room for saying that representations are perhaps "compiled," and this is why we have no conscious awareness of translating from representations to words. Rather we are not making decisions at all: *We have no choice, we are simply acting*. There are no "things stored in the brain" that we are searching or selecting between: ". . . the breakdown of a representation and jump to a new one happens independently of our will, as part of our coupling to the world we inhabit" (p. 99).

### 3.3. *All reasoning involves reinterpretation*

Another perspective on "direct action," or what Winograd and Flores following Heidegger call "readiness-to-hand," is that all intelligent reasoning is reinterpretation. This is far more advantageous than acting according to pre-conceived representations, and only finding out later that they are wrong. Yes, we make mistakes because we act inappropriately, but we are not following "plans." There is extreme plasticity in our behavior. Every action is an interpretation of the current situation, based on the entire history of our interactions. In some sense every action is automatically an inductive, adjusted process.

As an example of this phenomenon, close your eyes and consider how many windows are in your bedroom. Did you visualize the process of moving around your room? Possibly we are reactivating a motoric sequence, simulating that we are actually in the room and moving about. We replay the history and

articulate what we would see. But we are not necessarily remembering a particular walk through the room. We are constructing a coherent story, which is implicitly a generalization of our experience because it is based on all of our experience. We chain together a sequence of impressions and pretend that they occurred together. In this way, the chains of association are constructed freshly each time, as a reinterpretation of the unformalized background.

A functional simulation of the cognitive system in terms of manipulated representations cannot generate the range of reinterpretation an unformalized background allows. Winograd and Flores conclude, following Searle, that manipulating a representation formally is not understanding. Certainly there are formal games which we understand how to play. We can even accomplish our goals by playing formal games. But as soon as the interaction changes from the previous history, breakdown occurs, and a reinterpretation in language is required. We engage in a dialogue that articulates the basis of a representation and adjusts it to a new situation. To understand is to be able to make the commitment to do this reinterpretation.

### 3.4. *We assume commitment in other people*

As Weizenbaum pointed out about ELIZA, it is amazing that we are so convinced by so little and that we assume so much. Even when you are told how little ELIZA, SHRDLU, and MYCIN know, it still strains the imagination to appreciate the magnitude of their ignorance. We are all like children preferring to believe that the fantasy is real. Indeed, we don't have to "suspend disbelief" (in the theatrical sense), this is how cognition normally works: Attributing meaningfulness and assuming commitment go hand in hand. The situation is insidiousness: We don't normally articulate shared background, and computers don't have any. It plays right into our assumptions. If you use my words, I assume that you know what I mean. If you say you believe something, I assume that you are ready to convince me.

Weizenbaum stressed the lack of responsibility of computers because they are not part of the social fabric. The argument here is stronger: Computers cannot be responsible because they cannot even form commitments. When I speak with commitment I do more than just mouth words. I do not pretend. I am ready to defend what I say. I am committed. To speak the truth means to be willing and able to articulate why you believe it.

In providing explanations, we must determine why breakdown occurred. What is not obvious, not part of the shared background? What must be articulated? In constructing explanation programs, I have often concluded that we have not placed enough of the burden on the person asking the question. Unless there are systematic surprises that the explainer might guess, the questioner must articulate the nature of his surprise. The explainer must then be ready to form an interpretation that is contrary to his point of view.

However, we cannot completely model how a system's activity will perturb the interaction between user and machine. We can't anticipate every user's interpretations of what the machine is doing (p. 53). Given these limitations, we might focus instead on opening up the program's representations so they are easily browsed, making it easy for the questioner to figure out what he needs to know on his own.

Winograd and Flores provide an intriguing discussion of how computers might be best suited as coordinators of commitments, "the essential dimension of collective work." Following from their analysis of language, they propose that a program keep track of what we have to do, recording the status of our active commitments. Their proposal broadens our awareness from that of the individual working alone at a personal workstation to the social dimension of what is being written, computed, or recorded (p. 158). This idea is likely to have widespread appeal and could significantly enhance how we use computers.

### 3.5. *People do not carry models around in their heads*

Cognitive models explain patterns of behavior; they are developed by scientists. It is a strange and tremendous leap to say that these models actually exist in the heads of the people being modeled. While we commonly say that a person has knowledge, knowledge is not something that you can possess like an object. Knowledge is always an individual interpretation within a shared background. It is neither subjective, nor objective (p. 75). To say that someone knows something is not to say that he is in a certain state, but to explain his behavior over a sequence of interactions and to claim that he is predisposed to act in a similar way in similar situations (p. 47).

Few people believe that when we ride a bicycle we are manipulating internal representations of the handlebars and pedals, modeling their location internally, and computing trajectories. According to Winograd and Flores, speech is the same, a kind of skill coupled to the environment. While we may symbolize our utterances on paper in some calculus or written notation, there is nothing corresponding to these notations in the brain.

This has significant implications for understanding expert behavior. We are not modeling objects that exist inside an expert's head. This explains what is so patently obvious when you work with experts, namely that they have so much difficulty laying out consistent networks and describing relations among concepts in a principled way. If experts knew causal and subsumption networks as discrete concepts and relations, why would we find it so difficult to extract these statements from them? The concepts are often not defined, let alone related in a fixed, systematic way to one another. Experts know how to behave and they know formalizations that model how they behave.

The evidence in student modeling research is similar. Brown and VanLehn [1] found that student errors in subtraction, modeled as bugs in the procedure followed by students, changed over time. They called this "bug migration" and sought a systematic explanation for why the bugs changed. The key is that

there never was a bug in the student's head. When you realize this, you realize that you don't have to explain why the students *decided* that one procedure was wrong and another was better. This does not mean that there is no pattern to be found. Winograd and Flores would say that there is a mechanistic argument, it just isn't based on manipulating a representation.

This analysis might lead to an entirely different teaching method: not isolating the bug, but establishing an appropriate coupling and forcing breakdown. However, we still need to understand what articulation does to behavior. For example, in physical skills, such as playing the piano, attending directly to a faulty action—actually feeling that you are making it happen—allows you to get a grasp on the behavior and change it. The role of language in isolating where an undesirable action occurs automatically is perhaps similar, again, the possible value of psychoanalysis.

By this model, the most effective training occurs on the job site, hence the instructional strategy of getting the students to the workplace and minimizing the classroom blabber. Winograd and Flores' analysis provides a subtler understanding: Teaching involves establishing the history of interaction that constitutes a background that will lead to useful interpretations. Establishing a "structural coupling" means experiencing this history of interactions, not just being told what you would do if you had gone through the process. The problem is familiar: You don't understand me because you don't know where I've been. If all intelligent behavior flows from the unformalized background, teaching how to behave by saying what to do can only provide patterns to follow by rote. You have to try it yourself and get the feel of it. Thus, the repeated teacher's fallback, "You'll know better what I mean when you get out there."

While the implications for instructional computing are not clear, it seems at least theoretically important to realize that there is a difference between solving a problem and articulating a model (rationalizing the solution sequence). We knew this already, but Winograd and Flores provide a theoretical foundation for building up a new understanding, so we can view the inability to articulate a model as not a lack of self-knowledge, but the normal state of affairs.

### 3.6. *Engineering develops from recurrent breakdown*

Understanding and anticipating failure (breakdown) is at the heart of engineering. Essentially, Winograd and Flores have provided a theoretical background for understanding how programming, especially knowledge engineering, is like structural engineering. Programs don't always do what we expect because the designer did not anticipate how users would interact with the program and interpret its actions. A large part of this book concerns how programmers must anticipate the demands of the environment and prepare programs to cope with breakdown.

As in structural engineering, what makes knowledge engineering possible is

that breakdowns recur. These patterns lead the engineer to formulate "objective distinctions." Essentially what we take to be objective truth is what many people have articulated over a long period of time and we as observers expect to continue in the future.

However, constructing an autonomous agent is much more difficult than typical engineering problems. It is like building a bridge that will change its own structure as its interaction with the environment changes. This is the idea of an autopoietic system: It maintains its functions. This theory was developed in biology, and living organisms are the best examples we have. The idea now arises in computer systems engineering and plays an important role in the design of satellites and planetary probes.

A crucial point is that the organism adapts to meet the demands of its interactions. But this is not a process of representing the world: "The demands of continued autopoiesis shape this structure in a way that can be viewed as a reflection of an external world. But the correspondence is not one in which the form of the world is somehow mapped onto the structure of the organism" (p. 62).

3.7. *Systematic domains admit to formal representation*

Systematic domains are those in which there is a great regularity in relations over time and among people, so that there appears to be objective knowledge (p. 172). In modeling intelligent behavior within a systematic domain, we don't and needn't necessarily (indeed, can't) represent the meaning of terms. Rather we represent their systematic role within a network of requests and promises. Indeed, the operation of a program does not require that it represents anything at all. It's all in the eye of the beholder, who interprets input and output in terms of a systematic mapping in his world (p. 86). When MYCIN says the word "culture," you interpret it as something objective that you know about.

This analysis provides a fascinating handle on the nature of language, models, and formalization. It helps us understand why programs work as well as they do and what can go wrong. Thus, we can better understand what we are doing and perhaps how to go about it more systematically.

A surprising conclusion is that the Winograd and Flores analysis motivates a formal approach to knowledge representation. In short, the very nature of modeling is crystallizing our observations, the world as we know it, in terms of objects and properties. Formal representation methods are designed to attack this problem systematically.

The idea of systematic domains may also encourage us to adopt a less mystical view of knowledge acquisition. The task is to get at what's regular, recurrent. For example, in an interactive dialogue system, say a teaching program like GUIDON, we formalize the recurrent conversations that occur between a student and teacher. The idea of recurrence is very powerful for understanding what representation is all about.

Perhaps most important, Winograd and Flores lead us to see objects, properties, and relations as the essence of what a representation is, not just the present-day state of the art in AI research. We can never represent the meaning of terms. And because representation is a formalization process, we should use a logic-based approach, not because it's the solution to modeling intelligent behavior, but because it's precisely what we can do with computers. Formalization is what we're doing anyway, so we might as well be rigorous about it.

The idea that language arises in the need to action also fits very well with our experience in constructing programs. The relevant properties of a representation change as it is interpreted for different purposes such as problem solving, explanation, and cognitive modeling. For example, converting MYCIN to a hypothesize-and-test program required distinguishing between data and hypothesis "parameters" and classifying them. Similarly, an explanation program requires propositions in control rules to be classified as static and dynamic. A student modeling program needs to have control rules classified according to the constraints they satisfy. Thus, we build up representational structures according to the distinctions that are important for operating upon them. "Grounding of descriptions in action pervades all linguistic structure of objects, properties, and events" (p. 171).

The ideas of systematic domains and recurrent background have tremendous importance for knowledge engineering. For example, as we move from representing high-level patterns to representing generative theories for why these patterns occur, we must realize that this might not be possible in all cases.

For example, generating NEOMYCIN's control rules for diagnosis from more primitive concepts is probably impossible because they are based on a huge social context. Certainly, we could always generate a limited set of rules from more primitive concepts, and this might be worthwhile if the diagnostic procedure is relatively stable. But we must keep in mind that the primitives we choose are in an important sense ad hoc. The game can continue for several levels, but we can never get below the representation to something fixed and final. We will always have to assume a set of axioms. Of course, mathematicians have had to face this, and we should not have expected the formalization of medicine or other domains to be any different.

## 4. Unanswered questions

> One cannot construct machines that either exhibit or successfully model intelligent behavior (p. 11).

Winograd and Flores have chosen an uncompromising point of view about the nature of intelligence. Their definition contradicts the commonly accepted view that computer programs "which exhibit behavior we call 'intelligent behavior' when we observe it in human beings" are intelligent [4]. Even if a

computer program consistently wins games of chess, Winograd and Flores would say that this is not intelligence. This restrictive view is unfortunate, because it greatly complicates the problem of understanding this book.

All models are approximate and selective. Engineering models are successful within some practical setting. Rather than insisting that "computers cannot diagnose diseases," for example, which violates common sense, it would be more useful to carefully articulate when the models will fail. Little is gained by saying that such programs are not models of intelligent behavior.

While the book has clear implications for research in natural language and instructional research, as I have described, it is unclear just how well computer models might eventually perform in systematic domains. That is, how well can we circumscribe domains to construct useful and effective representations? Is it of any practical value to say that commitment and hence language cannot exist in a systematic domain (where the importance of unformalized background is minimized)? Is the shared background of people as great as Winograd and Flores suggest when they exclude computers from participating in language? When breakdown occurs, just how well do people resolve it? In this section I consider ways in which the arguments in the book appear to be incomplete and perhaps distort the nature of cognition.

4.1. *What is the mechanism of memory*?

To recapitulate, conceptual structures are not stored in the brain; the concepts of our language do not organize our memory. There are no stored associations, no conceptual network. Instead, we act: We speak, we associate. We don't do this by interpreting a network that mirrors the conceptual structure in what we say. Rather, the history of our behavior may exhibit recurrence that we can represent as such a network.

We are left with the image of some amorphous blob that speaks. How can we explain recurrence, if there is no structural predisposition to associate concepts in some way? Winograd and Flores believe that there is some mechanism behind jumps to new representations, but they provide no description of what it might be. They make no attempt to reinterpret models of memory and learning according to their theory.

For example, what accounts for our tendency to remember exceptions (as described by Schank [11])? Winograd and Flores acknowledge that Schank's work and ideas like "default reasoning" are closer to the nature of cognition, but they insist that these approaches are still limited by the need to distinguish the relevant objects and properties before doing any representation (p. 116).

Is there any evidence of a mechanism that generates the recurrence in our behavior? What about timing experiments involving discrimination and recall? Winograd and Flores claim that these experiments do not deal with "meaningful material" (p. 114). But aren't hierarchical relations meaningful? Couldn't systematicity in abstraction, for example the patterns in levels of "natural

kinds," be explained by structural properties of memory? The problem is that any admission of structure in the brain corresponding to conceptual relations undermines the argument that representations do not exist in the brain.

It appears obvious that the way the brain works favors categorization and association of certain kinds. From here it is but a short step to hierarchical search. Perhaps Winograd and Flores (and Maturana) have got the main idea right, that we aren't *examining* representations internally, but they have woefully ignored the problem of explaining recurrence in memory. It would have been helpful if the book included an appendix that at least acknowledged opposing arguments (such as Fodor's [5]).

4.2. *How much background is shared?*

> If a lion could talk, we could not understand him (Wittgenstein [15]).

According to Winograd and Flores, language requires being able to commit to articulating a shared background. If no common ground is found, then breakdown is not resolved. But the book seems to understate the difficulty of resolving breakdown. Isn't the normal state of affairs one in which individuals (and countries) frequently do not understand each other? We get by normally by making many assumptions and by ignoring differences. Perhaps resolving specific differences is not as important as sharing the goal to "work something out."

A few social rules (part of shared background) for coping with unresolved breakdown may be more important than having the shared background to resolve specific differences. Obviously, not much communication could occur on the basis of just agreeing to disagree. However, the book seems to adopt the opposite stance of idealizing language, suggesting that most breakdowns are resolved in the specific elements of shared background. The general agreement to adapt and cope with ambiguity and unresolved differences could be more important.

The book's idealized description of language is clear when we consider our interactions with animals and children. Shared background is minimized here, but communication is possible. It isn't necessary to be "fully human" to engage in language. Even if computers are programmed to be part of the social structure, even if they are our slaves, interaction with them can be consensual. Their speech acts can create commitment, just as much as a dog can request to play and then become engaged in a game of mock attack or chasing. Certainly there will be practical limitations, as we may not always understand a chimpanzee, and the differences between a dog's bark to play, to eat, or to warn may be too subtle for anyone but his master to discern. However, of what practical value is it to so narrowly define language and intelligence as to rule out the behavior of animals because they are not "fully human"? It is good to make

people more aware of the social dimensions of language, but Winograd and Flores have adopted an almost religious point of view that may overstate the requirements of shared background and the extent to which breakdown is typically resolved.

4.3. *What are the practical limits of formalization*?

The book states that we are "now witnessing a major breakdown in the design of computer technology" (p. 78). No evidence for this observation is given; just the inverse seems to be true. We are witnessing a major recognition of how much human knowledge is regular and can be usefully formalized. In the rapid growth of expert systems applications, engineers in particular are realizing the value of qualitative modeling techniques for describing recurrent objects, properties, and relations. Possibly there has been a misinterpretation of what computers are doing and the nature of intelligence, but the payoff is on the upswing and the limits appear to be years away, at least.

The book provides very little basis for determining the practical limits of formalization, particularly for applications of Artificial Intelligence to science and engineering. Perhaps by continuing to find structure within structure we can get programs that are very good, and even fool most people. Yes, they will fail sometimes, but so do people. There is little evidence that the practical limitations of formal reasoning are as serious as the book suggests.

Practical implications of the argument tend to return to conclusions we already knew, as I indicated in briefly considering explanation, teaching, and knowledge acquisition. However, the book gives us an improved understanding of what is difficult and why we might not succeed. The most important change might be a better understanding of what we are doing.

4.4. *Isn't reflection an essential part of reasoning*?

> Human cognition includes the use of representations, but is not based on representation. Experts do not need to have formalized representations in order to act. They may at times manipulate representations as one part of successful activity, but it is fruitless to search for a full formalization of the pre-understanding that underlies all thought and action (p. 99).
>
> The essence of our intelligence is our thrownness, not our reflection.

I believe that this book significantly understates the importance of reflection, to the point of distorting the nature of cognition. In reflection, we articulate our background in order to compare possible behaviors, anticipate consequences, and plan, rather than acting impulsively. Even granting the nature of unformalized background, readiness-to-hand, and the immediate nature of reflection (we don't decide to reflect), the valued action in a consensual

domain is one that anticipates ramifications. Human reasoning is immensely more successful by our ability to simulate what might happen, to visualize possible outcomes and prepare for them. We do this by reflecting, saying what we expect, and responding to what we say. (An excellent description of this imagination process appears in Jaynes' [7].)

We create representations by language, by acting. We make interpretations by what we say. Every representation is an interpretation. But isn't every representation therefore potentially crucial in our action? Granted that representations are not "inside" and that they are blind, once articulated don't they play a central role in intelligent behavior? We are always reinterpreting old representations. We are not just speaking like birds singing. The articulation is essential, it can change our behavior. Winograd and Flores fail to properly emphasize the loop: We are always listening to ourselves. Even if representations are not directly generated from representations, they are generated *in response to* representations. In particular, imagery and silent utterances are a form of "mental representation" which is part of the cognitive system. While these representations may be unnecessary for behavior, much behavior is mediated by them.

The weakness of the argument minimizing the centrality of representation is most clear in the example of the chairman who is always directly acting (p. 34). Winograd and Flores greatly understate the importance of making observations, forming hypotheses, and consciously choosing a course of action. In chairing a meeting, I attend, stop myself from saying something (anticipating a reaction), plan things to say, arrange a list of people to call upon, attempt to weigh alternative topics, watch the time, and suggest a revised agenda. The book seems to overgeneralize the nature of physical skills—as provided by the example of a hammer and how we attend to it—in suggesting that cognitive behavior generally has the same degree of automaticity and lack of reflection. Granted behavior must be immediate; there is no homunculus inside interpreting representations. But forming representations and reinterpreting them is where all of the action is! Cognitive behavior is strongly coupled to the representations it creates; as visualizations and silent utterances, they are "inside" the system as much as anything else.

Most of my day involves an inner conversation. Most of my awake activity is a long sequence of telling and asking statements to myself. Granted, I don't know where the questions come from (I don't have to work at firing neurons). Granted, I don't know where the answers come from. I just keep making requests and promises to myself. "What are all of the projects I'm working on now? I have to call Jan. What will I do when I finish this? I'll work on the review tomorrow. Where is my yellow pad?" Most of my life seems to involve responding to my own language, the representations I generate.

Winograd and Flores appear to have got the emphasis wrong. In emphasizing that TELL and ASK actions do not come from interpreted representations,

they ignore the crucial point that thinking involves the generation of representations and attending to them. We are constantly observers to our own thinking behavior. We are constantly responding to representations.

Most important, we tell ourselves what we *might do*. Then we react to this. And in our reaction we promise ourselves that we will do something different or make a request. We do not just simply act. We are engaged in a loop of imagining action and visualizing consequences. Yes, our words and motoric actions always proceed directly, but often not before intervening representational actions and sometimes not at all without them.

How then do we get to the state of reflecting? What, after all, changes us? Perhaps reflection is built in? Being able to place ourselves in a situation so we can know how we might behave is incredibly powerful. It means being able to simulate a structural coupling, to know what we are apt to do. This is much more than articulating a background; it is articulating the behavior that the background will elicit. By projecting forward in this way, admittedly with uncertainty, we can anticipate the consequences of behavior. This anticipation then has the potential of changing our background and resultant behavior.

By overemphasizing the direct, ready-to-hand, unreflective core of all behavior (including reflection itself), Winograd and Flores understate the importance of representation in intelligent behavior. That an expert can act without a representation is not very interesting in comparison to how impoverished his behavior would be if representations were not available for solving the difficult problems.

## 5. Recommendations

This book should make AI researchers more cautious about what they are doing, more aware of the nature of formalization, and more open to alternative views. By addressing the nature of representation and reasoning with examples familiar to most AI researchers, the book has the potential of being more influential than other criticisms of the field.

A scientific enterprise requires openness to blindness of all kinds. This book explains why blindness is inevitable and elevates our awareness of the origin of language and how breakdowns occur. When the Chenobyls and Challengers of AI occur, we can look back at this book to better understand why our programs failed. The book provides an important theoretical basis for the analysis of failure in knowledge engineering. Indirectly it tells us how to analyze domains: What are the recurrent dialogues? What breakdowns occur? What are the expert's methods for coping with breakdown? What are the shared sources of experience? Who can talk to whom and why?

Every AI researcher should read this book. Designers of interactive programs interested in theoretical aspects of language and improving their understanding of what they are doing will find this book to be fascinating, engrossing, and obstinately provocative. The title is apt: If you are interested in

understanding what computers can do, for example how you might use them in your business, and have a philosophical bent, you should definitely read this book. However, be forewarned that it points the way, rather than providing answers.

The authors state that the book is not intended to be a scholarly treatise, and it was probably a good idea to simplify the presentation in this way. However, I think the book will mostly appeal to researchers and academicians, and these readers should be aware that there are other books that adopt similar points of view. For example, I learned about "readiness-to-hand" from reading Polanyi [8], who calls the idea "tacit knowledge" (using the same hammer example). Yet, Winograd and Flores do not cite Polanyi, and Polanyi does not cite Heidegger. The intellectual development of the ideas is therefore obscure. I believe that Richard Rorty's [10] *Philosophy and the Mirror of Nature* (not cited by Winograd and Flores) is a good reference for readers who want a more complete understanding of the argument against the idea of internal representation. *Understanding Computers and Cognition* goes a long way towards making philosophical works like this more accessible to AI researchers.

In conclusion, even though the book is extremely well written, its arguments are so counterintuitive many readers are likely to remain confused and unconvinced. The book helps resolve foundational issues of AI, but the practical implications are unclear.

One goal for writing the book was to prevent a false view of computers from distorting our understanding of people. Ironically, the book's new view of cognition is a little scary, making reasoning seem limited and out of our personal control. The earth is not in the center; man is not in the center, and neither is his conscious mind. The relation of responsibility to reflection needs to be better developed and balanced against the core of automaticity that lies behind behavior. On the other hand, the book supports a humanist position, emphasizing our commonality, that what we are is mostly what we do together.

Certainly, this book might change how you think about the world. As I squashed a huge mosquito the other night, I thought, "So much for another structural coupling."

## REFERENCES

1. Brown, J.S. and VanLehn, K., Repair theory: A generative theory of bugs in procedural skills, *Cognitive Sci.* **4** (4) (1980) 379–415.
2. Clancey, W.J., From Guidon to Neomycin and Heracles, *AI Magazine* **7** (3) (1986) 40–60.
3. Dreyfus, H.L., *What Computers Can't Do: A Critique of Artificial Reason* (Harper and Row, New York, 1972).
4. Feigenbaum, E.A. and Feldman, J., *Computers and Thought* (Robert E. Krieger, Malabar, 1963).
5. Fodor, J.A., *The Language of Thought* (Harvard University Press, Cambridge, MA, 1975).
6. Gould, S.J., *Ever Since Darwin* (Norton, New York, 1977).
7. Jaynes, J., *The Origins of Consciousness in the Breakdown of the Bicameral Mind* (Houghton-Mifflin, Boston, MA, 1976).

8. Polanyi, M., *The Tacit Dimension* (Anchor Books, Garden City, NY, 1967).
9. Pribram, K.H., *Languages of the Brain* (Wadsworth, Monterey, CA, 1977).
10. Rorty, R., *Philosophy and the Mirror of Nature* (Princeton University Press, Princeton, NJ, 1979).
11. Schank, R.C., Failure-driven memory, *Cognition and Brain Theory* **4** (1) (1981) 41–60.
12. Weizenbaum, J., *Computer Power and Human Reason: From Judgment to Calculation* (Freeman, San Francisco, CA, 1976).
13. Wilber, K.W. (Ed.), *The Holographic Paradigm and Other Paradoxes* (Shambhala, Boulder, CO, 1982).
14. Winograd, T., *Understanding Natural Language* (Academic Press, New York, 1972).
15. Wittgenstein L., *Philosophical Investigations* (Macmillan, New York, 1958).

## On *Understanding Computers and Cognition: A New Foundation for Design*

### A response to the reviews

Terry Winograd
*Computer Science Department, Stanford University, Stanford, CA 94305, U.S.A.*

Fernando Flores
*Logonet, Inc., Berkeley, CA 94704, U.S.A.*

---

### 1. A theory of language

In *Understanding Computers and Cognition*, we presented a theory of language, on which we base our understanding of cognition and of computers. It includes some basic assertions about how language works:

(1) Language does not convey information. It evokes an understanding, or "listening," which is an interaction between what was said and the pre-understanding already present in the listener.

(2) An utterance produces different understanding for different listeners, since each person has a background of pre-understanding generated by a particular history. This background does not determine interpretation in a rigid way, but generates the domain of possibilities for how what is heard will be interpreted.

(3) The background that is relevant to understanding grows out of concerns, practices, and breakdowns in those practices. People interpret language in a way that makes sense for what they do.

(4) The background of concerns and practices is not purely individual, but is generated within a tradition. Each person is unique, sharing background to varying extents with other people. Some amount is universally human; more is shared with members of the same culture; more yet with those in the same line of work; and still more with partners in frequent conversation.