What Computers *Still* Can't Do

A Critique of Artificial Reason

Hubert L. Dreyfus

The MIT Press
Cambridge, Massachusetts
London, England

Introduction to the Revised Edition

What Computers Can't Do stirred up a controversy among all those interested in the possibility of formal models of man by arguing that, despite a decade of impressive print-outs and dire predictions of superintelligent robots, workers in artificial intelligence (AI) were, in 1967, facing serious difficulties which they tended to cover up with special-purpose solutions and rhetorical claims of generality. During the subsequent decade this critique has been more or less acknowledged. In the five-year period from 1967 to 1972 the *ad hoc* character of AI work was admitted and, indeed, elevated to a methodological principle. The study of artificially circumscribed gamelike domains was proclaimed a study of *micro-worlds* and was defended as a necessary first step toward broader and more flexible programs. Then, during the next five years (1972–1977) the micro-world "successes" were seen to be ungeneralizable, and in the best AI laboratories workers began to face the problem of representing the everyday general understanding which they had spent the first fifteen years of research trying to circumvent. Recently, even the wishful rhetoric characteristic of the field has been recognized and ridiculed by AI workers themselves.

My early outrage at the misleading names given to programs such as Newell, Shaw, and Simon's General Problem Solver (GPS) is now shared by M.I.T.'s Drew McDermott, who writes:

[I]n AI, our programs to a great degree are problems rather than solutions. If a researcher tries to write an "understanding" program, it isn't because he has thought of a better way of implementing this well-understood task, but because he hopes he can come closer to writing the *first* implementation. If he calls the main loop of his program "UNDERSTANDING", he is (until proven innocent) merely begging the question. He may mislead a lot of people, most prominently himself, and enrage a lot of others.[1]*§

McDermott also singled out overrated GPS:

Many instructive examples of wishful mnemonics by AI researchers come to mind once you see the point. Remember GPS? By now, "GPS" is a colorless term denoting a particularly stupid program to solve puzzles. But it originally meant "General Problem Solver", which caused everybody a lot of needless excitement and distraction. It should have been called LFGNS—"Local Feature-Guided Network Searcher".[2]

Even my earliest assessment that work in AI resembled alchemy more than science[3] has been accepted by Terry Winograd, formerly at M.I.T., now at Stanford:

In some ways, [AI] is akin to medieval alchemy. We are at the stage of pouring together different combinations of substances and seeing what happens, not yet having developed satisfactory theories. This analogy was proposed by Dreyfus (1965) as a condemnation of artificial intelligence, but its aptness need not imply his negative evaluation . . . it was the practical experience and curiosity of the alchemists which provided the wealth of data from which a scientific theory of chemistry could be developed.[4]

Winograd is right; as long as researchers in AI admit and learn from their failures their attempt to supply computers with human knowledge may in the end provide data for a totally different way of using computers to make intelligent artifacts. But until recently, admitting their failures so that others can learn from their mistakes—an essential part of any scientific field—has been virtually unknown in AI circles. McDermott reiterates my point that, as he puts it, ". . . AI as a field is starving for a few carefully documented failures." And he warns: "Remember,

§Notes begin on p. 307. [Citations are indicated by a superior figure. Substantive notes are indicated by a superior figure and an asterisk.]

though, if we can't criticize ourselves, someone else will save us the trouble."[5] I take this as my cue to return for a critical look at the research of the past ten years.[6]*

What strikes me, and has struck other writers reviewing the history of the field,[7] is how my views and those of workers interested in the theoretical issues in AI have gradually converged. In recent years the attempt to produce special-purpose programs tailored to narrowly restricted domains, with the concomitant principle that this should be achieved in whatever way is most efficient regardless of whether such methods are used by human beings, has been abandoned by AI *theorists* and frankly and quite successfully taken over by self-styled AI *engineers,* with no interest in making generally intelligent machines. Among those still interested in the theoretical issue of using computers to produce the full range of human intelligent behavior there is now general agreement that, as I argue in this book, intelligence requires understanding, and understanding requires giving a computer the background of common sense that adult human beings have by virtue of having bodies, interacting skillfully with the material world, and being trained into a culture.

Given the epistemological assumptions dictated by the information-processing model (see Chapter 4) this precondition of intelligent behavior necessarily appears to AI workers as the need to find a formal representation in which all the knowledge and beliefs of an average adult human being can be made explicit and organized for flexible use. Almost everyone now (with one exception we will deal with later) agrees that representing and organizing commonsense knowledge is incredibly difficult, and that facing up to this problem constitutes the moment of truth for AI. Either a way of representing and organizing everyday human know-how must be found, or AI will be swamped by the welter of facts and beliefs that must be made explicit in order to try to inform a disembodied, utterly alien computer about everyday human life. With this recognition, which characterizes the most recent five-year phase of AI research, unfounded optimism has given way to somewhat self-critical caution.

AI research has thus passed from stagnation to crisis during the decade since I concluded my research for this book. If I were to rewrite

the book today I would divide this decade into two phases and include them as Chapters 3 and 4 of Part I, so as to cover the full twenty years the field has been in existence. And I would modify the Conclusion to take into account the recent maturation of the field. But since the overall argument of the book is confirmed rather than contradicted by the latest developments, I would rather leave the original book intact—only reworking the material where a sentence or a paragraph has proved to be murky or misleading—while including what are, in effect, Chapters 3 and 4 and the new conclusion in this Introduction. The reader who wants to get a chronological sense of how research in artificial intelligence developed should skip ahead to Chapters 1 (Phase 1) and 2 (Phase 2), and then return to this critical survey of the past ten years. Moreover, since the arguments at the end of this Introduction presuppose and extend ideas which are more fully developed in the last half of the book, the conclusion of the Introduction to the Revised Edition might be best read after finishing Part III.

Phase III (1967–1972)
Manipulating Micro-Worlds

When *What Computers Can't Do* appeared in January 1972, making a case that after an exciting start which raised high hopes, work in artificial intelligence had been stagnating, reviewers within the field of AI were quick to point out that the research criticized was already dated and that my charge of stagnation did not take into account the "breakthroughs" which had occurred during the five years preceding the publication of my critique. Bruce Buchanan's reaction in *Computing Reviews* is typical:

> One would hope that a criticism of a growing discipline would mention work in the most recent one-third of the years of activity. . . . To this reviewer, and other persons doing AI research, programs developed in the last five years seem to outperform programs written in the tool-building period of 1957–1967.
>
> For example, it is dishonest to entitle the book a "critique" of AI when it dwells on the failure of early language translation programs (based primarily on syntactical analysis) without analyzing the recent work on understanding natural language (based on syntax, semantics, and context).[8]

If the point of these objections had been that my book did not take account of excellent programs such as M.I.T.'s MATHLAB (1970) for manipulating symbolic algebraic expressions, and Stanford's DENDRAL (1970) for inferring chemical structure from mass spectometry data, I would plead guilty. I would point out, however, that these programs, while solving hard technical problems and producing programs that compete with human experts, achieve success precisely because they are restricted to a narrow domain of facts, and thus exemplify what Edward Feigenbaum, the head of the DENDRAL project, has called "knowledge engineering."[9] They, thus, do not constitute, nor are they meant to constitute, progress toward producing general or generalizable techniques for achieving adaptable intelligent behavior.

Buchanan would presumably agree since the programs he mentions as giving the lie to my accusations of stagnation are not these engineering triumphs, but theoretically oriented projects such as Winograd's natural language understanding program, and the perception programs developed at M.I.T. and Stanford.[10] These, plus Patrick Winston's concept learning program, are the programs most often cited by those who claim that my book ignores a breakthrough which occurred around 1970. If these programs do constitute progress they roundly refute my charge of stagnation, so we must now take a detailed look at each of the programs, as well as a brief look at robot research, which was generating a great deal of excitement just as this book went to press.

Since by all accounts it is Winograd's natural language understanding program (SHRDLU—1971) which demonstrates what Papert and Minsky call "a major advance,"[11] we will turn to it first. SHRDLU simulates a robot arm which can move a set of variously shaped blocks and allows a person to engage in a dialogue with the computer, asking questions, making statements, issuing commands, about this simple world of movable blocks. The program is an integrated system which makes use of syntax, semantics, and facts about blocks. It is thus, as Winograd cautiously claims, "a computer program which 'understands' language in a limited domain, by including a model of the subject being talked about and a context of discourse. . . ."[12]* Furthermore, this program explores the new idea of representing this knowledge in the

form of procedures. The resulting dialogue is strikingly natural. Here is a sample. (The sentences in capital letters are computer print-out. The parentheses contain Winograd's comments.)

pick up a big red block.

OK.

(The system answers "OK" when it carries out a command. In order to pick up the red block, it had to clear it off by finding a space for the green one and moving the green one away.)

2 grasp the pyramid.

I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

(There are three pyramids in the scene, so the program does nothing.)

find a block which is taller than the one you are holding and put it into the box.

(The system has to figure out what is meant by "one" and "it", as well as using deduction to decide which block it is holding.)

BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

(Printed out to keep us informed.)

OK.

(The hand has to set down the red block to pick up the taller blue one.)

4. what does the box contain?

THE BLUE PYRAMID AND THE BLUE BLOCK.

what is the pyramid supported by?

(Earlier the system responded to "the pyramid" by saying "I don't understand which pyramid you mean". Now, it assumes that we mean the one just mentioned.)

THE BOX.

9. can the table pick up blocks?

NO.

(The semantic rules make this a nonsensical question which therefore must be false.)

10. can a pyramid be supported by a block?

YES.

(The deductive system finds an actual example, so it knows this is possible.)

11. can a pyramid support a pyramid?

I DON'T KNOW.

(The question makes sense, but there is no example in the scene, so there is no easy way to tell without being able to examine the robot's programs.)

12. stack up two pyramids.

I CAN'T.

(The robot tries and fails.)

13. the blue pyramid is mine.

I UNDERSTAND.

(This data is added to the data base.)

14. I own blocks which are not red, but I don't own anything which supports a pyramid.

I UNDERSTAND.

(Two new theorems are created for proving things about "owning".)

15. do I own the box?

NO.[13]

If being a major advance means being more general than its predecessors, then SHRDLU, since it no longer depends for its analysis of language on external constraints of the sort essential to Bobrow's STUDENT, is a major achievement. Winograd himself points out the superiority of his program over earlier language understanding programs:

. . . A program may succeed by carefully choosing the problem it will attack, so that some simple special-purpose solution will work. ELIZA (Weizenbaum, 1964) and STUDENT (Bobrow, 1967) are examples of programs which give impressive performances owing to a severe and careful restriction of the kind of

understanding they try to achieve. If a model is to be of broader significance, it must be designed to cover a large range of the things we mean when we talk of understanding. The principles should derive from an attempt to deal with the basic cognitive structures.[14]

If, however, "a major advance" means that a step has been made in dealing with the basic cognitive structures needed to cover everyday understanding—that thanks to SHRDLU there is now reason to be optimistic about the possibility of AI—then no progress at all can be claimed. To justify this negative judgment we must first find out how the optimists of the early seventies were able to convince themselves that, with SHRDLU, AI was at last on the right track.

If one holds, as some AI workers such as Winograd do, that there are various kinds of understanding so that whether an entity has understanding or not is just a question of degree, it may seem that each new program has a bit more understanding than the last, and that progress consists in inching out on the understanding continuum. If, on the other hand, one holds that "understanding" is a concept that applies only to entities exactly like human beings, that would stack the deck and make AI impossible. But it is not up to either side in the debate to stipulate what "understanding" means. Before talking of *degrees* of "understanding," one must note that the term "understand" is part of an interrelated set of terms for talking about behavior such as "ask," "answer," "know," etc. And some of these terms—such as "answer," for example—simply *do* have an all-or-nothing character. If one is tempted to say that the DENDRAL program, for example, literally *understands* mass spectroscopy, then one must be prepared to say that when it is fed a problem and types out the answer it has literally been asked and answered a question, and this, in turn, involves, among other things, that it *knows* that it has answered. But whatever behavior is required for us to say of an entity that it "knows" something, it should be clear that the computer does not now come near to meeting these conditions, so it has not answered even a little. If one is sensitive to the central meaning of these interconnected intentional terms it follows that the claim that programs like SHRDLU have a little bit of understanding is at best metaphorical and at most outright misleading.

Workers in AI were certainly not trying to cover up the fact that it

was SHRDLU's restricted domain which made apparent understanding possible. They even had a name for Winograd's method of restricting the domain of discourse. He was dealing with a micro-world. And in a 1970 internal memo at M.I.T., Minsky and Papert frankly note:

Each model—or "micro-world" as we shall call it—is very schematic; it talks about a fairyland in which things are so simplified that almost every statement about them would be literally false if asserted about the real world.[15]

But they immediately add:

Nevertheless, we feel that they [the micro-worlds] are so important that we are assigning a large portion of our effort toward developing a collection of these micro-worlds and finding how to use the suggestive and predictive powers of the models without being overcome by their incompatibility with literal truth.[16]

Given the admittedly artificial and arbitrary character of micro-worlds, why do Minsky and Papert think they provide a promising line of research?

To find an answer we must follow Minsky and Papert's perceptive remarks on narrative and their less than perceptive conclusions:

. . . In a familiar fable, the wily Fox tricks the vain Crow into dropping the meat by asking it to sing. The usual test of understanding is the ability of the child to answer questions like:

"Did the Fox think the Crow had a lovely voice?"

The topic is sometimes classified as "natural language manipulation" or as "deductive logic", etc. These descriptions are badly chosen. For the real problem is not to understand English; it is to *understand* at all. To see this more clearly, observe that nothing is gained by presenting the story in simplified syntax: CROW ON TREE. CROW HAS MEAT. FOX SAYS "YOU HAVE A LOVELY VOICE. PLEASE SING." FOX GOBBLES MEAT. The difficulty in getting a machine to give the right answer does not at all depend on "disambiguating" the words (at least, not in the usual primitive sense of selecting one "meaning" out of a discrete set of "meanings"). And neither does the difficulty lie in the need for unusually powerful logical apparatus. The main problem is that no one has constructed the elements of a body of knowledge about such matters that is adequate for understanding the story. Let us see what is involved.

To begin with, there is never a unique solution to such problems, so we do not ask what the Understander *must* know. But he will surely gain by having the concept of FLATTERY. To provide this knowledge, we imagine a "micro-theory" of flattery—an extendible collection of facts or procedures that describe conditions under which one might expect to find flattery, what forms it takes, what its consequences are, and so on. How complex this theory is depends on what is presupposed. Thus it would be very difficult to describe flattery to our Understander if he (or it) does not already know that statements can be made for purposes other than to convey literally correct, factual information. It would be almost impossibly difficult if he does not even have some concept like PURPOSE or INTENTION.[17]

The surprising move here is the conclusion that there could be a circumscribed "micro-theory" of flattery—somehow intelligible apart from the rest of human life—while at the same time the account shows an understanding of flattery opening out into the rest of our everyday world, with its understanding of purposes and intentions.

What characterizes the period of the early seventies, and makes SHRDLU seem an advance toward general intelligence, is the very concept of a micro-world—a domain which can be analyzed in isolation. This concept implies that although each area of discourse seems to open out into the rest of human activities its endless ramifications are only apparent and will soon converge on a self-contained set of facts and relations. For example, in discussing the micro-world of bargaining, Papert and Minsky consider what a child needs to know to understand the following fragment of conversation:

Janet: "That isn't a very good ball you have. Give it to me and I'll give you my lollipop.[18]

And remark

. . . we conjecture that, eventually, the required micro-theories can be made reasonably compact and easily stated (or, by the same token, *learned*) once we have found an adequate set of structural primitives for them. When one begins to catalogue what one needs for just a little of Janet's story, it seems at first to be endless:

Time	Things	Words
Space	People	Thoughts

Talking: Explaining. Asking. Ordering. Persuading. Pretending
Social relations: Giving. Buying. Bargaining. Begging. Asking. Presents. Stealing . . .
Playing: Real and Unreal, Pretending
Owning: Part of, Belong to, Master of, Captor of
Eating: How does one compare the values of foods with the values of toys?
Liking: good, bad, useful, pretty, conformity
Living: Girl. Awake. Eats. Plays.
Intention: Want. Plan. Plot. Goal. Cause. Result. Prevent.
Emotions: Moods. Dispositions. Conventional expressions.
States: asleep. angry. at home.
Properties: grown-up. red-haired. called "Janet".
Story: Narrator. Plot. Principal actors.
People: Children. Bystanders.
Places: Houses. Outside.
Angry: State
caused by: Insult
deprivation
assault
disobedience
frustration
spontaneous
Results not cooperative
lower threshold
aggression
loud voice
irrational
revenge
Etc.[19]

They conclude

But [the list] is not endless. It is only large, and one needs a large set of concepts to organize it. After a while one will find it getting harder to add new concepts, and the new ones will begin to seem less indispensable.[20]

This totally unjustified belief that the seemingly endless reference to other human practices will converge so that simple micro-worlds can be studied in relative isolation reflects a naive transfer to AI of methods that have succeeded in the natural sciences. Winograd characteristically describes his work in terms borrowed from physical science:

We are concerned with developing a formalism, or "representation," with which to describe . . . knowledge. We seek the "atoms" and "particles" of which it is built, and the "forces" that act on it.[21]

It is true that physical theories about the universe can be built up by studying relatively simple and isolated systems and then making the model gradually more complex and integrating it with other domains of phenomena. This is possible because all the phenomena are presumably the result of the lawlike relations of a set of basic elements, what Papert and Minsky call "structural primitives." This belief in local success and gradual generalization was clearly also Winograd's hope at the time he developed SHRDLU.

The justification for our particular use of concepts in this system is that it is thereby enabled to engage in dialogs that simulate in many ways the behavior of a human language user. For a wider field of discourse, the conceptual structure would have to be expanded in its details, and perhaps in some aspects of its overall organization.[22]

Thus, for example, it might seem that one could "expand" SHRDLU's concept of owning, since in the above sample conversation SHRDLU seems to have a very simple "micro-theory" of owning blocks. But as Simon points out in an excellent analysis of SHRDLU's limitations, the program does not understand owning at all because it cannot deal with meanings. It has merely been given a set of primitives and their possible relationships. As Simon puts it:

The SHRDLU system deals with problems in a single blocks world, with a fixed representation. When it is instructed to "pick up a big red block", it needs only to associate the term "pick up" with a procedure for carrying out that process; identify, by applying appropriate tests associated with "big", "red", and "block", the argument for the procedure and use its problem-solving capabilities to carry out the procedure. In saying "it needs only", it is not my intention to demean the capabilities of SHRDLU. It is precisely because the program possesses stored programs expressing the intensions of the terms used in inquiries and instructions that its interpretation of those inquiries and instructions is relatively straightforward.[23]

In understanding, on the other hand, "the problem-understanding subsystem will have a more complicated task than just mapping the input

language onto the intentions stored in a lexicon. It will also have to create a representation for the information it receives, and create meanings for the terms that are consistent with the representation."[24] So, for example, in the conversation concerning owning:

. . . although SHRDLU's answer to the question is quite correct, the system cannot be said to understand the meaning of "own" in any but a sophistic sense. SHRDLU's test of whether something is owned is simply whether it is tagged "owned". There is no intensional test of ownership, hence SHRDLU knows what it owns, but doesn't understand what it is to own something. SHRDLU would understand what it meant to own a box if it could, say, test its ownership by recalling how it had gained possession of the box, or by checking its possession of a receipt in payment for it; could respond differently to requests to move a box it owned from requests to move one it didn't own; and, in general, could perform those tests and actions that are generally associated with the determination and exercise of ownership in our law and culture.[25]

Moreover, even if it satisfied all these conditions it still wouldn't understand, unless it also understood that it (SHRDLU) couldn't own anything, since it isn't a part of the community in which owning makes sense. Given our cultural practices which constitute owning, a computer cannot own something any more than a table can.

This discussion of owning suggests that, just as it is misleading to call a program UNDERSTAND when the problem is to find out what understanding is, it is likewise misleading to call a set of facts and procedures concerning blocks a micro-*world,* when what is really at stake is the understanding of what a world is. A set of interrelated facts may constitute a *universe,* a domain, a group, etc., but it does not constitute a *world,* for a world is an organized body of objects, purposes, skills, and practices in terms of which human activities have meaning or make sense. It follows that although there is a children's world in which, among other things, there are blocks, there is no such thing as a blocks world. Or, to put this as a critique of Winograd, one cannot equate, as he does, a program which deals with "a tiny bit of the world," with a program which deals with a "mini-world."[26]

In our everyday life we are, indeed, involved in various "sub-worlds" such as the world of the theater, of business, or of mathematics, but each

of these is a "mode" of our shared everyday world.[27]* That is, sub-worlds are not related like isolable physical systems to larger systems they *compose;* rather they are local elaborations of a whole which they *presuppose.* If micro-worlds *were* sub-worlds one would not have to extend and combine them to reach the everyday world, because the everyday world would have to be included already. Since, however, micro-worlds are *not* worlds, there is no way they can be combined and extended to the world of everyday life. As a result of failing to ask what a world is, five more years of stagnation in AI was mistaken for progress.

Papert and Minsky's 1973 grant proposal is perhaps the last time the artificially isolated character of the micro-world is defended as a scientific virtue—at least at M.I.T.:

> Artificial Intelligence, as a new technology, is in an intermediate stage of development. In the first stages of a new field, things have to be simplified so that one can *isolate* and study the *elementary* phenomena. In most successful applications, we use a strategy we call "working within a Micro-World".[28]

SHRDLU is again singled out as the most successful version of this research method. "A good example of a suitably designed Micro-world is shown in the well-known project of Winograd, which made many practical and theoretical contributions to Understanding Natural Language."[29] But while gestures are still made in the direction of generalization it is obvious that SHRDLU is running into difficulty.

> Since the Winograd demonstration and thesis, several workers have been adding new elements, regulations, and features to that system. That work has not gone very far, however, because the details of implementation of the original system were quite complex.[30]

Such failures to generalize no doubt lie behind the sober evaluation in a proposal two years later:

> . . . Artificial Intelligence has done well in tightly constrained domains—Winograd, for example, astonished everyone with the expertise of his blocks-world natural language system. Extending this kind of ability to larger worlds has not proved straightforward, however. . . . The time has come to treat the problems involved as central issues.[31]

But typically, it is only from the vantage point of the next phase of research, with its new hopes, that the early seventies' illusion that one can generalize work done in narrowly constrained domains is finally diagnosed and laid to rest. Winograd himself acknowledges that:

> The AI programs of the late sixties and early seventies are much too literal. They deal with meaning as if it were a structure to be built up of the bricks and mortar provided by the words, rather than a design to be created based on the sketches and hints actually present in the input. This gives them a "brittle" character, able to deal well with tightly specified areas of meaning in an artificially formal conversation. They are correspondingly weak in dealing with natural utterances, full of bits and fragments, continual (unnoticed) metaphor, and reference to much less easily formalizable areas of knowledge.[32]

Another supposed breakthrough mentioned by Buchanan is Adolfo Guzman's program, SEE (1968), which analyzes two-dimensional projections of complicated scenes involving partially occluded three-dimensional polyhedra. (See Figure 1). Already as developed by Guzman this program could outdo human beings in unscrambling some classes of complicated scenes, and as generalized by David Waltz it is even more impressive. It not only demonstrates the power gained by restricting the domain analyzed, but it also shows the kind of generalization that *can* be obtained in micro-world work, as well as indirectly showing the kind of generalization that is precluded by the very nature of special-purpose heuristics.

Guzman's program analyzes scenes involving cubes and other such rectilinear solids by merging regions into bodies using evidence from the vertices. Each vertex suggests that two or more of the regions around it belong together depending on whether the vertex is shaped like an L, an arrow, a T, a K, an X, a fork, a peak, or an upside-down peak. With these eight primitives and commonsense rules for their use, Guzman's program did quite well. But it had certain weaknesses. According to Winston, "The program could not handle shadows, and it did poorly if there were holes in objects or missing lines in the drawing."[33] Waltz then generalized Guzman's work and showed that by introducing three more such primitives, a computer can be programmed to decide if a particular line in a drawing is a shadow, a crack, an obscuring edge, or an internal

seam in a way analogous to the solution of sets of algebraic equations. As Winston later sums up the change:

> Previously it was believed that only a program with a complicated control structure and lots of explicit reasoning power could hope to analyze scenes like that in figure [1]. Now we know that understanding the constraints the real world imposes on how boundaries, concave and convex interiors, shadows, and cracks can come together at junctions is enough to make things much simpler. A table which contains a list of the few thousand physically possible ways that line types can come together accompanied by a simple matching program are all that is required. Scene analysis is translated into a problem resembling a jigsaw puzzle or a set of linear equations. No deep problem solving effort is required; it is just a matter of executing a very simple constraint dependent, iterative process that successively throws away incompatible line arrangement combinations.[34]

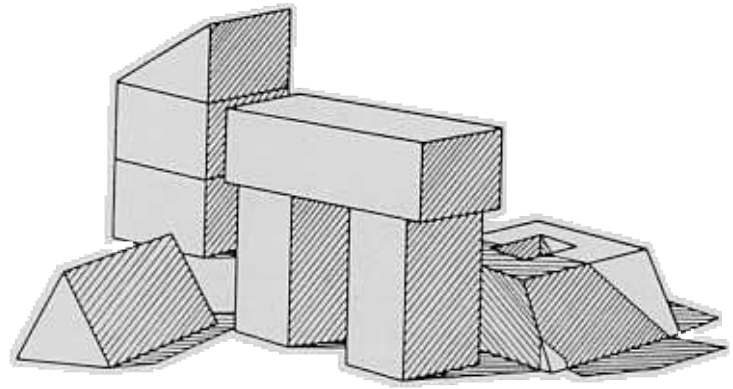
This is just the kind of mathematical generalization within a domain one might expect in micro-worlds where the rule-governed relation of the primitives (in this case the set of vertices) are under some external constraint (in this case the laws of geometry and optics). What one would not expect is that the special-purpose heuristics which depend on corners for segregating rectilinear objects could in any way be generalized so as to make possible the recognition of other sorts of objects. And, indeed, none of Guzman's or Waltz's techniques, since they rely on the intersection of straight lines, have any use in analyzing a scene involving curved objects. What one gains in narrowing a domain, one loses in breadth of significance. Winston's evaluation covers up this lesson:

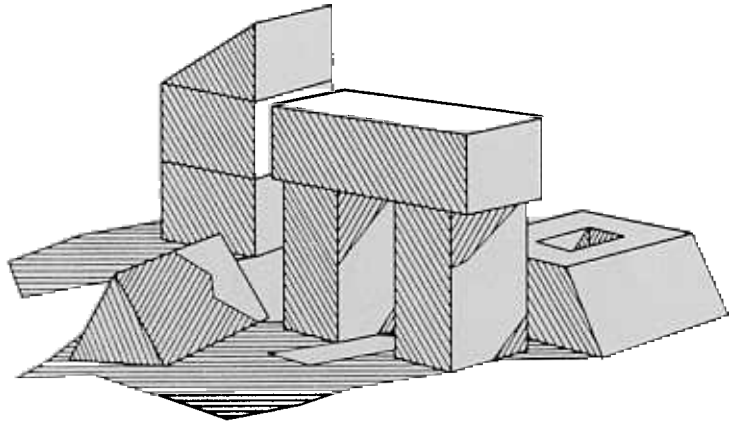
> . . . It is wrong to think of Waltz's work as only a statement of the epistemology of line drawings of polyhedra. Instead I think it is an elegant case study of a paradigm we can expect to see again and again, and as such, it is a strong metaphoric tool for guiding our thinking, not only in vision but also in the study of other systems involving intelligence.[35]

But in a later grant proposal he acknowledges that:

> To understand the real world, we must have a different set of primitives from the relatively simple line trackers suitable and sufficient for the blocks world.[36]

Waltz's work is a paradigm of the kind of generalization one can strive for *within* a micro-world all right, but for that very reason it provides





Figure

no way of thinking about general intelligent systems. In the light of these later evaluations my assumption that work in the early seventies did not refute my accusation of stagnation seems vindicated.

The nongeneralizable character of the programs so far discussed makes them engineering feats, not steps toward generally intelligent systems, and they are, therefore not at all promising as contributions to psychology. Yet Winston includes Waltz's work in his claim that " . . . making machines see is an important way to understand how we animals see. . . ."[37] and Winograd makes similar claims for the psychological relevance of his work:

> The gain from developing AI is not primarily in the usefulness of the programs we create, but in the set of concepts we develop, and the ways in which we can apply them to understanding human intelligence.[38]

These comments suggest that in the early seventies an interesting change was taking place at M.I.T. In previous papers Minsky and his co-workers sharply distinguished themselves from workers in Cognitive Simulation, such as Simon, who presented their programs as psychological theories, insisting that the M.I.T. programs were "an attempt to build intelligent machines without any prejudice toward making the system . . . humanoid."[39] Now in their book, *Artificial Intelligence,*[40] a summary of work done at M.I.T. during the period 1967–1972, Minsky and Papert present the M.I.T. research as a contribution to psychology. They first introduce the notion of a symbolic description:

> What do we mean by "description"? We do not mean to suggest that our descriptions must be made of strings of ordinary-language words (although they might be). The simplest kind of description is a structure in which some features of a situation are represented by single ("primitive") symbols, and relations between those features are represented by other symbols—or by other features of the way the description is put together.[41]

They then defend the role of symbolic descriptions in a psychological account of intelligent behavior by a constant polemic against behaviorism and gestalt theory which have opposed the use of formal models of the mind.

One can detect, underlying this change, the effect of the proliferation of micro-worlds, with their reliance on symbolic descriptions, and the disturbing failure to produce even the hint of a system with the flexibility

of a six-month-old child. Instead of concluding from this frustrating situation that the special-purpose techniques which work in context-free, gamelike, micro-worlds may in no way resemble general-purpose human and animal intelligence, the AI workers seem to have taken the less embarrassing if less plausible tack of suggesting that even if they could not succeed in building intelligent systems, the *ad hoc* symbolic descriptions successful in micro-world analysis could be justified as a valuable contribution to psychology.

Such a line, however, since it involves a stronger claim than the old slogan that as long as the machine was intelligent it did not matter at all whether it performed in a humanoid way, runs the obvious risk of refutation by empirical evidence. An information-processing model must be a formal symbolic structure, however, so Minsky and Papert, making a virtue of necessity, revive the implausible intellectualist position according to which concrete perception is assimilated to the rule-governed symbolic descriptions used in abstract thought.

> The Gestaltists look for simple and fundamental principles about how perception is organized, and then attempt to show how symbolic reasoning can be seen as following the same principles, while we construct a complex theory of how knowledge is applied to solve intellectual problems and then attempt to show how the symbolic description that is what one "sees" is constructed according to similar processes.[42]

Some recent work in psychology, however, points in the exactly opposite direction. Rather than showing that perception can be analyzed in terms of formal features, Erich Goldmeier's extention of early Gestalt work on the perception of similarity of simple perceptual figures—arising in part in response to "the frustrating efforts to teach pattern recognition to [computers]"[43]—has revealed sophisticated distinctions between figure and ground, matter and form, essential and accidental aspects, norms and distortions, etc., which he shows cannot be accounted for in terms of any known formal features of the phenomenal figures. They can, however, according to Goldmeier, perhaps be explained on the neurological level, where the importance of Prägnanz—i.e., singularly salient shapes and orientations—suggests underlying physical phenomena such as "regions of resonance"[44] in the brain.

Recent work in neurophysiology has suggested new mechanisms which might confirm the Gestaltist's intuition that other sorts of process than the manipulation of formal representations of the sort required by digital computers underlie perception. While still nothing definite is known about how the brain "processes information," computer models look even less likely now than in 1970, while models based on the properties of optical holograms look perhaps more promising. As John Haugeland summarizes the evidence:

> First, [optical holograms] are prepared from the light bouncing off an ordinary object, and can subsequently be used to reconstruct a full three-dimensional colored image of that object. Second, the whole image can be reconstructed from any large enough portion of the hologram (i.e., there's no saying which portion of the hologram "encodes" which portion of the image). Third, a number of objects can be separately recorded on the same hologram, and there's no saying which portion records which object. Fourth, if a hologram of an arbitrary scene is suitably illuminated with the light from a reference object, bright spots will appear indicating (virtually instantaneously) the presence and location of any occurrences of the reference object in the scene (and dimmer spots indicate "similar" objects). So some neurophysiological holographic encoding might account for a number of perplexing features of visual recall and recognition, including their speed, some of their invariances, and the fact that they are only slightly impaired by large lesions in relevant areas of the brain. . . .
>
> Another interesting property of optical holograms is that if a hologram [combining light from two separate] objects is illuminated with the light from one of them, an image of the other (absent) object appears. Thus, such a hologram can be regarded as a kind of "associator" of (not ideas, but) visual patterns. . . .[45]

Haugeland adds:

> . . . Fairly detailed hypothetical models have been proposed for how holograms might be realized in neuronal structures; and there is some empirical evidence that some neurons behave in ways that would fit the models.[46]

Of course, it is still possible that the Gestaltists went too far in trying to assimilate thought to the same sort of concrete, holistic, processes they found necessary to account for perception. Thus, even though the exponents of symbolic descriptions have no account of perceptual processes, they might be right that the mechanism of everyday thinking and learning consists in constructing a formal description of the world and transforming this representation in a rule-governed way. Such a formal model of everyday learning and categorization is proposed by Winston in his 1970 thesis, "Learning Structural Descriptions from Examples."[47] Given a set of positive and negative instances, Winston's self-proclaimed "classic" program can, for example, use a descriptive repertoire to construct a formal description of the class of arches. Since, as we mentioned earlier, Winston's program (along with those of Winograd and Guzman) is often mentioned as a success of the late sixties, we must examine it in detail.

Is this program a plausible general theory of learning? Winston's commitment to a computer model dictates the conclusion that it must be:

> Although this may seem like a very special kind of learning, I think the implications are far ranging, because I believe that learning by examples, learning by being told, learning by imitation, learning by reinforcement and other forms are much like one another. In the literature of learning there is frequently an unstated assumption that these various forms are fundamentally different. But I think the classical boundaries between the various kinds of learning will disappear once superficially different kinds of learning are understood in terms of processes that construct and manipulate descriptions.[48]

Yet Winston's program works only if the "student" is saved the trouble of what Charles Sanders Peirce called abduction, by being "told" a set of context-free features and relations—in this case a list of possible spacial relationships of blocks such as "left-of," "standing," "above," and "supported by"—from which to build up a description of an arch. Minsky and Papert presuppose this preselection when they say that "to eliminate objects which seem atypical . . . the program lists all relationships exhibited by more than half of the candidates in the set."[49] Lurking behind this claim is the supposition that there are only a finite number of relevant features; but without preselected features all objects share an indefinitely large number of relationships. The work of discriminating, selecting, and weighting a limited number of relevant features is the result of repeated experience and is the first stage of learning. But since in Winston's work the programmer selects and preweights the primitives, his program gives us no idea how a computer could make this selection and assign these weights. (In this respect Winston's program shows no

progress beyond Newell, Shaw, and Simon's 1958 proposal; see p. 83 of this book.) Thus the Winston program, like every micro-world program, works only because it has excluded from its task domain the very ability it is supposed to explain.

If not a theory of learning, is Winston's program at least a plausible theory of categorization? Consider again the arch example. Once it has been given what Winston disarmingly calls a "good description"[50] and carefully chosen examples, the program does conclude that an arch is a structure in which a prismatic body is supported by two upright blocks that do not touch each other. But, since arches function in various ways in our everyday activity, there is no reason to suppose that these are the necessary and sufficient conditions for being an arch, or that there are any such defining features. Some prominent characteristics shared by most everyday arches are "helping to support something while leaving an important open space under it," or "being the sort of thing one can walk under and through at the same time." How does Winston propose to capture such contextual characteristics in terms of the context-free features required by his formal representation?

Winston admits that having two supports and a flat top does not begin to capture even the geometrical structure of arches. So he proposes "generalizing the machine's descriptive ability to acts and properties required by those acts"[51] by adding a functional predicate, "something to walk through."[52] But it is not at all clear how a functional predicate which refers to implicit knowledge of the bodily skill of walking through is to be formalized. Indeed, Winston himself provides a *reductio ad absurdum* of this facile appeal to formal functional predicates:

> To a human, an arch may be something to walk through, as well as an appropriate alignment of bricks. And certainly, a flat rock serves as a table to a hungry person, although far removed from the image the word table usually calls to mind. But the machine does not yet know anything of walking or eating, so the programs discussed here handle only some of the physical aspects of these human notions. There is no inherent obstacle forbidding the machine to enjoy functional understanding. It is a matter of generalizing the machine's descriptive ability to acts and properties required by those acts. Then chains of pointers can link TABLE to FOOD as well as to the physical image of a table, and the machine will be perfectly happy to draw up its chair to a flat rock with the human given that there is something on that table which it wishes to eat.[53]

Progress on recognition of arches, tables, etc., must, it seems, either wait until we have captured in an abstract symbolic description much of what human beings implicitly know about walking and eating simply by having a body, or else until computers no longer have to be told what it is to walk and eat, because they have human bodies and appetites themselves!

Despite these seemingly insurmountable obstacles Winston boasts that "there will be no contentment with [concept learning] machines that only do as well as humans."[54] But it is not surprising that Winston's work is nine years old and there has been little progress in machine learning, induction, or concept formation. In their account Minsky and Papert admit that "we are still far from knowing how to design a powerful yet subtle and sensitive inductive learning program."[55] What is surprising is that they add: "but the schemata developed in Winston's work should take us a substantial part of the way."[56] The lack of progress since Winston's work was published, plus the use of predigested weighted primitives from which to produce its rigid, restricted, and largely irrelevant descriptions, makes it hard to understand in what way the program is a substantial step.

Moreover, if Winston claims to "shed some light on [the question:] How do we recognize examples of various concepts?"[57] his theory of concepts as definitions must, like any psychological theory, be subject to empirical test. It so happens that contrary to Winston's claims, recent evidence collected and analyzed by Eleanor Rosch on just this subject shows that human beings are not aware of classifying objects as instances of abstract rules but rather group objects as more or less distant from an imagined paradigm. This does not exclude the possibility of unconscious processing, but it does highlight the fact that there is no empirical evidence at all for Winston's formal model. As Rosch puts it:

> Many experiments have shown that categories appear to be coded in the mind neither by means of lists of each individual member of the category, nor by means

of a list of formal criteria necessary and sufficient for category membership, but, rather, in terms of a prototype of a typical category member. The most cognitively economical code for a category is, in fact, a *concrete image* of an average category member.[58]

One paradigm, it seems, is worth a thousand rules. As we shall soon see, one of the characteristics of the next phase of work in AI is to try to take account of the implications of Rosch's research.

Meanwhile, what can we conclude concerning AI's contribution to the science of psychology? No one can deny Minsky and Papert's claim that "Computer Science has brought a flood of . . . ideas, well defined and experimentally implemented, for thinking about thinking. . . ."[59] But all of these ideas can be boiled down to ways of constructing and manipulating symbolic descriptions, and, as we have seen, the notion that human cognition can be explained in terms of formal representations does not seem at all obvious in the face of actual research on perception, and everyday concept formation. Even Minsky and Papert show a commendable new modesty. They as much as admit that AI is still at the stage of astrology (not unlike alchemy), and that the much heralded breakthrough still lies in the future:

> Just as astronomy succeeded astrology, following Kepler's discovery of planetary regularities, the discoveries of these many principles in empirical explorations of intellectual processes in machines should lead to a science, eventually.[60]

Happily, "should" has replaced "will" in their predictions. Indeed, this period's contribution to psychology suggests an even more modest hope: As more psychologists like Goldmeier are frustrated by the limitations of formal computer models, and others turn to investigating the function of images as opposed to symbolic representations, the strikingly limited success of AI may come to be seen as an important disconfirmation of the information processing approach.

To complete our survey of the state of AI research as it entered its second decade we need to consider briefly the state of robot research, both because work in this area received a lot of misleading publicity during this period and because, as we have just seen in discussing Winston's claims, workers in AI often take refuge in the idea that computers will finally achieve human understanding when they have humanoid bodies.

Our account will have to be brief because there is not much to report. After the usual optimistic start, the M.I.T. robot arm was stopped cold by just the problem of representing its own body space which I suspected would be its undoing (see p. 251). In the 1968–1969 AI Progress Report this problem is clearly an embarrassment:

> . . . [H]ow should one represent a machine's body image? For the problem of a single, not-too-complicated arm, one can doubtless get by with cleverly coded, sparse, three-dimensional arrays, but one would like something more symbolic. And one wonders what happens in the nervous system; we have not seen anything that might be considered to be a serious theory. Consider that a normal human can place an object on a table, turn about and make a gross change in his position and posture, and then reach out and grasp within one or two inches of the object, all with his eyes closed! It seems unlikely that his cerebellum could perform the appropriate vector calculations to do this. . . .[61]

However, rather than see this as evidence that their attempt to represent the robot's arm as one more object in physical space was misguided, the authors of the report get into deeper trouble defending their faith.

> . . . We would presume that this complex motor activity is made up, somehow, of a large library of *stereotypical programs,* with some heuristic interpolation scheme that fits the required action to some collection of *reasonably similar stored actions.* But we have found nowhere any serious proposal about neurological mechanisms for this, and one can hope that some plausible ideas will come out of robotics research itself.[62]

Neurophysiology offers, admittedly speculative, accounts of such similarity, but these are holographic not information processing models. As for the AI approach, it merely raises the further problem of recognizing similarity, which is discussed in connection with chess-playing programs in the next section. In the light of these problems, when the report adds: "Unfortunately, at present this area is somewhat dormant,"[63] we can only take "dormant" as a polite synonym for stagnant or even comatose.

In spite of its better press (see p. 300) the SRI robot, Shakey, was in no better shape. As Bertram Raphael frankly sums up the situation in response to exaggerated coverage by the media:

> . . . Many experiments were performed with Shakey between 1968 and 1972 . . . [but] we made much less progress than various press reports might suggest toward the creation of an independent sentient robot capable of meaningful performance in a normal human environment. Responsible scientists consider this intriguing idea premature, probably by at least several decades.[64]

In effect, Shakey is another case of a micro-world success which turned into a real-world failure.

> At his peak, Shakey could only function in a sterile "play-pen" environment of walls, doorways, carefully painted baseboards (so he could "see" where the walls met the floor), and a few simply-shaped wooden blocks; he had only about a dozen pre-programmed "instinctive" abilities, such as TURN, PUSH, GO-THROUGH-DOORWAY, and CLIMB-RAMP, which could be combined in various ways by the planning programs. . . . The scientists who worked on Shakey developed a deep appreciation of how difficult it is to produce a robot even with relatively trivial abilities, let alone the true science-fiction-like independent competence.[65]

According to Raphael, Shakey and the SRI robot project have been "temporarily put aside" and there will be no interesting robot work to report until AI workers solve the basic problem of knowledge representation:

> Surprisingly, the issues of how to acquire, represent, and make use of a broad store of knowledge has been the most neglected part of past robot research. The developers of the laboratory robot systems were so busy patching together existing capabilities (in vision, language, and problem solving), and filling in essential new areas (representing the physical world, providing for error recovery), that they did not attend to the fundamental issue of knowledge structures.[66]

So now we have the overall picture. In all those areas where enthusiasts saw signs of success at just the time this book appeared—language understanding, scene analysis, concept learning, and robot building—the work turned out to be based on brilliant but *nongeneralizable* exploitation of specific features of the task domain. With this realization AI finally had to face the problem of representing everyday knowledge—a difficult, decisive, and philosophically fascinating task with which it is still struggling today.

Phase IV (1972–1977) Facing the Problem of Knowledge Representation

As the restricted interest of work in restricted domains became apparent, the distinction between specific applications and research on basic principles became sharper. Feigenbaum comes to refer to his work on DENDRAL and his more recent program for inferring the rules of mass spectometry, META-DENDRAL, as "knowledge engineering"[67] while Winograd and his associates call their work "cognitive science."[68]* At M.I.T., a grant proposal from this period distinguishes between "no-holds-barred, special purpose, domain-dependent work" and "no-tricks basic study."[69] And it seems to be generally accepted that every program we discussed in Phase III, and, indeed, the whole micro-world concept, was in this straightforward sense, a trick.

We shall now see that in Phase IV the special-purpose work makes steady progress, while the basic study faces a crisis. Everyday human know-how is increasingly acknowledged to be presupposed by intelligent behavior, yet it turns out to be incredibly difficult, perhaps in principle impossible, to program.

The areas in which knowledge engineering has been successful are just those in which the first edition of *What Computers Can't Do* predicted that progress could be expected. (See Column III, of my breakdown of the field, p. 292.) As long as the domain in question can be treated as a game, i.e., as long as what is relevant is fixed, and the possibly relevant factors can be defined in terms of context-free primitives, then computers can do well in the domain. And they will do progressively better relative to people as the amount of domain-specific knowledge required is increased. In such special-purpose programs the form of knowledge representation can be limited to situation → action rules in which the situation is defined in terms of a few parameters and indicates the conditions

under which a specific heuristic rule is relevant. Again, because relevance is defined beforehand, reasoning can be by inference chains with no need for reasoning by analogy.

All these features can be found in one of the most impressive practical programs to date: Shortliffe's MYCIN program (1976) for diagnosing blood infections and meningitis infections and recommending drug treatment. The rules in this case are of the form:

RULE 85

IF:

1. The site of the culture is blood, and
2. The gram stain of the organism is gramneg, and
3. The morphology of the organism is rod, and
4. The patient is a compromised host

THEN:

There is suggestive evidence (.6) that the identity of the organism is pseudomonas-aeruginosa.[70]

The program has been tested by a panel of judges:

. . . In 90% of the cases submitted to the judges, a majority of the judges said that the program's decisions were the-same-as or as-good-as the decisions they would have made.[71]*

This approach, although successful as an engineering feat, involves several assumptions which may conceal potential limitations. Feigenbaum, in his analysis of MYCIN, assumes that acquiring expert skill is acquiring rules for recognizing situations and rules for evaluating evidence.

. . . In most "crafts or branches of learning" what we call "expertise" is the essence of the art. And for the domains of knowledge that we touch with our art, it is the "rules of expertise" or the rules of "good judgment" of the expert practitioners of the domain that we seek to transfer to our programs.[72]

He conscientiously notes that the experts themselves are not aware of using rules:

. . . Experience has also taught us that much of this knowledge is private to the expert, not because he is unwilling to share publicly how he performs, but because he is unable. He knows more than he is aware of knowing. (Why else is the Ph.D. or the Internship a guild-like apprenticeship to a presumed "master of the craft"? What the masters really know is not written in the textbooks of the masters.)[73]

But Feigenbaum with his assumption that expert performance must result exclusively from following rules, is nonetheless convinced that by suitable questioning he can get the expert, as Plato would say, to "recollect" the complete set of unconscious heuristics:

. . . But we have learned also that this private knowledge can be uncovered by the careful, painstaking analysis of a second party, or sometimes by the expert himself, operating in the context of a large number of highly specific performance problems.[74]

If internship and the use of examples play an *essential* role in expert judgment, i.e., if there is a limit to what can be understood by rules, Feigenbaum would never see it—especially in domains such as medicine where there is a very large and rapidly increasing body of factual information concerning drugs and their side effects and interactions, so that the computer can make up in data-processing capacity for what it lacks in judgment. Yet, the fact remains that in each field where "knowledge engineering" has made its valuable contribution and rivaled the experts, there are still masters who do better than the machine. To determine whether this is an accident, or whether skill may involve more than rule following, it is helpful to look at developments in chess, where the domain is restricted, factual knowledge is at a minimum, and where we have some psychological evidence of what master players actually do.

Chess is an ideal micro-world in which relevance is restricted to the narrow domain of the kind of chess piece (pawn, knight, etc.), its color, and the position of the piece on the board. But while the game's circumscribed character makes a world champion chess program in principle possible, there is a great deal of evidence that human beings play chess quite differently from computers; and I was not surprised to find that up to 1971 computers played fairly low-level chess (see pp. 82–85). In July 1976, however, the Northwestern University chess program, called

CHESS 4.5, won the class B section of the Paul Masson American Chess Championship with an impressive 5 wins and no losses. It then went on in February 1977 to win the 84th Minnesota Open Tournament against experts and high-class A players.[75] Such unexpected impressive results require a reexamination of the difference between human and computer chess playing.

A chess program has the sort of situation → action rules discussed above. A situation is characterized in terms of context-free features: the position and color of each piece on the board. All possible legal moves and the positions which result are then defined in terms of these features. To evaluate and compare positions, rules are provided for calculating scores on attributes such as "material balance" (where a numerical value is assigned to each piece on the board and the total score is computed for each player), or "center control" (where the number of pieces bearing on each centrally located square is counted). Finally, there must be a formula for evaluating alternative positions on the basis of these scores. Using this approach and looking at a tree of around 3 million potential positions CHESS 4.5 can beat some players at the expert level, but a chess master generally looks at the results of less than 100 possible moves (see p. 102) and yet plays a far better game. How can this be?

In Chapter 1, I note that human beings avoid the counting out of large numbers of alternatives characteristic of a computer program by "zeroing in" on the appropriate area in which to look for a move and I suggest that this ability is the result of having a sense of the developing game. While no doubt correct, this now seems to me an inadequate account, for it does not take into consideration the fact that to develop this ability to zero in, chess masters must play thousands of actual and book games. What does this apprenticeship add to their skill?

By playing over book games chess masters presumably develop the ability to recognize present positions as similar to positions which occurred in classic games. These previous positions have already been analyzed in terms of their significant aspects. Aspects of a chess position include such overall characteristics as "control of the situation" (the extent to which a player's opponent's moves can be forced by making threatening moves), "crampedness of the position" (the amount of freedom of maneuver inherent in both the player's position and the opponent's position), or "overextendedness" (the fact that while the position might be superficially quite strong, one is not in sufficient control of the situation to follow through and, with correct play by the opponent, a massive retreat will be required). The already analyzed remembered positions focus the player's attention on critical aspects of the current position, and the master can thus zero in on these critical areas before beginning to count out specific moves.

The distinction between features and aspects is central here. *Aspects* play a role in an account of human play similar to that of *features* in the computer model, but there is a crucial difference. In the computer model the *situation is* DEFINED IN TERMS OF *the features,* whereas in human play *situational understanding is* PRIOR TO *aspect specification.* For example, the numerical value of a feature such as material balance can be calculated independently of any understanding of the game, whereas an aspect like overextendedness cannot be calculated simply in terms of the position of the pieces, since the same board position can have different aspects depending on its place in the long-range strategy of a game. In a game in which white's long-range strategy is an attack on the opponent's king, the advanced position of white's pieces does not constitute overextension, whereas otherwise it would. No present or envisaged chess program attempts to include such long-range strategy, yet to recognize aspects requires some such overall interpretation of the game.

For the same reason some sort of *feature-based* matching of the present position against a stored library of previous positions won't help account for a master player's ability to use past experience to zero in. It is astronomically unlikely that two positions will ever turn out to be *identical,* so that what has to be compared are *similar* positions. But similarity cannot be defined as having a large number of pieces on identical squares. Two positions which are identical except for one pawn moved to an adjacent square can be totally different, while two positions can be similar although no pieces are on the same square in each. Thus similarity depends on the player's sense of the issues at stake, not merely on the position of the pieces. Seeing two positions as similar is exactly what requires a deep understanding of the game. And structuring the situation in terms of aspects of remembered similar situations in turn

enables the human player to avoid the massive counting out required when the positions are characterized only in terms of context-free features.

Aspects also enable masters to formulate heuristic *maxims* which play a role in this account analogous to heuristic *rules* in the computer model. Polanyi calls attention to the difference between strict rules and maxims:

> Maxims are rules, the correct application of which is part of the art which they govern. The true maxims of golfing or of poetry increase our insight into golfing or poetry and may even give valuable guidance to golfers and poets; but these maxims would instantly condemn themselves to absurdity if they tried to replace the golfer's skill or the poet's art. Maxims cannot be understood, still less applied by anyone not already possessing a good practical knowledge of the art.[76]

At present computers using exhaustive search, and masters using selective search guided by aspect analysis and maxims, can each look ahead about six or seven ply.[77]* Given the exponential growth of alternative moves it will not be possible without better tree-searching heuristics to significantly increase the computer's power to look ahead. Thus with present programs what is really at stake is how far computers which must use tactics based on context-free features can make up by sheer brute force for the use of long-range strategy, the recognition of similarity to other preanalyzed games, and the zeroing in on crucial aspects characteristic of advanced human play.

In general being able to see similarity to prototypical cases and to recognize shared aspects in terms of this similarity, as well as the possibility of profiting from maxims formulated in terms of these aspects, all seem to play an essential role in the acquisition and utilization of expertise. But since these abilities are not based on context-free features but depend on the overall situation they cannot be captured in the situation → action rule formalism. Thus we can expect in every area where expertise is based on experience to continue to find some experts who outperform even the most sophisticated programs.

Although chess programs and knowledge engineering in general have made remarkable progress during the past two years, discourse understanding, despite the introduction of interesting new ideas, is still in the

same state of stagnation as it was in 1972. While this has led some researchers to ever more extravagant promises and claims, it has led others to sober thoughts on the difficulty of programming human understanding. In order to form a reasonable opinion about what has yet to be done to make computers intelligent, we must turn from the computer's successes in restricted domains to the stag/flation afflicting the field of discourse understanding.

The difference between programs like MYCIN and CHESS 4.5, and programs for understanding discourse, is precisely the difference between domain-specific knowledge and general intelligence; between anything-goes engineering and no-tricks basic study; or, as we can now see, the difference between areas in which relevance has been decided beforehand (Area III in my chart, p. 292), and areas in which determining what is relevant is precisely the problem (Area IV).

In the past five years, the problem of how to structure and retrieve data in situations when anything might be relevant has come to be known as the knowledge representation problem. As Patrick Winston, head of the M.I.T. AI Laboratory, puts it in a section of a 1975 research proposal entitled "The Need for Basic Studies":

> . . . We believe that proper representation is the key to advanced vision, common sense reasoning, and expert problem solving, just as it is to many other aspects of Artificial Intelligence.[78]

Of course, the representation of knowledge was always a central problem for work in AI, but earlier periods were characterized by an attempt to repress it by seeing how much could be done with as little knowledge as possible. Now, the difficulties are being faced. As Roger Schank of Yale recently remarked:

> . . . Researchers are starting to understand that tour-de-forces in programming are interesting but non-extendable . . . the AI people recognize that how people use and represent knowledge is the key issue in the field. . . .[79]

Papert and Goldstein explain the problem:

> It is worthwhile to observe here that the goals of a knowledge-based approach to AI are closely akin to those which motivated Piaget to call . . . himself an "epistemologist" rather than a psychologist. The common theme is the view that

the process of intelligence is determined by the knowledge held by the subject. The deep and primary questions are to understand the operations and data structures involved.[80]

Another memo illustrates how ignoring the background knowledge can come back to haunt one of AI's greatest tricks in the form of nongeneralizability:

> . . . Many problems arise in experiments on machine intelligence because things obvious to any person are not represented in any programs. One can pull with a string, but one cannot push with one. One cannot push with a thin wire, either. A taut inextensible cord will break under a very small lateral force. Pushing something affects first its speed; only indirectly its position! Simple facts like these caused serious problems when Charniak attempted to extend Bobrow's "Student" program to more realistic applications, and they have not been faced up to until now.[81]

The most interesting current research is directed toward the underlying problem of developing new, flexible, complex data types which will allow the representation of background knowledge in large, more structured units.

In 1972, drawing on Husserl's phenomenological analysis, I pointed out that it was a major weakness of AI that no programs made use of expectations (see pp. 241, 242, and 250). Instead of modeling intelligence as a passive receiving of context-free facts into a structure of already stored data, Husserl thinks of intelligence as a context-determined, goal-directed activity—as a *search* for anticipated facts. For him the noema, or mental representation of any type of object, provides a context or "inner horizon" of expectations or "predelineations" for structuring the incoming data: a "rule governing *possible* other consciousness of [the object] as identical—possible, as exemplifying essentially predelineated types."[82]* As I explain in Chapter 7:

> . . . We perceive a house, for example, as more than a façade—as having some sort of back—some inner horizon. We respond to this whole object first and then, as we get to know the object better, fill in the details as to inside and back. [p. 241]

The noema is thus a symbolic description of all the features which can be expected with certainty in exploring a certain type of object—features

which remain "inviolably the same: as long as the objectivity remains intended as *this* one and of this kind"[83] . . . plus "predelineations" of those properties which are possible but not necessary features of this type of object.

A year after my objection, Minsky proposed a new data structure remarkably similar to Husserl's for representing everyday knowledge:

> A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. . . .
>
> We can think of a frame as a network of nodes and relations. The "top levels" of a frame are fixed, and represent things that are always true about the supposed situation. The lower levels have many *terminals*—"slots" that must be filled by specific instances or data. Each terminal can specify conditions its assignments must meet. . . .
>
> Much of the phenomenological power of the theory hinges on the inclusion of expectations and other kinds of presumptions. A frame's terminals are normally already filled with "default" assignments.[84]

In Minsky's model of a frame, the "top level" is a developed version of what in Husserl's terminology "remains inviolably the same" in the representation, and Husserl's predelineations have been made precise as "default assignments"—additional features that can normally be expected. The result is a step forward in AI techniques from a passive model of information processing to one which tries to take account of the context of the interactions between a knower and his world. Husserl thought of his method of transcendental-phenomenological constitution, i.e., "explicating" the noema for all types of objects, as the beginning of progress toward philosophy as a rigorous science, and Patrick Winston has hailed Minsky's proposal as "the ancestor of a wave of progress in AI."[85] But Husserl's project ran into serious trouble and there are signs that Minsky's may too.

During twenty years of trying to spell out the components of the noema of everyday objects, Husserl found that he had to include more and more of what he called the "outer horizon," a subject's total knowledge of the world:

> . . . To be sure, even the tasks that present themselves when we take single types of objects as restricted clues prove to be extremely complicated and always lead to extensive disciplines when we penetrate more deeply. That is the case, for

example, with a transcendental theory of the constitution of a spatial object (to say nothing of a Nature) as such, of psycho-physical being and humanity as such, cultures as such.[86]

He sadly concluded at the age of seventy-five that he was "a perpetual beginner" and that phenomenology was an "infinite task"—and even that may be too optimistic. His successor, Heidegger, pointed out that since the outer horizon or background of cultural practices was the condition of the possibility of determining relevant facts and features and thus prerequisite for structuring the inner horizon, as long as the cultural context had not been clarified the proposed analysis of the inner horizon of the *noema* could not even claim progress.

There are hints in an unpublished early draft of the frame paper that Minsky has embarked on the same misguided "infinite task" that eventually overwhelmed Husserl:

> Just constructing a knowledge base is a major intellectual research problem. . . . We still know far too little about the contents and structure of common-sense knowledge. A "minimal" common-sense system must "know" something about cause-effect, time, purpose, locality, process, and types of knowledge. . . . We need a serious epistemological research effort in this area.[87]

Minsky's naïveté and faith are astonishing. Philosophers from Plato to Husserl, who uncovered all these problems and more, have carried on serious epistemological research in this area for two thousand years without notable success. Moreover, the list Minsky includes in this passage deals only with natural objects, and their positions and interactions. As Husserl saw, and as I argue in Chapter 8, intelligent behavior also presupposes a background of cultural practices and institutions. Observations in the frame paper such as:

> Trading normally occurs in a social context of law, trust, and convention. Unless we also represent these other facts, most trade transactions will be almost meaningless[88]

show that Minsky has understood this too. But Minsky seems oblivious to the hand-waving optimism of his proposal that programmers rush in where philosophers such as Heidegger fear to tread, and simply make explicit the totality of human practices which pervade our lives as water encompasses the life of a fish.

To make this essential point clear it helps to take an example used by Minsky and look at what is involved in understanding a piece of everyday equipment as simple as a chair. No piece of equipment makes sense by itself. The physical object which is a chair can be defined in isolation as a collection of atoms, or of wood or metal components, but such a description will not enable us to pick out chairs. What makes an object a *chair* is its function, and what makes possible its role as equipment for sitting is its place in a total practical context. This presupposes certain facts about human beings (fatigue, the ways the body bends), and a network of other culturally determined equipment (tables, floors, lamps), and skills (eating, writing, going to conferences, giving lectures, etc.). Chairs would not be equipment for sitting if our knees bent backwards like those of flamingos, or if we had no tables as in traditional Japan or the Australian bush.

Anyone in our culture understands such things as how to sit on kitchen chairs, swivel chairs, folding chairs; and in arm chairs, rocking chairs, deck chairs, barber's chairs, sedan chairs, dentist's chairs, basket chairs, reclining chairs, wheel chairs, sling chairs, and beanbag chairs—as well as how to get out of them again. This ability presupposes a repertoire of bodily skills which may well be indefinitely large, since there seems to be an indefinitely large variety of chairs and of successful (graceful, comfortable, secure, poised, etc.) ways to sit in them. Moreover, understanding chairs also includes social skills such as being able to sit appropriately (sedately, demurely, naturally, casually, sloppily, provocatively, etc.) at dinners, interviews, desk jobs, lectures, auditions, concerts (intimate enough for there to be chairs rather than seats), and in waiting rooms, living rooms, bedrooms, courts, libraries, and bars (of the sort sporting chairs, not stools).

In the light of this amazing capacity, Minsky's remarks on chairs in his frame paper seem more like a review of the difficulties than even a hint of how AI could begin to deal with our commonsense understanding in this area.

> There are many forms of chairs, for example, and one should choose carefully the chair-description frames that are to be the major capitols of chair-land. These are used for rapid matching and assigning priorities to the various differences. The lower priority *features* of the *cluster* center then serve . . . as properties of the chair *types*. . . .[89]

There is no argument why we should expect to find elementary context-free *features* characterizing a chair *type,* nor any suggestion as to what these features might be. They certainly cannot be legs, back, seat, etc., since these are not context-free characteristics defined apart from chairs which then "cluster" in a chair representation, but rather legs, back, etc. come in all shapes and variety and can only be recognized as *aspects* of already recognized chairs. Minsky continues:

> Difference pointers could be "functional" as well as geometric. Thus, after rejecting a first try at "chair" one might try the functional idea of "something one can sit on" to explain an unconventional form.[90]

But, as we already saw in our discussion of Winston's concept-learning program, a function so defined is not abstractable from human embodied know-how and cultural practices. A functional description such as "something one can sit on" treated merely as an additional context-free descriptor cannot even distinguish conventional chairs from saddles, thrones, and toilets. Minsky concludes:

> Of course, that analysis would fail to capture toy chairs, or chairs of such ornamental delicacy that their actual use would be unthinkable. These would be better handled by the method of excuses, in which one would bypass the usual geometrical or functional explanation in favor of responding to *contexts* involving *art* or *play.*[91]

This is what is required all right, but by what elementary features are *these* contexts to be recognized? There is no reason at all to suppose that one can avoid the difficulty of formally representing our knowledge of chairs by abstractly representing even more holistic, concrete, culturally determined, and loosely organized human practices such as art and play.

Minsky in his frame article claims that: "the frame idea . . . is in the tradition of . . . the 'paradigms' of Kuhn,"[92] so it is appropriate to ask whether a theory of formal representation such as Minsky's, even if it can't account for everyday objects like chairs, can do justice to Thomas Kuhn's analysis of the role of paradigms in the practice of science. Such a comparison might seem more promising than testing the ability of frames to account for our everyday understanding, since science is a theoretical enterprise which deals with context-free data whose lawlike

relations can in principle be grasped by any sufficiently powerful "pure-intellect," whether human, Martian, digital, or divine.

Paradigms, like frames, serve to set up expectations. As Kuhn notes: "In the absence of a paradigm or some candidate for paradigm, all the facts that could possibly pertain to the development of a given science are likely to seem equally relevant."[93] Minsky interprets as follows:

> According to Kuhn's model of scientific evolution 'normal' science proceeds by using established *descriptive schemes.* Major changes result from new 'paradigms', new ways of describing things. Whenever our customary viewpoints do not work well, whenever we fail to find effective frame systems in memory, we must construct new ones that bring out the right *features.*[94]

But what Minsky leaves out is precisely Kuhn's claim that a paradigm or exemplar is *not* an *abstract explicit descriptive scheme* utilizing formal *features,* but rather a shared *concrete case,* which dispenses with features altogether:

> The practice of normal science depends on the ability, acquired from exemplars, to group objects and situations into similarity sets which are primitive in the sense that the grouping is done without an answer to the question, "Similar with respect to what?"[95]

Thus, although it is the job of scientists to find abstractable, exact, symbolic descriptions, and *the subject matter of science* consists of such formal accounts, the *thinking* of scientists themselves does not seem to be amenable to this sort of analysis. Kuhn explicitly repudiates any formal reconstruction which claims that the scientists must be using symbolic descriptions:

> I have in mind a manner of knowing which is misconstrued if reconstructed in terms of rules that are first abstracted from exemplars and thereafter function in their stead.[96]

Indeed, Kuhn sees his book as raising just those questions which Minsky refuses to face:

> Why is the *concrete* scientific achievement, as a locus of professional commitment, prior to the various concepts, laws, theories, and points of view that may be *abstracted* from it? In what sense is the shared paradigm a fundamental unit

for the student of scientific development, a unit that *cannot* be fully reduced to logically *atomic components* which might function in its stead?"

Although research based on frames cannot deal with this question and so cannot account for commonsense or scientific knowledge, the frame idea did bring the problem of how to represent our everyday knowledge into the open in AI. Moreover, it provided a model so vague and suggestive that it could be developed in several different directions. Two alternatives immediately presented themselves: either to use frames as part of a special-purpose micro-world analysis dealing with commonsense knowledge as if everyday activity took place in preanalyzed specific domains, or else to try to use frame structures in "a no-tricks basic study" of the open-ended character of everyday know-how. Of the two most influential current schools in AI, Roger Schank and his students at Yale have tried the first approach, Winograd, Bobrow, and their research group at Stanford and Xerox, the second.

Schank's version of frames are called "scripts." Scripts encode the essential steps involved in stereotypical social activities. Schank uses them to enable a computer to "understand" simple stories. Like the micro-world builders of Phase III, Schank believes he can start with isolated stereotypical situations described in terms of primitive actions and gradually work up from there to all of human life.

To carry out this project, Schank invented an event description language consisting of eleven primitive acts such as: ATRANS—the transfer of an abstract relationship such as possession, ownership, or control; PTRANS—the transfer of physical location of an object; INGEST—the taking of an object by an animal into the inner workings of that animal, etc.,[98] and from these primitives he builds gamelike scenarios which enable his program to fill in gaps and pronoun reference in stories.

Such primitive acts, of course, make sense only when the context is already interpreted in a specific piece of discourse. Their artificiality can easily be seen if one compares one of Schank's context-free primitive acts to real-life actions. Take PTRANS, the transfer of physical location of an object. At first it seems an interpretation-free fact if ever there was one. After all, either an object moves or it doesn't. But in real life things

are not so simple; even what counts as physical motion depends on our purposes. If someone is standing still in a moving elevator on a moving ocean liner, is his going from A to B deck a PTRANS? What about when he is just sitting on B deck? Are we all PTRANSing around the sun? Clearly the answer depends on the situation in which the question is asked.

Such primitives can, however, be used to describe fixed situations or scripts once the relevant purposes have already been agreed upon. Schank's definition of a script emphasizes its predetermined, bounded, gamelike character:

> We define a script as a *predetermined* causal chain of conceptualizations that describe the *normal sequence of things* in a familiar situation. Thus there is a restaurant script, a birthday-party script, a football game script, a classroom script, and so on. Each script has in it a *minimum number of players* and objects that assume certain roles within the script . . . [E]ach *primitive* action given stands for the most important *element* in a *standard set* of actions."[99]

His illustration of the restaurant script spells out in terms of primitive actions the rules of the restaurant game:

Script: restaurant
Roles: customer; waitress; chef; cashier
Reason: to get food so as to go down in hunger and up in pleasure

Scene 1 entering

PTRANS—go into restaurant
MBUILD—find table
PTRANS—go to table
MOVE—sit down

Scene 2 ordering

ATRANS—receive menu
ATTEND—look at it
MBUILD—decide on order
MTRANS—tell order to waitress

Scene 3 eating

ATRANS—receive food
INGEST—eat food

Scene 4 exiting

MTRANS—ask for check
ATRANS—give tip to waitress
PTRANS—go to cashier
ATRANS—give money to cashier
PTRANS—go out of restaurant[100]

No doubt many of our social activities are stereotyped and there is nothing in principle misguided in trying to work out primitives and rules for a restaurant game, the way the rules of Monopoly are meant to capture a simplified version of the typical moves in the real estate business. But Schank claims that he can use this approach to understand stories about *actual* restaurant-going—that in effect he can treat the sub-world of restaurant going as if it were an isolated micro-world. To do this, however, he must artificially limit the possibilities; for, as one might suspect, no matter how stereotyped, going to the restaurant is not a self-contained game but a highly variable set of behaviors which open out into the rest of human activity. What "normally" happens when one goes to a restaurant can be preselected and formalized by the programmer as default assignments, but the background has been left out so that a program using such a script cannot be said to understand going to a restaurant at all. This can easily be seen by imagining a situation that deviates from the norm. What if when one tries to order he finds that the item in question is not available, or before paying he finds that the bill is added up wrongly? Of course, Schank would answer that he could build these normal ways restaurant-going breaks down into his script. But there are always *abnormal* ways everyday activities can break down: the juke box might be too noisy, there might be too many flies on the counter, or as in the film *Annie Hall,* in a New York delicatessen one's girl friend might order a pastrami sandwich on white bread with mayonnaise. When we understand going to a restaurant we understand how to cope with even these abnormal possibilities because going to a restaurant is part of our everyday activities of going into buildings, getting things we want, interacting with people, etc.

To deal with this sort of objection Schank has added some general rules for coping with unexpected disruptions. The general idea is that in a story "it is usual for non-standard occurrences to be explicitly mentioned"[101] so the program can spot the abnormal events and understand the subsequent events as ways of coping with them. But here we can see that dealing with stories allows Schank to bypass the basic problem, since it is the *author's* understanding of the situation which enables him to decide which events are disruptive enough to mention.

This *ad hoc* way of dealing with the abnormal can always be revealed by asking further questions, for the program has not understood a restaurant story the way people in our culture do, until it can answer such simple questions as: When the waitress came to the table did she wear clothes? Did she walk forward or backward? Did the customer eat his food with his mouth or his ear? If the program answers, "I don't know," we feel that all of its right answers were tricks or lucky guesses and that it has not understood anything of our everyday restaurant behavior.[102*] The point here, and throughout, is not that there are subtle things human beings can do and recognize which are beyond the low-level understanding of present programs, but that in any area there are simple taken-for-granted responses central to human understanding, lacking which a computer program cannot be said to have any understanding at all.

Schank's claim, then, that "the paths of a script are the possibilities that are extant in a situation"[103] is insidiously misleading. Either it means that the script accounts for the possibilities in the restaurant game defined by Schank, in which case it is true but uninteresting; or he is claiming that he can account for the possibilities in an everyday restaurant situation which is impressive but, by Schank's own admission, false.

Real short stories pose a further problem for Schank's approach. In a script what the primitive actions and facts are is determined beforehand, but in a short story *what counts as the relevant facts depends on the story itself.* For example, a story which describes a bus trip contains in its script that the passenger thanks the driver (a Schank example). But the fact that the passenger thanked the driver would not be important in a story in which the passenger simply took the bus as a part of a longer journey, while it might be crucially important if the story concerned a misanthrope who had never thanked anyone before, or a very law-abiding young man who had courageously broken the prohibition against

speaking to drivers in order to speak to the attractive woman driving the bus. Overlooking this point, Schank claimed at a recent meeting that his program which can extract death statistics from newspaper accident reports had answered my challenge that a computer would count as intelligent only if it could summarize a short story.[104] But Schank's newspaper program cannot provide a clue concerning judgments of what to include in a story summary because it works only where relevance and significance have been predetermined, and thereby avoids dealing with the world built up in a story in terms of which judgments of relevance and importance are made.

Another way to see that script analysis of story understanding leaves out something essential is to consider the question: In reading a story how do we call up the appropriate script? In discussing this question Schank points out:

> . . . While the restaurant script can be a subpart of a larger script (such as $TRIP) [In Schank's notation the dollar sign indicates a script.] it must be marked as not being capable of being subsumed by $DELIVERY.[105]

But this "solution" raises the problem of negative information which dogs a proposal like Schank's. It seems implausible to suppose that one could mark the restaurant script as *not* subsumed under such other scripts as making a phone call, answering a call for help, retrieving a lost object, looking for a job, getting signatures for a petition, repairing equipment, coming to work, doing an inspection, leaving a bomb, arranging a banquet, collecting for the Mafia, looking for change for the meter, buying cigarettes, hiding from the police, etc., etc., which might lead one to enter a restaurant *without* intending to eat. It would be more manageable to write a program which, *whenever* someone in a story enters a restaurant, follows the restaurant script until the understander's expectations fail to be fulfilled. Presumably because he thinks of his programs as having psychological reality, Schank neglects this alternative, and on this point he is right. Normally in reading a story we do not suppose that a person who enters a restaurant for a purpose that does not involve eating is preparing to eat; so we do not have to be jolted out of this hypothesis by the fact that the waitress does not bring him a menu. But

Schank's proposal leaves completely unanswered the problem of how we *do* choose the right script.

Schank's latest book does have some interesting ideas about how to go beyond scripts, since he readily admits that much of our everyday activities is not scripted. He introduces "plans" as our way of dealing with stories about situations which don't have fixed scripts. And he points out that plans are made up of subplans or planboxes, which are useful in many situations. For example,

> one kind of instrumental goal is a general building block in many planning processes. In a plan for satisfying hunger, one of the crucial steps is to go to where food is. Going to an intended location is a very general process, useful in all sorts of specific plans.[106]

> Thus a planbox is used whenever no script is available. If a planbox is used often enough, it will generate a script that eliminates the need for the planbox *as long as the surrounding context stays the same.*[107]

But here the persistent problem of recognizing similarity again arises. How can we tell whether the surrounding context is the same? It won't be identical, and Schank gives us no theory of how to recognize contexts as similar.

Finally, Schank has to deal with the short-term goals which motivate everyday plans, the long range goals which generate the short term ones, and the life themes, in terms of which people organize their goal-oriented activities.

> . . . The expectations that we generate from themes are an important part of understanding stories because they generate the goals that generate the plans that we expect to be carried out.[108]

Here Schank has to face the important way desires, emotions, and a person's interpretation of what it means to be a human being open up endless possibilities for human life. If the themes which organize our lives turn out to be unprogrammable Schank is in trouble and so is all of AI. But Schank again imperturbably uses his engineering approach and starts making lists of life themes. This leads to what would seem to be an in-principle problem:

Because life themes are continuous goal generators, it is not really possible to delimit a set of possible life themes. There are as many life themes as there are possible long term goals.[109]

But Schank passes over this difficulty, as he does all others, by stipulating a few more *ad hoc* primitives.

> . . . As understanders we attempt to type people we hear about in terms of one of our standard life themes. As we hear of differences from the normal type we create a private life theme for the individual we are hearing about. The infinity of possible life themes comes from this possibility of the unique combination of goals for any individual. What makes life themes manageable is that the number of life theme types is small (six) and the number of standard life themes within those typings is a tractable size (say 10 to 50 for each type).[110]

If these primitives don't account for our understanding of the variety of possible human lives, Schank is ready, as always, to add a few more.

Nothing could ever call into question Schank's basic assumption that all human practice and know-how is represented in the mind as a system of beliefs constructed from context-free primitive actions and facts, but there are signs of trouble. Schank does admit that an individual's "belief system" cannot be fully elicited from him; although he never doubts that it exists and that it could in principle be represented in his formalism. He is therefore led to the desperate idea of a program which could learn about everything from restaurants to life themes the way people do. In a recent paper he concludes:

> We hope to be able to build a program that can learn, as a child does, how to do what we have described in this paper instead of being spoon-fed the tremendous information necessary. In order to do this it might be necessary to await an effective automatic hand-eye system and an image processor.[111]

For Schank's *ad hoc* approach there is no way of ever facing an interesting failure, but the fact that robot makers such as Raphael report that progress in their area must await an adequate scheme for knowledge representation, and that those like Schank who hope to provide such representation systems finally fall back on robots as a means for acquiring them, suggests that the field is in a loop—the computer world's conception of a crisis.

In any case, Schank's appeal to learning is at best another evasion. Developmental psychology has shown that children's learning does not consist merely in acquiring more and more information about specific routine situations by adding new primitives and combining old ones as Schank's view would lead one to expect. Rather learning of specific details takes place on a background of shared practices which seem to be picked up in everyday interactions not as facts and beliefs but as bodily skills for coping with the world. Any learning presupposes this background of implicit know-how which gives significance to details. Since Schank admits that he cannot see how this background can be made explicit so as to be given to a computer, and since the background is presupposed for the kind of script learning Schank has in mind, it seems that his project of using preanalyzed primitives to capture commonsense understanding is doomed.

A more plausible, even if in the last analysis perhaps no more promising, approach would be to use the new theoretical power of frames or stereotypes to dispense with the need to preanalyze everyday situations in terms of a set of primitive features whose *relevance is independent of context.* This approach starts with the recognition that in everyday communication " 'Meaning' is multi-dimensional, formalizable only in terms of the entire complex of goals and knowledge [of the world] being applied by both the producer and understander."[112] This knowledge, of course, is assumed to be "A body of specific beliefs (expressed as symbol structures . . .) making up the person's 'model of the world'."[113] Given these assumptions Terry Winograd and his co-workers are developing a new knowledge representation language (KRL), which they hope will enable programmers to capture these beliefs in symbolic descriptions of multidimensional prototypical objects whose *relevant aspects are a function of their context.*

Prototypes would be structured so that any sort of description from proper names to procedures for recognizing an example could be used to fill in any one of the nodes or slots that are attached to a prototype. This allows representations to be defined in terms of each other, and results in what the author calls "a *wholistic* as opposed to

reductionistic view of representation."[114] For example, since any description could be part of any other, chairs could be described as having aspects such as seats and backs, and seats and backs in turn could be described in terms of their function in chairs. Furthermore, each prototypical object or situation could be described from many different perspectives. Thus nothing need be defined in terms of its necessary and sufficient features in the way Winston and traditional philosophers have proposed, but rather, following Rosch's research on prototypes, objects would be classified as more or less resembling certain prototypical descriptions.

Winograd illustrates this idea using the traditional philosophers' favorite example:

> The word "bachelor" has been used in many discussions of semantics, since (save for obscure meanings involving aquatic mammals and medieval chivalry) it seems to have a formally tractable meaning which can be paraphrased "an adult human male who has never been married". . . . In the realistic use of the word, there are many problems which are not as simply stated and formalized. Consider the following exchange:
>
> Host: I'm having a big party next weekend. Do you know any nice bachelors I could invite?
> Friend: Yes, I know this fellow X. . . .
>
> The problem is to decide, given the facts below, for which values of X the response would be a reasonable answer in light of the normal meaning of the word "bachelor". A simple test is to ask for which ones the host might fairly complain "You lied. You said X was a bachelor.":
>
> A: Arthur has been living happily with Alice for the last five years. They have a two year old daughter and have never officially married.
>
> B: Bruce was going to be drafted, so he arranged with his friend Barbara to have a justice of the peace marry them so he would be exempt. They have never lived together. He dates a number of women, and plans to have the marriage annulled as soon as he finds someone he wants to marry.
>
> C: Charlie is 17 years old. He lives at home with his parents and is in high school.
>
> D: David is 17 years old. He left home at 13, started a small business, and is now a successful young entrepreneur leading a playboy's life style in his penthouse apartment.

> E: Eli and Edgar are homosexual lovers who have been living together for many years.
>
> F: Faisal is allowed by the law of his native Abu Dhabi to have three wives. He currently has two and is interested in meeting another potential fiancee.
>
> G: Father Gregory is the bishop of the Catholic cathedral at Groton upon Thames.
>
> [This] cast of characters could be extended indefinitely, and in each case there are problems in deciding whether the word "bachelor" could appropriately be applied. In normal use, a word does not convey a clearly definable combination of primitive propositions, but evokes an *exemplar* which possesses a number of properties. This exemplar is not a specific individual in the experience of the language user, but is more abstract, representing a conflation of typical properties. A prototypical bachelor can be described as:
>
> 1. a person
> 2. a male
> 3. an adult
> 4. not currently officially married
> 5. not in a marriage-like living situation
> 6. potentially marriageable
> 7. leading a bachelor-like life style
> 8. not having been married previously
> 9. having an intention, at least temporarily, not to marry
> 10. . . .
>
> Each of the men described above fits some but not all of these characterizations. Except for narrow legalistic contexts, there is no significant sense in which a subset of the characteristics can be singled out as the "central meaning" of the word. In fact, among native English speakers there is little agreement about whether someone who has been previously married can properly be called a "bachelor" and fairly good agreement that it should not apply to someone who is not potentially marriageable (e.g. has taken a vow of celibacy).
>
> Not only is this list [of properties] open-ended, but the individual terms are themselves not definable in terms of primitive notions. In reducing the meaning of 'bachelor' to a formula involving 'adult' or 'potentially marriageable', one is led into describing these in terms of exemplars as well. 'Adult' cannot be defined in terms of years of age for any but technical legal purposes and in fact even in this restricted sense, it is defined differently for different aspects of the law. Phrases such as 'marriage-like living situation' and 'bachelor-like life style' reflect directly in their syntactic form the intention to convey stereotyped exemplars rather than formal definitions.[115]

Obviously if KRL succeeds in enabling AI researchers to use such prototypes to write flexible programs, such a language will be a major breakthrough and will avoid the *ad hoc* character of the "solutions" typical of micro-world programs. Indeed, the future of AI depends on some such work as that begun with the development of KRL. But there are problems with this approach. Winograd's analysis has the important consequence that in comparing two prototypes, what counts as a match and thus what counts as the relevant aspects which justify the match will be a result of the program's understanding of the current context.

> The result of a matching process is not a simple true/false answer. It can be stated in its most general form as: "Given the set of alternatives which I am currently considering . . . and looking in order at those stored structures which are most accessible in the *current context,* here is the best match, here is the degree to which it seems to hold, and here are the specific detailed places where match was not found. . . ."
>
> The selection of the order in which sub-structures of the description will be compared is a function of their current accessibility, which depends both on the form in which they are stored and the *current context.*[116]

This raises four increasingly grave difficulties. First, for there to be "a class of cognitive 'matching' processes which operate on the descriptions (symbol structures) available for two entities, looking for correspondences and differences"[117] there must be a finite set of prototypes to be matched. To take Winograd's example:

> A single object or event can be described with respect to several prototypes, with further specifications from the perspective of each. The fact that last week *Rusty flew to San Francisco* would be expressed by describing the event as a typical instance of *Travel* with the mode specified as *Airplane,* destination *San Francisco,* etc. It might also be described as a *Visit* with the actor being *Rusty,* the friends a particular group of people, the interaction warm, etc.[118]

But *etc.* covers what might, without predigestion for a specific purpose, be a hopeless proliferation. The same flight might also be a test flight, a check of crew performance, a stopover, a mistake, a golden opportunity, not to mention a visit to brother, sister, thesis adviser, guru, etc., etc., etc. Before the program can function at all the total set of possible alternatives must be pre-selected by the programmer.

Second, the matching makes sense only *after* the current candidates for comparison have been found. In chess, for example, positions can be compared only after the chess master calls to mind past positions the current board positions might plausibly resemble. And, as we saw in the chess case, the discovery of the relevant candidates which makes the matching of aspects possible requires experience and intuitive association.

We saw also, in both the chess and the robot cases, that the discovery of this prior similarity seems to point to some entirely different sort of processing than symbolic description—perhaps the sort of processing provided by some brain equivalent of holograms in which similarity is basic. The only way a KRL-based program (which must use symbolic descriptions) could proceed would be to guess some frame on the basis of what was already "understood" by the program, and then see if that frame's features could be matched to some current description. If not, the program would have to backtrack and try another prototype until it found one into whose slots or default terminals the incoming data could be fitted. This seems an altogether implausible and inefficient model of how we perform, and only rarely occurs in our conscious life (see p. 248 of this book for a Husserlian discussion of this problem). Of course, cognitive scientists could answer the above objection by maintaining, in spite of the implausibility, that we try out the various prototypes very quickly and are simply not aware of the frantic shuffling of hypotheses going on in our unconscious. But, in fact, most would agree with Winograd that at present the frame selection problem is unsolved.

> The problem of choosing the frames to try is another very open area. There is a selection problem, since we cannot take all of our possible frames for different kinds of events and match them against what is going on.[119]

There is, moreover, a third and more basic question which may pose an in-principle problem for any formal holistic account in which the significance of any fact, indeed what counts as a fact, always depends on context. Winograd stresses the critical importance of context:

> The results of human reasoning are *context dependent,* the structure of memory includes not only the long-term storage organization (what do I know?) but also

a current context (what is in focus at the moment?). We believe that this is an important feature of human thought, not an inconvenient limitation.[120]

He further notes that "the problem is to find a formal way of talking about . . . current attention focus and goals. . . ."[121] Yet he gives no formal account of how a computer program written in KRL could determine the current context.

Winograd's work does contain suggestive claims such as his remark that "the procedural approach formalizes notions such as 'current context' . . . and 'attention focus' in terms of the processes by which cognitive state changes as a person comprehends or produces utterances."[122] There are also occasional parenthetical references to "current goals, focus of attention, set of words recently heard, etc."[123] But reference to recent words has proven useless as a way of determining what the current context is, and reference to current goals and focus of attention is vague and perhaps even question-begging. If a human being's current goal is, say, to find a chair to sit on, his current focus might be on recognizing whether he is in a living room or a warehouse. He will also have short-range goals like finding the walls, longer-range goals like finding the light switch, middle-range goals like wanting to write or rest; and what counts as satisfying these goals will in turn depend on his ultimate goals and interpretation of himself as, say, a writer, or merely as easily exhausted and deserving comfort. So Winograd's appeal to "current goals and focus" covers too much to be useful in determining what specific situation the program is in.

To be consistent, Winograd would have to treat each type of situation the computer could be in as an object with *its* prototypical description; then in recognizing a specific situation, the situation or context in which *that* situation was encountered would determine which foci, goals, etc. were relevant. But where would such a regress stop? Human beings, of course, don't have this problem. They are, as Heidegger puts it, already in a situation, which they constantly revise. If we look at it genetically, this is no mystery. We can see that human beings are gradually trained into their cultural situation on the basis of their embodied precultural situation, in a way no programmer using KRL is trying to capture. But

for this very reason a program in KRL is not always-already-in-a-situation. Even if it represents all human knowledge in its stereotypes, including all possible types of human situations, it represents them from the outside like a Martian or a god. It isn't situated in any one of them, and it may be impossible to program it to behave as if it were.

This leads to my fourth and final question: Is the know-how which enables human beings constantly to sense what specific situation they are in, the sort of know-how which can be represented as a kind of knowledge in *any* knowledge representation language no matter how ingenious and complex? It seems that our sense of our situation is determined by our changing moods, by our current concerns and projects, by our long-range self-interpretation and probably also by our sensory-motor skills for coping with objects and people—skills we develop by practice without ever having to represent to ourselves our body as an object, our culture as a set of beliefs, and our propensities as situation → action rules. All these uniquely human capacities provide a "richness" or a "thickness" to our way of being-in-the-world and thus seem to play an essential role in situatedness, which in turn underlies all intelligent behavior.

There is no reason to suppose that moods, mattering, and embodied skills can be captured in any formal web of belief, and except for Kenneth Colby, whose view is not accepted by the rest of the AI community, no current work assumes that they can. Rather, all AI workers and cognitive psychologists are committed, more or less lucidly, to the view that such noncognitive aspects of the mind can simply be ignored. This belief that a significant part of what counts as intelligent behavior can be captured in purely cognitive structures defines cognitive science and is a version of what, in Chapter 4, I call the psychological assumption. Winograd makes it explicit:

> AI is the general study of those aspects of cognition which are common to all physical symbol systems, including humans and computers.[124]*

But this definition merely delimits the field; it in no way shows there is anything to study, let alone guarantees the project's success.

Seen in this light, Winograd's grounds for optimism contradict his

own basic assumptions. On the one hand, he sees that a lot of what goes on in human minds cannot be programmed, so he only hopes to program a significant part:

> [C]ognitive science . . . does not rest on an assumption that the analysis of mind as a physical symbol system provides a *complete* understanding of human thought. . . . For the paradigm to be of value, it is only necessary that there be *some significant aspects* of thought and language which can be profitably understood through analogy with other symbol systems we know how to construct.[125]

On the other hand, he sees that human intelligence is "wholistic" and that meaning depends on "the entire complex of goals and knowledge." What our discussion suggests is that all aspects of human thought, including nonformal aspects like moods, sensory-motor skills, and long-range self-interpretations, are so interrelated that one cannot substitute an abstractible web of explicit beliefs for the whole cloth of our concrete everyday practices.

What lends plausibility to the cognitivist position is the conviction that such a web of beliefs must finally fold back on itself and be complete, since we can know only a finite number of facts and procedures describable in a finite number of sentences. But since facts are descriminated and language is used only in a context, the argument that the web of belief must in principle be completely formalizable does not show that such a belief system can account for intelligent behavior. This would be true only if the context could also be captured in the web of facts and procedures. But if the context is determined by moods, concerns, and skills, then the fact that our beliefs can in principle be completely represented does not show that representations are sufficient to account for cognition. Indeed, if nonrepresentable capacities play an essential role in situatedness, and the situation is presupposed by all intelligent behavior, then the "aspects of cognition which are common to all physical symbol systems" will not be able to account for any cognitive *performance* at all.

In the end the very idea of a holistic information processing model in which the relevance of the facts depends on the context may involve a contradiction. To recognize any context one must have already selected from the indefinite number of possibly descriminable features the possibly relevant ones, but such a selection can be made only after the context has already been recognized as similar to an already analyzed one. The holist thus faces a vicious circle: relevance presupposes similarity and similarity presupposes relevance. The only way to avoid this loop is to be always-already-in-a-situation without representing it so that the problem of the priority of context and features does not arise, or else to return to the reductionist project of preanalyzing all situations in terms of a fixed set of possibly relevant primitives—a project which has its own practical problems, as our analysis of Schank's work has shown, and, as we shall see in the conclusion, may have its own internal contradiction as well.

Whether this is, indeed, an in-principle obstacle to Winograd's approach only further research will tell. Winograd himself is admirably cautious in his claims:

> If the procedural approach is successful, it will eventually be possible to describe the mechanisms at such a level of detail that there will be a verifiable fit with many aspects of detailed human performance . . . but we are nowhere near having explanations which cover language processing as a whole, including meaning.[126]

If problems do arise because of the necessity in any formalism of isolating beliefs from the rest of human activity, Winograd will no doubt have the courage to analyze and profit from the discovery. In the meantime everyone interested in the philosophical project of cognitive science will be watching to see if Winograd and company can produce a moodless, disembodied, concernless, already adult surrogate for our slowly acquired situated understanding.

Conclusion

Given the fundamental supposition of the information processing approach that all that is relevant to intelligent behavior can be formalized in a structured description, all problems must appear to be merely problems of complexity. Bobrow and Winograd put this final faith very clearly at the end of their description of KRL:

> The system is complex, and will continue to get more so in the near future. . . . [W]e do not expect that it will ever be reduced to a very small set of mechanisms. Human thought, we believe, is the product of the interaction of a fairly large set of interdependent processes. Any representation language which is to be used in modeling thought or achieving "intelligent" performance will have to have an extensive and varied repertoire of mechanisms.[127]

Underlying this mechanistic assumption is an even deeper assumption which has gradually become clear during the past ten years of research. During this period AI researchers have consistently run up against the problem of representing everyday context, just as I predicted they would in the first edition of this book. Work during the first five years (1967–1972) demonstrated the futility of trying to evade the importance of everyday context by creating artificial gamelike contexts preanalyzed in terms of a list of fixed-relevance features. More recent work has thus been forced to deal directly with the background of commonsense know-how which guides our changing sense of what counts as the relevant facts. Faced with this necessity researchers have implicitly tried to treat the broadest context or background as an object with its own set of preselected descriptive features. This assumption, that the background can be treated as just another object to be represented in the same sort of structured description in which everyday objects are represented, is essential to our whole philosophical tradition. Following Heidegger, who is the first to have identified and criticized this assumption, I will call it the metaphysical assumption.

The obvious question to ask in conclusion is: Is there any evidence besides the persistent difficulties and history of unfulfilled promises in AI for believing that the metaphysical assumption is unjustified? It may be that no argument can be given against it, since facts put forth to show that the background of practices is unrepresentable are in that very act shown to be the sort of facts which *can* be represented. Still, since the value of this whole dialogue is to help each side to become as clear as possible concerning its presuppositions and their possible justification, I will attempt to lay out the argument which underlies my antiformalist, and, therefore, antimechanist convictions.

My thesis, which owes a lot to Wittgenstein,[128*] is that whenever human behavior is analyzed in terms of rules, these rules must always

contain a *ceteris paribus* condition, i.e., they apply "everything else being equal," and what "everything else" and "equal" means in any specific situation can never be fully spelled out without a regress. Moreover, this *ceteris paribus* condition is not merely an annoyance which shows that the analysis is not yet complete and might be what Husserl called an "infinite task." Rather the *ceteris paribus* condition points to a background of practices which are the condition of the possibility of all rulelike activity. In explaining our actions we must always sooner or later fall back on our everyday practices and simply say "this is what we do" or "that's what it is to be a human being." Thus in the last analysis all intelligibility and all intelligent behavior must be traced back to our sense of what we *are,* which is, according to this argument, necessarily, on pain of regress, something we can never explicitly *know.*

This argument can be best worked out in terms of an example. Back in 1972 when Minsky was working on the frame concept, one of his students, Eugene Charniak, was developing a scriptlike approach for dealing with children's stories. Papert and Goldstein provide a revealing analysis of this approach:

> . . . [C]onsider the following story fragment from Charniak,
>
> Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take *it* back."
>
> The goal is to construct a theory that explains how the reader understands that *"it"* refers to the new kite, not the one Jack already owns. Purely syntactic criteria (such as assigning the referent of *"it"* to the last mentioned noun) are clearly inadequate, as the result would be to mistakenly understand the last sentence of the story as meaning that Jack will make Janet take back the kite *he already owns.* . . . [I]t is clear that one cannot know that *"it"* refers to the new kite without knowledge about the trading habits of our society. One could imagine a different world in which newly bought objects are never returned to the store, but old ones are. The question we raise here is how this knowledge might be represented, stored and made available to the process of understanding Charniak's story.[129]

Their answer to this question is, of course, dictated by the metaphysical assumption. They try to make the background of practices involved explicit as a set of beliefs:

Charniak's formal realization of a *frame* was in the form of *base-knowledge* about a large variety of situations that arise in the context of these stories. The mechanism of his program was for the content of sentences to evoke this base knowledge with the following effect: demons ("frame-keepers" in our terminology) were created to monitor the possible occurrence in later sentences of likely (but not inevitable) consequences of the given situation. Thus, for our story fragment the birthday knowledge creates expectations about the need for participants of the party to buy presents and the possible consequence of having to return these gifts. Hence, these demons expect the possibility of Jack already possessing the present and the resulting need for Janet to return *it,* where *it* is known to be the present.[130]

But once games and micro-worlds are left behind, a yawning abyss threatens to swallow up those who try to carry out such a program. Papert and Goldstein march bravely in:

> . . . But the story does not include explicitly all important facts. Look back at the story. Some readers will be surprised to note that the text itself does not state (a) that the presents bought by Penny and Janet were for *Jack,* (b) that the [kite] bought by Janet was intended as a present, and (c) that having an object implies that one does not want another. All of the above facts are inserted into the database by other demons made activated by the birthday frame.[131]

Our example turns on the question: How does one store the "facts" mentioned in (c) above about returning presents? To begin with there are perhaps indefinitely many reasons for taking a present back. It may be the wrong size, run on the wrong voltage, be carcinogenic, make too much noise, be considered too childish, too feminine, too masculine, too American, etc., etc. And each of these facts requires further facts to be understood. But we will concentrate on the reason mentioned in (c): that normally, i.e., *everything else being equal,* if one has an object, one does not want another just like it. Of course, this cannot simply be entered as a true proposition. It does not hold for dollar bills, cookies, or marbles. (It is not clear it even holds for kites.) Papert and Goldstein would answer that, of course, once we talk of the norm we must be prepared to deal with exceptions:

> [T]he typical situation in comprehension is to be faced with a set of clues that evoke a rich and detailed knowledge structure, the frame, that supplies the unstated details. Naturally, these defaults may be inappropriate for some situations and, in those cases, the text must supply the exceptions.[132]

But here the desperate hand waving begins, for the text need not explicitly mention the exceptions at all. If the gift were marbles or cookies, the text surely would not mention that these were exceptions to the general rule that one of a kind is enough. So the data base would have to contain *an account of all possible exceptions* to augment the text—if it even makes sense to think of this as a definite list. Worse, even if one listed all the exceptional cases where one would be glad to possess more than one specimen of a certain type of object, there are situations which allow an exception to this exception: already having one cookie is more than enough if the cookie in question is three feet in diameter; one thousand marbles is more than a normal child can handle, etc. Must we then list the situations which lead one to expect exceptions to the exceptions? But these exceptions too can be overridden in the case of, say, a cookie monster or a marble freak, and so it goes. . . . The computer programmer writing a story understander must try to list all possibly relevant information, and once that information contains appeals to the *normal* or *typical* there is no way to avoid an infinite regress of qualifications for applying that knowledge to a specific situation.

The only "answer" the M.I.T. group offers is the metaphysical assumption that the background of everyday life is a set of rigidly defined situations in which the relevant facts are as clear as in a game:

> The fundamental frame assumption is the thesis that . . . [m]ost situations in which people find themselves *have sufficient in common* with previously encountered situations for the salient features to be *pre-analyzed* and stored in a *situation-specific* form.[133]

But this "solution" is untenable for two reasons:[134*]

1. Even if the current situation is, indeed, *similar* to a preanalyzed one, we still have the problem of deciding which situation it is similar to. We have already seen that even in games such as chess no two positions are likely to be identical so a deep understanding of what is going on is required to decide what counts as a similar position in any two games. This should be even more obvious in cases where the problem is to decide

which preanalyzed situation a given real-world situation most resembles, for example whether a situation where there are well-dressed babies and new toys being presented has more in common with a birthday party or a beauty contest.

2. Even if all our lives *were* lived in identical stereotypical situations, we have just seen that any real-world frame must be described in terms of the normal, and that appeal to the normal necessarily leads to a regress when we try to characterize the conditions which determine the applicability of the norm to a specific case. Only our *general* sense of what is typical can decide here, and *that* background understanding by definition cannot be "situation-specific."

This is the other horn of the dilemma facing the information-processing model. We have seen in discussing KRL that the holistic approach leads to a circle as to which comes first, similarity or relevant aspects, now it turns out that the reductionist alternative leads to a regress.

Still, to this dilemma the AI researchers might plausibly respond: "Whatever the background of shared interests, feelings, and practices necessary for understanding specific situations, that knowledge *must* somehow be represented in the human beings who have that understanding. And how else could such knowledge be represented but in some explicit data structure?" Indeed, the kind of computer programming accepted by all workers in AI would require such a data structure, and so would philosophers who hold that all knowledge must be explicitly represented in our minds, but there are two alternatives which would avoid the contradictions inherent in the information-processing model by avoiding the idea that everything we know must be in the form of some explicit symbolic representation.

One response, shared by existential phenomenologists such as Merleau-Ponty and ordinary language philosophers such as Wittgenstein, is to say that such "knowledge" of human interests and practices need not be represented at all. Just as it seems plausible that I can learn to swim by practicing until I develop the necessary patterns of responses, without representing my body and muscular movements in some data structure, so too what I "know" about the cultural practices which enables me to recognize and act in specific situations has been gradually acquired

through training in which no one ever did or could, again on pain of regress, make explicit what was being learned.

Another possible account would allow a place for representations, at least in special cases where I have to stop and reflect, but such a position would stress that these are usually nonformal representations, more like images, by means of which I explore what I *am*, not what I *know*. On this view I don't normally represent to myself that I have desires, or that standing up requires balance, or, to take an example from Schank's attempt to make explicit our interpersonal knowledge, that:

> [I]f two people are positively emotionally related, then a negative change in one person's state will cause the other person to develop the goal of causing a positive change in the other's state.[135]

Still, when it is helpful, I can picture myself in a specific situation and ask myself what would I do or how would I feel—if I were in Jack's place how would I react to being given a second kite—without having to make explicit all that a computer would have to be told to come to a similar conclusion. We thus appeal to *concrete* representations (images or memories) based on our own experience without having to make explicit the strict rules and their spelled out *ceteris paribus* conditions required by *abstract* symbolic representations.

Indeed, it is hard to see how the subtle variety of ways things can matter to us could be exhaustively spelled out. We can anticipate and understand Jack's reaction because we remember what it feels like to be amused, amazed, incredulous, disappointed, disgruntled, saddened, annoyed, disgusted, upset, angry, furious, outraged, etc., and we recognize the impulses to action associated with these various degrees and kinds of concerns. A computer model would have to be given a description of each shade of feeling as well as each feeling's normal occasion and likely result.

The idea that feelings, memories, and images *must* be the conscious tip of an unconscious framelike data structure runs up against both *prima facie* evidence and the problem of explicating the *ceteris paribus* conditions. Moreover, the formalist assumption is not supported by one shred of scientific evidence from neurophysiology or psychology, or from

the past successes of AI, whose repeated failures required appeal to the metaphysical assumption in the first place.

AI's current difficulties, moreover, become intelligible in the light of this alternative view. The proposed formal representation of the background of practices in symbolic descriptions, whether in terms of situation-free primitives or more sophisticated data structures whose building blocks can be descriptions of situations, would, indeed, look more and more complex and intractable if minds were not physical symbol systems. If belief structures are the result of abstraction from the concrete practical context rather than the true building blocks of our world, it is no wonder the formalist finds himself stuck with the view that they are endlessly explicatable. On my view "the organization of world knowledge provides the largest stumbling block"[136] to AI precisely because the programmer is forced to treat the world as an object, and our know-how as knowledge.

But this metaphysical assumption definitive of cognitive science is never questioned by its practitioners. John McCarthy notes that "it is quite difficult to formalize the facts of common knowledge,"[137] but he never doubts that common knowledge can be accounted for in terms of facts.

> The epistemological part of AI studies what kinds of *facts* about the world are available to an observer with given opportunities to observe, how these facts can be represented in the memory of a computer, and what *rules* permit legitimate conclusions to be drawn from these facts.[138]

When AI workers finally face and analyze their failures it might well be this metaphysical assumption that they will find they have to reject.

Looking back over the past ten years of AI research we might say that the basic point which has emerged is that *since intelligence must be situated it cannot be separated from the rest of human life.* The persistent denial of this seemingly obvious point cannot, however, be laid at the door of AI. It starts with Plato's separation of the intellect or rational soul from the body with its skills, emotions, and appetites. Aristotle continued this unlikely dichotomy when he separated the theoretical

from the practical, and defined man as a rational animal—as if one could separate man's rationality from his animal needs and desires. If one thinks of the importance of the sensory-motor skills in the development of our ability to recognize and cope with objects, or of the role of needs and desires in structuring all social situations, or finally of the whole cultural background of human self-interpretation involved in our simply knowing how to pick out and use chairs, the idea that we can simply ignore this know-how while formalizing our intellectual understanding as a complex system of facts and rules is highly implausible.

However incredible, this dubious dichotomy now pervades our thinking about everything including computers. In the *Star Trek* TV series, the episode entitled "The Return of the Archons" tells of a wise statesman named Landru who programmed a computer to run a society. Unfortunately, he could give the computer only his abstract intelligence, not his concrete wisdom, so it turned the society into a rational plannified hell. No one stops to wonder how, without Landru's embodied skills, feelings, and concerns, the computer could understand everyday situations and so run a society at all.

In *Computer Power and Human Reason,*[139] Joseph Weizenbaum, a well-known contributor to work in AI (see pp. 218 ff.) makes this same mistake. Indeed, the radical separation of intelligence and wisdom is the basic assumption which seems to support but actually undermines the thesis of his otherwise eloquent book. Weizenbaum warns that we demean ourselves if we come to think of human beings on the AI model as devices for solving technical problems. But to make the argument that we are not such devices he embraces the very dichotomy which gives plausibility to AI. Weizenbaum argues, for example, that since a computer cannot understand loneliness it cannot *fully* understand the sentence " 'Will you come to dinner with me this evening' . . . to mean a shy young man's desperate longing for love"[140*] (a point which workers in AI would readily admit), while at the same time Weizenbaum grants the dubious AI assumption that "it may be possible, following Schank's procedures, to construct a conceptual structure that corresponds to the meaning of the sentence."[141] Stressing these extremes of emphathetic wisdom and formalized meaning leads Weizenbaum to overlook the

essential point that all meaningful discourse must take place in a shared context of concerns.

Ironically, Weizenbaum was the first major contributor to AI to recognize the essential relation of meaning and pragmatic context. As he put it in 1968: "[I]n real conversation global context assigns meaning to what is being said. . . ."[142] But once he overlooks this essential connection there is no way he can resist the conclusions of his AI colleagues. Thus, in spite of his well-documented claim that each culture has what Justice Oliver W. Holmes called its "tacit assumptions" and "unwritten practices,"[143] and his commitment to the strong thesis argued for in this book that these practices "cannot be explicated in any form but life itself,"[144]* Weizenbaum, like Minsky, concludes: "I see no way to put a bound on the degree of intelligence such an organism [i.e., a computer] could, at least in principle attain."[145]

This surprising admission can be explained only if Weizenbaum holds the AI view that the unexplicatable assumptions and unwritten practices of a culture play no essential role in the intelligent behavior of its members. Indeed, at times Weizenbaum seems to embrace the most implausible implications of this implausible view, viz., that these tacit assumptions and practices play no role in everyday linguistic communication, for he concedes that:

> It is technically feasible to build a computer system that will interview patients applying for help at a psychiatric out-patient clinic and will produce their psychiatric profiles complete with charts, graphs, and natural-language commentary.[146]

Consistent with this view that intelligence and natural language communication—as distinct from intuition and wisdom—are in-principle completely formalizable, Weizenbaum further allows that:

> . . . the view of man as a species of the more general genus "information-processing system" does concentrate our attention on one aspect of man. . . .[147]*

He calls to aid in justifying this claim the latest "scientific" version of the Platonic dichotomy—the split brain. This is a natural association, since pop literature on the split brain seems to support the science-fiction illusion of the separation of intuition and pure intelligence. As Weizenbaum explains it:

> The LH [Left Hemisphere] thinks, so to speak, in an orderly, sequential, and, we might call it, *logical fashion.* The RH [Right Hemisphere], on the other hand, appears to think in terms of *holistic* images. *Language processing* appears to be almost exclusively centered in the LH. . . .[148]

Here again linguistic capacity is isolated and equated with context-free logicality, forgetting, what Weizenbaum was the first AI worker to see, that when language is used in communication (and the Left Hemisphere alone is perfectly able to use language to communicate), "a global [holistic] context assigns meaning to what is being said. . . ."[149]

After these damaging admissions Weizenbaum is left with only the moralistic position that "however intelligent machines may be made to be, there are some acts of thought that ought to be attempted only by humans."[150]* This stricture presumably follows from the notion that although the background of cultural practices plays no essential role in intelligent behavior, including everyday conversation, it does play a role in the wisdom required in making sound legal decisions and psychiatric evaluations—although even here Weizenbaum is wary of making any in-principle claim. And he has good reason for caution, since once everyday activity has been admitted to be a technical problem amenable to the powers of pure formal intelligence it is impossible to draw a line limiting what computers may ultimately be able to do. All Weizenbaum has left is the high-minded platitude that "since we do not now have any ways of making computers wise, we ought not now to give computers tasks which demand wisdom."[151]*

From the perspective we have been laying out here the real problem is that Weizenbaum accepts the metaphysical assumption that whatever is required for everyday intelligence can be objectified and represented in a belief system. Whether this assumption takes the form of the deep philosophical claim that goes back to Leibniz and is still made by Husserl that the perceptions and practices required for situated intelligence can all be represented in a symbolic description, or the shallow technological view, shared by Weizenbaum and the "artificial intelligentsia" he opposes, that everyday understanding and natural language communication does not essentially involve our embodied, socialized skills, this assumption distorts our perception of our humanity.

Great artists have always sensed the truth, stubbornly denied by both

philosophers and technologists, that the basis of human intelligence cannot be isolated and explicitly understood. In *Moby Dick* Melville writes of the tattooed savage, Queequeg, that he had "written out on his body a complete theory of the heavens and the earth, and a mystical treatise on the art of attaining truth; so that Queequeg in his own proper person was a riddle to unfold; a wondrous work in one volume; but whose mysteries not even himself could read. . . ."[152] Yeats puts it even more succinctly: "I have found what I wanted—to put it in a phrase, I say, 'Man can embody the truth, but he cannot know it'."

Hubert L. Dreyfus
1979

Introduction

Since the Greeks invented logic and geometry, the idea that all reasoning might be reduced to some kind of calculation—so that all arguments could be settled once and for all—has fascinated most of the Western tradition's rigorous thinkers. Socrates was the first to give voice to this vision. The story of artificial intelligence might well begin around 450 B.C. when (according to Plato) Socrates demands of Euthyphro, a fellow Athenian who, in the name of piety, is about to turn in his own father for murder: "I want to know what is characteristic of piety which makes all actions pious . . . that I may have it to turn to, and to use as a standard whereby to judge your actions and those of other men."[1]§ Socrates is asking Euthyphro for what modern computer theorists would call an "effective procedure," "a set of rules which tells us, from moment to moment, precisely how to behave."[2]

Plato generalized this demand for moral certainty into an epistemological demand. According to Plato, all knowledge must be stateable in explicit definitions which anyone could apply. If one could not state his know-how in terms of such explicit instructions—if his knowing *how*

§Notes begin on p. 307. [Citations are indicated by a superior figure. Substantive notes are indicated by a superior figure and an astersik.]