From Micro-Worlds to Knowledge Representation: AI at an Impasse

6

Hubert L. Dreyfus
1979

Since the completion, near the beginning of the decade, of the first edition of *What Computers Can't Do* (Dreyfus 1972/92), there has been considerable change and development in artificial intelligence. None of this more recent work leads me to relax my overall skepticism, but it does call for new and detailed discussion. In this article I review and examine a representative sample of AI as it has developed in the seventies. The work discussed falls into two categories, which can be assigned to the first and second halves of the decade, respectively.

1 The early seventies: micro-worlds

1.1 SHRDLU: understanding natural language

We begin with Winograd's (1972) natural-language understanding program, SHRDLU, inasmuch as it was the most impressive new achievement at the time. 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, and 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" (1973). 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 small caps are computer print-out; the parentheses contain Winograd's comments.)

(1) Pick up a big red block.

OKAY.

(The system answers "OKAY" 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.)

(3) 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.)

OKAY.

(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.

(5) 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 questions 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 datum 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.

(Winograd 1972, pp. 8–11)

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 (1967), 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 things we mean when we talk of understanding. The principles should derive from an attempt to deal with the basic cognitive structures. (1973, p. 167)

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. 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 MIT, 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. (p. 39)

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.

Given the admittedly artificial and arbitrary character of micro-worlds, why do Papert and Minsky 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", and the like. 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.* (1970, pp. 42–44)

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. (p. 48)

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, stealing; presents.
Playing:	Real and unreal; pretending.
Owning:	Part of; belongs 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; or spontaneous.
Results:	Not cooperative; lower threshold; aggression; loud voice; irrational; revenge.

And so on. (pp. 50–52)

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. (p. 52)

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. (1976, p. 9)

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. (1972, p. 26)

Thus, 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. (1977, p. 1062)

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. (p. 1063)

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 intentional 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. (p. 1064)

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 (compare McDermott 1976, p. 4), 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 (1974, p. 20), a program that deals with a "tiny bit of the world", with a program that deals with a "mini-world".

In our everyday life we are, indeed, involved in such various "sub-worlds" as the world of the theater, of business, or of mathematics, but each of these is a "mode" of our shared everyday world.[1] 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 years of stagnation in AI was mistaken for progress.

1.2 "Scene parsing" and computer vision

A second major application of the micro-world technique was in computer vision. Already in 1968, Adolfo Guzman's SEE program could analyze two-dimensional projections of complicated three-dimensional "scenes", consisting of piles of polyhedra. Even this early program correctly analyzed certain classes of scenes which people find difficult to figure out; but it had serious limitations. In 1972/75, Waltz generalized Guzman's methods, and produced a much more powerful vision system. Together, these programs provide a case study not only in how much can be achieved with the micro-worlds approach, but also in the kind of generalization that is possible within that approach—and, by implication, the kind that isn't.

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 common-sense 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" (1975, p. 8). 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

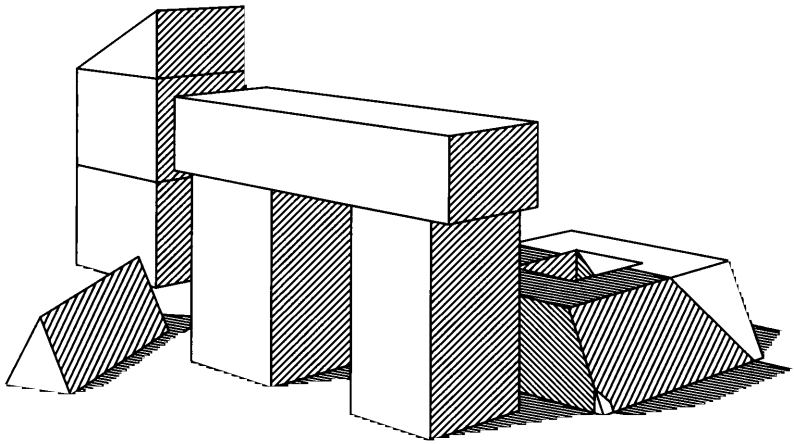

Figure 6.1: Sample blocks-world scene "parsed" by Waltz's program.

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 [6.1]. 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. (1976, pp. 77–78)

This is just the kind of mathematical generalization within a domain that one might expect in a micro-world, where the rule-governed relations 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. (1975, p. 8)

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. (1976, p. 39)

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 no way of thinking about general intelligent systems.

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" (1975, p. 2), 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. (1976, p. 3)

These comments suggest that in the early seventies an interesting change was taking place at MIT 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 MIT programs were "an attempt to build intelligent machines without any prejudice toward making the system ... humanoid" (Minsky 1969, p. 7). Now, in their book *Artificial Intelligence*, a summary of work done at MIT during the period 1967–72, Minsky and Papert (1973) present the MIT 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. (p. 11)

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 symbol 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. (1973, p. 34)

Some recent work in psychology, however, points exactly in the opposite direction. Rather than showing that perception can be analyzed in terms of formal features, Erich Goldmeier's (1972) extension 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]" (p. 1)—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*—that is, singularly salient shapes and orientations—suggests underlying physical phenomena such as "regions of resonance" (p. 128) in the brain.

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.

1.3 Learning new concepts or categories

Just such a formal model of everyday learning and categorization is proposed by Winston in his 1970 thesis, "Learning Structural Descriptions from Examples" (see Winston 1975). 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 Winston's program (along with those of Winograd, Guzman, and Waltz) is often mentioned as a major success of the micro-worlds technique, we must examine it in detail.

This program, too, illustrates the possibilities and essential limitations of micro-worlds. Is it the basis of a plausible general approach to learning? Winston thinks so.

> 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 on 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. (1975, p. 185)

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 spatial 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" (1973, p. 56). 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. 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" (p. 158) 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" (p. 194) by adding a functional predicate, "something to walk through" (p. 193). 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 these 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.
>
> (pp. 193–194)

Progress on recognition of arches, tables, and the like, 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" (p. 160). 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 (1973) admit that "we are still far from knowing how to design a powerful yet subtle and sensitive inductive learning program" (p. 56). What is surprising is that they add: "but the schemata developed in Winston's work should take us a substantial part of the way". 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?" (1975, p. 157), 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. (1977, p. 30)

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" (1973, p. 25). 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, 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 on intellectual processes in machines should lead to a science, eventually. (1973, p. 25)

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.

Before concluding our discussion of this research phase, it should be noted that some problem domains are (nearly enough) micro-worlds already; so they lend themselves to AI techniques without the need for artificial restrictions, and, by the same token, nongeneralizability is not the same kind of Waterloo. Game playing, particularly chess, is the most conspicuous example. Though some extravagant early predictions were not fulfilled, large computers now play fairly high caliber chess, and small machines that play creditable amateur games are being marketed as toys. But game players are not the only examples; excellent programs have been written for analyzing certain kinds of mass-spectroscopy data (Feigenbaum 1977), and for assisting in the diagnosis and treatment of some diseases (Shortliffe 1976). Such work is both impressive and important; but it shouldn't give the *wrong* impression. In each case, it succeeds because (and to the extent that) the relevant domain is well circumscribed in advance, with all the significant facts, questions, and/or options already laid out, and related by a comparatively small set of explicit rules—in short, because it's a micro-world. This is not to belittle either the difficulty or the value of spelling out such domains, or designing programs which perform well in them. But we should not see them as any closer to the achievement of genuine artificial intelligence than we do the "blocks-world" programs. In principle, interpreting mass spectrograms or batteries of specific symptoms has as little to do with the general intelligence of physicists and physicians, as disentangling vertices in projections of polyhedra does with vision. The real, theoretical problems for AI lie elsewhere.

2 The later seventies: knowledge representation

In roughly the latter half of the decade, the problem of how to structure and retrieve information, in situations where *anything* might be relevant, has come to the fore as the "knowledge-representation problem". 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 *tours de force* 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. (1977, pp. 1007–1008)

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. (1975/76, p. 7)

Another memorandum 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 program. 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.
>
> (Papert and Minsky 1973, p. 77)

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 larger, 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 (1972/92, pp. 153f/241f and 162/250). 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" (Husserl 1960, p. 53). As I explained 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. 153/241)

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, as Husserl puts it, "inviolably the same: as long as the objectivity remains intended as *this* one and of this kind" (p. 51)—plus "predelineations" of those properties which are possible but not necessary features of this type of object.

2.1 Frames and knowledge representation

Then, in 1974, 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. (1974 [chapter 5 of this volume], pp. 1f [111f])

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—that is, "explicating" the noemata 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" (1975, p. 16). 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 nature) as such, of psycho-physical being and humanity as such, cultures as such. (1960, pp. 54–55)

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 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 and effect, time, purpose, locality, process, and types of knowledge ... We need a serious epistemological research effort in this area. (p. 74 [138])

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, 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" (p. 34 [117]) 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). 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, barbers' chairs, sedan chairs, dentists' chairs, basket chairs, reclining chairs, wheel chairs, sling chairs, and beanbag chairs—as well as how to get off/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) ways to sit on/in them. Moreover, understanding chairs also includes social skills such as being able to sit appropriately (sedately, demurely, naturally, casually, sloppily, provocatively) 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 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 common sense 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 capitals 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* ... (p. 52 [132]; emphasis added)

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, and so on, since these are not context-free characteristics defined apart from chairs which then "cluster" in a chair representation; rather, legs, back, and the rest, 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.

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*. (emphasis added)

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 a art and play.

Minsky in his frame article claims that "the frame idea ... is in the tradition of ... the 'paradigms' of Kuhn" (p. 3 [113]); so it's appropri-

ate 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 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" (1962/70, p. 15). 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*.
>
> (p. 58 [135]; emphasis added)

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?"
>
> (Kuhn 1962/70, p. 200)

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. (p. 192)

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?
>
> (p. 11; emphasis added)

Although research based on frames cannot deal with this question, and so cannot account for common-sense 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 common-sense 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 group at Stanford and Xerox, the second.

2.2 Scripts and primitive actions

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, 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, and so forth. (1975a, p 39); 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 we compare 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 were 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 PTRANS-ing around the sun? Clearly the answer depends on the situation in which the question is asked.

Such primitives can be used, however, 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 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.
>
> (1975b, p. 131; emphasis added)

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 (1975b, p. 131)

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, one finds that the item in question is not available, or before paying one 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, and so on.

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 nonstandard occurrences to be explicitly mentioned" (Schank and Abelson 1977, p. 51); 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.[2] 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" (1975b, p. 132) 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 that 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.[3] 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.

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 than an individual's "belief system" cannot be fully elicited from him—though 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 one of his papers 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.
>
> (1972, pp. 553–554)

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 common sense understanding is doomed.

2.3 KRL: a knowledge-representation language

Winograd and Bobrow propose a more plausible, even if in the last analysis perhaps no more promising, approach that would 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 multidimensional, formalizable only in terms of the entire complex of goals and knowledge [of the world] being applied by both the producer and understander." (Winograd 1976b, p. 262) 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'" (p. 268). Given these assumptions, Winograd and his coworkers 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 authors call "a *holistic* as opposed to *reductionistic* view of representation" (Bobrow and Winograd 1977, p. 7). 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 by 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?

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 fiancée.

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 (for instance, 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 the syntactic form the intention to convey stereotyped exemplars rather than formal definitions. (1976b, 276–278)

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 count 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 substructures 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.* (1976b, p. 281–282; emphasis added)

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" (p. 280), 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*, and so on. It might also be described as a *Visit* with the actor being *Rusty*, the friends a particular group of people, the interaction warm, and so on. (Bobrow and Winograd 1977, p. 8)

But *"and so on"* 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, *and so on, and so on, and so on.* 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 that the current board positions might plausibly resemble. And (as in the chess case) the discovery of the relevant candidates which make the matching of aspects possible requires experience and intuitive association.

The only way a KRL-based program (which must use symbolic descriptions) could proceed, in chess or anywhere else, 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. 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 still agree with Winograd's (1974) assessment that the frame selection problem remains 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. (p. 80)

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 the context. Bobrow and Winograd stress 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. (1977, p. 32)

Winograd further notes that "the problem is to find a formal way of talking about ... current attention focus and goals" (1976b, p. 283). 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 like 'current context' ... and 'attention focus' in terms of the processes by which cognitive state changes as a person comprehends or produces utterances" (pp. 287–288). There are also occasional parenthetical references to "current goals, focus of attention, set of words recently heard, and so on" (p. 282). 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, and the like, were relevant. But where would such a regress stop? Human beings, of course, don't have this problem. They are, as Heidegger puts it, *always 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 that enables human beings constantly to sense what specific situation they are in the sort of know-how that 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, or 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 I call the *psychological assumption* (1972/92, chapter 4). 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. (see Schank et al. 1977, p. 1008)

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. (1976b, p. 264)

On the other hand, he sees that human intelligence is "holistic" 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 abstractable 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 discriminated, 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 discriminable 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.
>
> (1976b, p. 297)

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.

3 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.
>
> (Bobrow and Winograd 1977, p. 43)

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. 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 common-sense 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, I will attempt to lay out the argument which underlies my antiformalist, and therefore, antimechanist convictions.

My thesis, which owes a lot to Wittgenstein (1953), is that whenever human behavior is analyzed in terms of rules, these rules must always contain a *ceteris paribus* condition, that is, they apply "everything else being equal"; and what "everything else" and "equal" mean in any specific situation can never be fully spelled out without a regress. Moreover, the ceteris paribus condition is not merely an annoyance which shows that the analysis is not yet complete and might be what Husserl called and "infinite task". Rather the ceteris paribus condition points to a background of practices which are the condition of the possibility of all rule-like 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*.

Still, to this dilemma the AI researchers might plausibly respond: "*Whatever* background of shared interests, feelings, and practices is 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 enable 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 would

stress that these are usually nonformal representations—more like images, by means of which I explore what I *am*, not what I *know*. 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 as required by *abstract* symbolic representations.

The idea that feelings, memories, and images *must* be the conscious tip of an unconscious frame-like 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 explicable. On my view, the organization of world knowledge provides the largest stumbling block to AI precisely because the programmer is forced to treat the world as an object, and our know-how as knowledge.

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.

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 he himself could read" (1851/1952, p. 477). 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'."

Notes

1. This view is worked out further in Heidegger (1927/62); see especially p. 93 and all of section 18.
2. This is John Searle's way of formulating this important point. In a talk at the University of California at Berkeley (October 19, 1977), Schank agreed with Searle that to understand a visit to a restaurant, the computer needs more than a script; it needs to know everything that people know. He added that he is unhappy that as it stands his program cannot distinguish "degrees of weirdness". Indeed, for the program it is equally "weird" for the restaurant to be out of food as it is for the customer to respond by devouring the chef. Thus Schank seems to agree that without some understanding of degree of deviation from the norm, the program does not understand a story even when in that story events follow a completely normal stereotyped script. It follows that although scripts capture a necessary condition of everyday understanding, they do not provide a sufficient condition.
3. At the Society for Interdisciplinary Study of the Mind, Symposium for Philosophy and Computer Technology, State University College, New Paltz, NY, March 1977.