MAKING A MIND vs. MODELING THE BRAIN: AI BACK AT A BRANCHPOINT

[N]othing seems more possible to me than that people some day will come to the definite opinion that there is no copy in the ... nervous system which corresponds to a *particular* thought, or a *particular* idea, or memory. Ludwig Wittgenstein (1948)

[I]*nformation* is not stored anywhere in particular. Rather it is stored everywhere. Information is better thought of as "evoked" than "found". D. Rumelhart & D. Norman (1981)

In the early 1950s, as calculating machines were coming into their own, a few pioneer thinkers began to realize that digital computers could be more than number crunchers. At that point two opposed visions of what computers could be, each with its correlated research program, emerged and struggled for recognition. One faction saw computers as a system for manipulating mental symbols; the other, as a medium for modeling the brain. One sought to use computers to instantiate a formal representation of the world; the other, to simulate the interactions of neurons. One took problem solving as its paradigm of intelligence; the other, learning. One utilized logic, the other statistics. One school was the heir to the rationalist, reductionist tradition in philosophy; the other viewed itself as idealized, holistic neuro-science.

The rallying cry of the first group was that both minds and digital computers were physical symbol systems. By 1955 Allen Newell and Herbert Simon, working at the RAND Corporation, had concluded that strings of bits manipulated by a digital computer could stand for anything -- numbers, of course, but also features of the real world. Moreover, programs could be used as rules to represent relations between these symbols, so that the system could infer further facts about the represented objects and their relations. As Newell put it recently in his account of the history of issues in AI:

The digital-computer field defined computers as machines that manipulated numbers. The great thing was, adherents said, that everything could be encoded into numbers, even instructions. In contrast, the scientists in AI saw computers as machines that manipulated symbols. The great thing was, they said, that everything could be encoded into symbols, even numbers. Allen Newell

This way of looking at computers became the basis of a way of looking at minds. Newell and Simon hypothesized that the human brain and the digital computer, while totally different in structure and mechanism, had, at the appropriate level of abstraction, a common functional description. At this level, both the human brain and the appropriately programmed digital computer could be seen as two different instantiations of a single species of device -- one which generated intelligent behavior by manipulating symbols by means of formal rules. Newell and Simon stated their view as an hypothesis:

<u>The Physical Symbol System Hypothesis</u>. A physical symbol system has the necessary and sufficient means for general intelligent action. . By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence.

Newell and Simon trace the roots of their hypothesis back to Frege, Russell, and Whitehead but, of course, Frege and company were themselves heirs to a long atomistic, rationalist tradition. Descartes already assumed that all understanding consisted in forming and manipulating appropriate representations, that these representations could be analyzed into primitive elements (naturas simplices), and that all phenomena could be understood as a complex combinations of these simple elements. Moreover, at the same time, Hobbes implicitly assumed that the elements were formal elements related by purely syntactic operations, so that reasoning could be reduced to calculation. "When a man *reasons*, he does nothing else but conceive a sum total from addition of parcels," Hobbes wrote, "for REASON ... is nothing but reckoning. Finally Leibniz, working out the classical idea of mathesis -- the formalization of everything --, sought support to develop a universal symbol system, so that "we can assign to every object its determined characteristic number". According to Leibniz, in understanding we analyze concepts into more simple elements. In order to avoid a regress of simpler and simpler elements, there must be ultimate simples in terms of which all complex concepts can be understood. Moreover, if concepts are to apply to the world, there must be simple features which these elements represent. Leibniz envisaged "a kind of alphabet of human thoughts".

Ludwig Wittgenstein, drawing on Frege and Russell, stated the pure form of this syntactic, representational view of the relation of the mind to reality in his <u>Tractatus</u> <u>Logico-Philosophicus</u>. He defined the world as the totality of logically independent atomic facts:

1. The world is the totality of facts, not of things. Facts, in turn, were exhaustively analyzable into primitive objects.

2.01. An atomic fact is a combination of objects

2.0124. If all objects are given, then *thereby* all atomic facts are given. These facts, their constituents, and their logical relations were represented in the mind. 2.1. We make to ourselves pictures of facts.

2.15. That the elements of the picture are combined with one another in a definite way, represents that the things are so combined with one *another*

AI can be thought of as the attempt to find the primitive elements and logical relations in the subject (man or computer) which mirror the primitive objects and their relations which make up the world. Newell and Simon's physical symbol system hypothesis in effect turns the Wittgensteinian vision -- which is itself the culmination of the classical rationalist philosophical tradition -- into an empirical claim, and bases a research program on it.

The opposed intuition, that we should set about creating artificial intelligence by modeling the brain not the mind's symbolic representation of the world, drew its inspiration not from philosophy but from what was soon to be called neuro-science. It was directly inspired by the work of D.O. Hebb who in 1949 suggested that a mass of neurons could learn if, when neuron A and neuron B were simultaneously excited, that increased the strength of the connection between them.

This lead was followed by Frank Rosenblatt who reasoned that since intelligent behavior based on our representation of the world was likely to be hard to formalize, AI should rather attempt to automate the procedures by which a network of neurons learns to discriminate patterns and respond appropriately. As Rosenblatt put it:

The implicit assumption [of the symbol manipulating research program] is that it is relatively easy to specify the behavior that we want the system to perform, and that the challenge is then to design a device or mechanism which will effectively carry out this behavior ... [I]t is both easier and more profitable to axiomatize the *physical system* and then investigate this system analytically to determine its behavior, than to axiomatize the *behavior* and then design a physical system by techniques of logical synthesis.

Another way to put the difference between the two research programs is that those seeking symbolic representations were looking for a formal structure that would give the computer the ability to solve a certain class of problems or discriminate certain types of patterns. Rosenblatt, on the other hand, wanted to build a physical device, or to simulate such a device on a digital computer, which then could generate its own abilities.

Many of the models which we have heard discussed are concerned with the question of what logical structure a system must have if it is to exhibit some property, X. This is essentially a question about a static system. ... An alternative way of looking at the question is: what kind of a system can *evolve* property X? I think we can show in a number of interesting cases that the second question can be solved without having an answer to the first.

Both approaches met with immediate and startling success. Newell and Simon succeeded by 1956 in programming a computer using symbolic representations to solve

simple puzzles and prove theorems in the propositional calculus. On the basis of these early impressive results it looked like the physical symbol system hypothesis was about to be confirmed, and Newell and Simon were understandably euphoric. Simon announced:

It is not my aim to surprise or shock you ... But the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until -- in a visible future -- the range of problems they can handle will be coextensive with the range to which the human mind has been applied.

He and Newell explained:

[W]e now have the elements of a theory of heuristic (as contrasted with algorithmic) problem solving; and we can use this theory both to understand human heuristic processes and to simulate such processes with digital computers. Intuition, insight, and learning are no longer exclusive possessions of humans: any large high-speed computer can be programmed to exhibit them also.

Rosenblatt put his ideas to work in a type of device which he called a perceptron. By 1956 Rosenblatt was able to train a perceptron to classify certain types of patterns as similar and to separate these from other patterns which were dissimilar. By 1959 he too was jubilant and felt his approach had been vindicated:

It seems clear that the ... perceptron introduces a new kind of information processing automaton: For the first time, we have a machine which is capable of having original ideas. As an analogue of the biological brain, the perceptron, more precisely, the theory of statistical separability, seems to come closer to meeting the requirements of a functional explanation of the nervous system than any system previously proposed. ... As concept, it would seem that the perceptron has established, beyond doubt, the feasibility and principle of non-human systems which may embody human cognitive functions ... The future of information processing devices which operate on statistical, rather than logical, principles seems to be clearly indicated.

In the early sixties both approaches looked equally promising, and both made themselves equally vulnerable by making exaggerated claims. Yet the result of the internal war between the two research programs was surprisingly asymmetrical. By 1970 the brain simulation research which had its paradigm in the perceptron was reduced to a few, lonely, underfunded efforts, while those who proposed using digital computers as symbol manipulators had undisputed control of the resources, graduate programs, journals, symposia, etc. that constitute a flourishing research program. Reconstructing how this came about is complicated by the myth of manifest destiny an on-going research program generates. Thus it looks to the victors as if symbolic information processing won out because it was on the right track, while the neural net approach lost because it simply didn't work. But this account of the history of the field is a retroactive illusion. Both research programs had ideas worth exploring and both had deep, unrecognized problems.

Each position had its detractors and what they said was essentially the same: each approach had shown that it could solve certain easy problems but that there was no reason to think that either group could extrapolate its methods to real world complexity. Indeed, there was evidence that as problems got more complex the computation required by both approaches would grow exponentially and so soon become intractable. Marvin Minsky and Seymour Papert said in 1969 of Rosenblatt's perceptron:

Rosenblatt's schemes quickly took root, and soon there were perhaps as many as a hundred groups, large and small, experimenting with the model ... The results of these hundreds of projects and experiments were generally disappointing, and the explanations inconclusive. The machines usually work quite well on very simple problems but deteriorate very rapidly as the tasks assigned to them get harder.

Three years later, Sir James Lighthill, after reviewing work using heuristic programs such as Simon's and Minsky's reached a strikingly similar negative conclusion:

Most workers in AI research and in related fields confess to a pronounced feeling of disappointment in what has been achieved in the past 25 years. Workers entered the field around 1950, and even around 1960, with high hopes that are very far from having been realized in 1972. In no part of the field have the discoveries made so far produced the major impact that was then promised. ... [O]ne rather general cause for the disappointments that have been experienced: failure to recognize the implications of the `combinatorial explosion'. This is a general obstacle to the construction of a ... system on a large knowledge base which results from the explosive growth of any combinatorial expression, representing numbers of possible ways of grouping elements of the knowledge base according to particular rules, as the base's size increases.

As David Rumelhart succinctly sums it up: "Combinatorial explosion catches you sooner or later, although sometimes in different ways in parallel than in serial. Both sides had, as Jerry Fodor once put it, walked into a game of three-dimensional chess thinking it was tic-tac-toe. Why then, so early in the game, with so little known and so much to learn, did one team of researchers triumph at the total expense of the other? Why, at this crucial branchpoint, did the symbolic representation project become the only game in town?

Everyone who knows the history of the field will be able to point to the proximal cause. About 1965 Minsky and Papert, who were running a laboratory at MIT dedicated to the symbol manipulation approach and therefore competing for support with the perceptron projects, began circulating drafts of a book directly attacking perceptrons. In the book they made clear their scientific position:

Perceptrons have been widely publicized as "pattern recognition" or "learning" machines and as such have been discussed in a large number of books, journal articles, and voluminous "reports". Most of this writing ... is without scientific value.

But their attack was also a philosophical crusade. They rightly saw that traditional reliance on reduction to logical primitives was being challenged by a new holism.

Both of the present authors (first independently and later together) became involved with a somewhat therapeutic compulsion: to dispel what we feared to be the first shadows of a "holistic" or "Gestalt" misconception that would threaten to haunt the fields of engineering and artificial intelligence as it had earlier haunted biology and psychology.

They were quite right. Artificial neural nets may, but need not, allow an interpretation of their hidden nodes in terms of features a human being could recognize and use to solve the problem. While neural network modeling itself is committed to neither view, it can be demonstrated that association does not *require* that the hidden nodes be interpretable. Holists like Rosenblatt happily assumed that individual nodes or patterns of nodes were not picking out fixed features of the domain.

Minsky and Papert were so intent on eliminating all competition and so secure in the atomistic tradition that runs from Descartes to early Wittgenstein, that the book suggests much more than it actually demonstrates. They set out to analyze the capacity of a one-layer perceptron while completely ignoring in the mathematical portion of their book Rosenblatt's chapters on multilayer machines and his proof of the convergence of an (inefficient) probabilistic learning algorithm based on back propagation of errors. According to Rumelhart and McClelland:

Minsky and Papert set out to show which functions can and cannot be computed by [one-layer] machines. They demonstrated, in particular, that such perceptrons are unable to calculate such mathematical functions as parity (whether an odd or even number of points are on in the retina) or the topological function of connectedness (whether all points that are on are connected to all other points that are on either directly or via other points that are also on) without making use of absurdly large numbers of predicates. The analysis is extremely elegant and demonstrates the importance of a mathematical approach to analyzing computational systems.

But the implications of the analysis are quite limited. Rumelhart and McClelland continue:

Essentially ... although Minsky and Papert were exactly correct in their analysis of the *one-layer perceptron*, the theorems don't apply to systems which are even a little more complex. In particular, it doesn't apply to multilayer systems nor to systems that allow feedback loops

Yet, in the conclusion to <u>Perceptrons</u>, when Minsky and Papert ask themselves the question: "Have you considered perceptrons with many layers?", they give the impression, while rhetorically leaving the question open, of having settled it.

Well, we have considered Gamba machines, which could be described as "two layers of perceptron." We have not found (by thinking or by studying the literature) any other really interesting class of multilayered machine, at least none whose principles seem to have a significant relation to those of the perceptron. ... [W]e consider it to be an important research problem to elucidate (or reject) our intuitive judgment that the extension is sterile.

Their attack of gestalt thinking in A.I. succeeded beyond their wildest dreams. Only an unappreciated few, among them S. Grossberg, J.A. Anderson and T. Kohonen, took up the "important research problem". Indeed, almost everyone in AI assumed that neural nets had been laid to rest forever. Rumelhart and McClelland note:

Minsky and Papert's analysis of the limitations of the one-layer perceptron, coupled with some of the early successes of the symbolic processing approach in artificial intelligence, was enough to suggest to a large number of workers in the field that there was no future in perceptron-like computational devices for artificial intelligence and cognitive psychology.

But why was it enough? Both approaches had produced some promising work and some unfounded promises. It was too early to close accounts on either approach. Yet something in Minsky and Papert's book struck a responsive chord. It seemed AI workers shared the quasi-religious philosophical prejudice against holism which motivated the attack. One can see the power of the tradition, for example, in Newell and Simon's article on physical symbol systems. The article begins with the scientific hypothesis that the mind and the computer are intelligent by virtue of manipulating discrete symbols, but it ends with a revelation. "The study of logic and computers has revealed to us that intelligence resides in physical-symbol systems. Holism could not compete with such intense philosophical convictions. Rosenblatt was discredited along with the hundreds of less responsible network research groups that his work had encouraged. His research money dried up, he had troubled getting his work published, he became depressed, and one day his boat was found empty at sea. Rumor had it that he had committed suicide. Whatever the truth of that rumor, one thing is certain: by 1970, as far as AI was concerned, neural nets were dead. Newell, in his history of AI, says the issue of symbols versus numbers "is certainly not alive now and has not been for a long time. Rosenblatt is not even mentioned in John Haugeland's or in Margaret Boden's histories of the AI field.

But blaming the rout of the connectionists on an anti-holistic prejudice is too simple. There was a deeper way philosophical assumptions influenced intuition and led to an overestimation of the importance of the early symbol processing results. The way it looked at the time was that the perceptron people had to do an immense amount of mathematical analysis and calculating to solve even the most simple problems of pattern recognition such as discriminating horizontal from vertical lines in various parts of the receptive field, while the symbol manipulating approach had relatively effortlessly solved hard problems in cognition such as proving theorems in logic and solving puzzles such as the cannibal-missionary problem. Even more importantly, it seemed that given the computing power available at the time, the neural net researchers could only do speculative neuro-science and psychology, while the simple programs of symbolic representationists were on their way to being useful. Behind this way of sizing up the situation was the assumption that thinking and pattern recognition are two distinct domains and that thinking is the more important of the two. As we shall see later in our discussion of the common sense knowledge problem, this way of looking at things ignores both the preeminent role of pattern discrimination in human expertise and also the background of common sense understanding which is presupposed in real world, everyday thinking. Taking account of this background may well require pattern recognition.

This gets us back to the philosophical tradition. It was not just Descartes and his descendants which stood behind symbolic information processing, but all of Western philosophy. According to Heidegger, traditional philosophy is defined from the start by its focusing on facts in the world while "passing over" the world as such. This means that philosophy has from the start systematically ignored or distorted the everyday context of human activity. That branch of the philosophical tradition that descends from Socrates, to Plato, to Descartes, to Leibniz, to Kant, to conventional AI takes it for granted, in addition, that understanding a domain consists in having a *theory* of that domain. A theory formulates the relationships between objective, *context-free* elements (simples, primitives, features, attributes, factors, data points, cues, etc.) in terms of abstract principles (covering laws, rules, programs, etc.).

Plato held that in theoretical domains such as mathematics and perhaps ethics, thinkers apply explicit, context-free rules or theories they learned in another life, outside the everyday world. Once learned, such theories function in this world by controlling the thinker's mind whether he is conscious of them or not. Plato's account did not apply to everyday skills but only to domains in which there is *a priori* knowledge. The success of theory in the natural sciences, however, reinforced the idea that in any orderly domain there must be some set of context-free elements and some abstract relations between those elements which accounts for the order of that domain and for man's ability to act intelligently in it. Thus Leibniz boldly generalized the rationalist account to all forms of intelligent activity, even everyday practice.

[T]he most important observations and turns of skill in all sorts of trades and professions are as yet unwritten. This fact is proved by experience when passing from theory to practice we desire to accomplish something. *Of course, we can also write up this practice, since it is at bottom just another theory more complex and particular.* . .

The symbolic information processing approach gains its assurance from this transfer to all domains of methods that were developed by philosophers and which have succeeded in the natural sciences. Since, on this view, any domain must be formalizable, the way to do AI in any area is obviously to find the context-free elements and principles and base a formal, symbolic representation on this theoretical analysis. Terry Winograd characteristically describes his AI 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.

No doubt theories about the universe are often built up gradually by modeling relatively simple and isolated systems and then making the model gradually more complex and integrating it with models of other domains. This is possible because all the phenomena are presumably the result of the law-like relations between what Papert and Minsky call "structural primitives." Since no one argues for atomistic reductionism in A.I. it seems that A.I. workers must implicitly assume that the abstraction of elements from their everyday context, which defines philosophy and works in natural science, must also work in AI. This would account for the way the physical symbol system hypothesis so quickly turned into a revelation and for the ease with which Papert's and Minsky's book triumphed over the holism of the perceptron.

Teaching philosophy at M.I.T. in the mid-sixties, Hubert was soon drawn into the debate over the possibility of AI. It was obvious to him that researchers such as Newell, Simon, and Minsky were the heirs to the philosophical tradition. But given his understanding of later Wittgenstein and early Heidegger, that did not seem to be a good omen for the reductionist research program. Both these thinkers had called into question the very tradition on which symbolic information processing was based. Both were holists, both were struck by the importance of everyday practices, and both held that one could not have a theory of the everyday world.

It is one of the ironies of intellectual history that Wittgenstein's devastating attack on his own <u>Tractatus</u>, his <u>Philosophical Investigations</u>, was published in 1953 just as AI took over the abstract, atomistic tradition he was attacking. After writing the <u>Tractatus</u> Wittgenstein spent years doing what he called "phenomenology" -- looking in vain for the atomic facts and basic objects his theory required. He ended by abandoning his <u>Tractatus</u> and all rationalistic philosophy. He argued that the analysis of everyday situations into facts and rules (which is where most traditional philosophers and AI researchers think theory must begin) is itself only meaningful in some context and for some purpose. Thus the elements chosen already reflect the goals and purposes for which they are carved out. When we try to find the ultimate context-free, purpose-free elements, as we must if we are going to find the primitive symbols to feed a computer, we are in effect trying to free aspects of our experience of just that pragmatic organization which makes it possible to use them intelligibly in coping with everyday problems.

In the <u>Philosophical Investigations</u> Wittgenstein directly criticizes the logical atomism of the <u>Tractatus</u>.

"What lies behind the idea that names really signify simples"? -- Socrates says in the <u>Theaetetus</u>: "If I make no mistake, I have heard some people say this: there is no definition of the primary elements -- so to speak -- out of which we and everything else are composed. ... But just as what consists of these primary elements is itself complex, so the names of the elements become descriptive language by being compounded together." Both Russell's `individuals' and my `objects' (<u>Tractatus Logico-Philosophicus</u>) were such primary elements. But what are the simple constituent parts of which reality is composed? ... It makes no sense at all to speak absolutely of the `simple parts of a chair.'

Already in the 1920s Martin Heidegger had reacted in a similar way against his mentor, Edmund Husserl, who regarded himself as the culmination of the Cartesian tradition and was, therefore, the grandfather of AI. Husserl argued that an act of consciousness or *noesis* does not, on its own, grasp an object; rather, the act has intentionality (directedness) only by virtue of an "abstract form" or meaning in the *noema* correlated with the act.

This meaning or symbolic representation, as conceived by Husserl, was a complex entity that had a difficult job to perform. In <u>Ideas</u> Husserl bravely tries to explain how the *noema* gets the job done. Reference is provided by predicate-*senses* which, like Fregean *Sinne*, just have the remarkable property of picking out objects' atomic properties. These predicates are combined into complex "descriptions" of complex objects, as in Russell's theory of descriptions. For Husserl, who is close to Kant on this point, the *noema* contains a hierarchy of strict rules. Since Husserl thought of intelligence as a context-determined, goal-directed activity, the mental representation of any type of object had to provide a context or "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. The *noema* must contain a rule describing all the features which can be expected with certainty in exploring a certain *type* of object--features which remain "inviolably the same: as long as the objectivity remains intended as *this* one and of this kind The rule must also prescribe "predelineations" of properties that are possible but not necessary features of this type of object: "Instead of a completely determined sense, there is always, therefore, a *frame of empty sense*.

In 1973 Marvin 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*

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 become "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 interactions between a knower and the world. The task of AI thus converges with the task of transcendental phenomenology. Both must try in everyday domains to find frames constructed from a set of primitive predicates and their formal relations.

Heidegger, before Wittgenstein, carried out, in response to Husserl, a phenomenological description of the everyday world and everyday objects like chairs and hammers, and like Wittgenstein he found that the everyday world could not be represented by a set of context-free elements. It was Heidegger who forced Husserl to face precisely this problem. He pointed out that there are other ways of "encountering" things than relating to them as objects defined by a set of predicates. When we use a piece of equipment like a hammer, Heidegger pointed out, we actualize a skill (which need not be represented in the mind) in the context of a socially organized nexus of equipment, purposes, and human roles (which need not be represented as a set of facts). This context or world, and our everyday ways of skillful coping in it which Heidegger called *circumspection*, are not something we *think* but, as part of our socialization, forms the way we *are*. Heidegger concluded:

The context ... can be taken formally in the sense of a system of relations. But ... [t]he phenomenal content of these `relations' and `relata' . . . is such that they resist any sort of mathematical functionalization; nor are they merely something thought, first posited in an `act of thinking.' They are rather relationships in which concernful circumspection as such already dwells.

This defines the splitting of the ways between Husserl and AI on the one hand, and Heidegger and later Wittgenstein on the other. The crucial question becomes: Can there be a theory of the everyday world as rationalist philosophers have always held? Or is the common sense background rather a combination of skills, practices, discriminations, etc., which are not intentional states, and so, *a fortiori*, do not have any representational content to be explicated in terms of elements and rules?

Husserl tried to avoid the problem posed by Heidegger by making a move soon to become familiar in AI circles. He claimed that the world, the background of significance, the everyday context, was merely a very complex system of facts correlated with a complex system of beliefs, which, since they have truth conditions, he called "validities". Thus one could, in principle, suspend one's dwelling in the world and achieve a detached, description of the human belief system. One could thus complete the task that had been implicit in philosophy since Socrates. One could make explicit the beliefs and principles underlying all intelligent behavior. As Husserl put it:

[E]ven the background ... of which we are always concurrently conscious but which is momentarily irrelevant and remains completely unnoticed, still functions according to its implicit validities

Since he firmly believed that the shared background could be made explicit as a belief system Husserl was ahead of his time in raising the question of the possibility of AI. After discussing the possibility of a formal axiomatic system describing experience, and pointing out that such a system of axioms and primitives -- at least as we know it in geometry -- could not describe everyday shapes such as "scalloped" and "lens-shaped," Husserl leaves open the question whether these everyday concepts could nonetheless be formalized. (This is like raising and leaving open the A.I. question whether one can axiomatize common sense physics.) Picking up Leibniz's dream of a mathesis of all experience, Husserl remarks:

The pressing question is .. whether there could not be ... an idealizing procedure that substitutes pure and strict ideals for intuited data and that would ... serve ... as the basic medium for a mathesis of experience.

But, as Heidegger predicted, the task of writing out a complete theoretical account of everyday life turned out to be much harder than initially expected. Husserl's project ran into serious trouble, and there are signs that Minsky's has too. During twenty-five years of trying to spell out the components of the subject's representation of everyday objects, Husserl found that he had to include more and more of a subject's commonsense understanding of the everyday 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 ... spatial objects (to say nothing of a Nature) as such, of psychophysical being and humanity as such, culture as such.

He spoke of the noema's "huge concreteness and of its "tremendous complication, and he sadly concluded at the age of seventy-five that he was a perpetual beginner and that phenomenology was an "infinite task.

here are hints in his frame paper that Minsky has embarked on the same "infinite task" that eventually overwhelmed Husserl:

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

To a student of contemporary philosophy Minsky's naivete and faith were astonishing. Husserl's phenomenology *was* just such a research effort. Indeed, philosophers, from Socrates to Leibniz, to early Wittgenstein, had carried on serious epistemological research in this area for two thousand years without notable success.

In the light of Wittgenstein's reversal and Heidegger's devastating critique of Husserl, Hubert predicted trouble for symbolic information processing. As Newell notes in his history of AI, Hubert's warning was ignored:

Dreyfus's central intellectual objection ... is that the analysis of the context of human action into discrete elements is doomed to failure. This objection is grounded in phenomenological philosophy. Unfortunately, this appears to be a nonissue as far as AI is concerned. The answers, refutations, and analyses that

have been forthcoming to Dreyfus's writings have simply not engaged this issue -which indeed would be a novel issue if it were to come to the fore.

The trouble was not long in coming to the fore, however, as the everyday world took its revenge on AI as it had on traditional philosophy. As we see it, the research program launched by Newell and Simon has gone through three ten-year stages.

From 1955-1965 two research themes, representation and search, dominated the field then called Cognitive Simulation. Newell and Simon showed, for example, how a computer could solve the cannibal and missionary problem, using the general heuristic search principle known as means-end analysis, viz. use any available operation that reduces the distance between the description of the current situation and the description of the goal. They then abstracted this heuristic technique and incorporated it into their General Problem Solver (GPS).

The second stage (1965-1975), led by Marvin Minsky and Seymour Papert at M.I.T., was concerned with what facts and rules to represent. The idea was to develop methods for dealing systematically with knowledge in isolated domains called micro-worlds. Famous programs written around 1970 at M.I.T. include Terry Winograd's SHRDLU which could obey commands given in a subset of natural language about a simplified blocks-world, Thomas Evan's Analogy Problem Program, David Waltz's Scene Analysis Program and Patrick Winston's program which learned concepts from examples.

The hope was that the restricted and isolated "micro-worlds" could be gradually made more realistic and combined so as to approach real world understanding. But researchers confused two domains which, following Heidegger, we shall distinguish as universe and world. A set of interrelated facts may constitute a *universe*, like the physical universe, but it does not constitute a world. The latter, like the world of business, the world of theater, or the world of the physicist, is an organized body of objects, purposes, skills, and practices on the basis of which human activities have meaning or make sense. To see the difference one can contrast the *meaningless* physical *universe* with the *meaningful world* of the discipline of physics. The world of physics, the business world, and the theater world, make sense only against a background of common human concerns. They are local elaborations of the one common-sense world we all share. That is, sub-worlds are not related like isolable physical systems to larger systems they *compose*, but are rather, local elaborations of a whole, which they *presuppose*. Micro-worlds were *not* worlds but isolated meaningless domains, and it has gradually become clear that there was no way they could be combined and extended to arrive at the world of everyday life.

In its third and so far final stage, roughly from 1975 to the present, AI has been wrestling with what has come to be called the common-sense knowledge problem. The representation of knowledge was always a central problem for work in AI, but the two earlier periods -- cognitive simulation and micro-worlds -- were characterized by an attempt to avoid the problem of common-sense knowledge by seeing how much could

be done with as little knowledge as possible. By the middle 1970s, however, the issue had to be faced. Various data structure such as Minsky's frames and Roger Schank's scripts have been tried without success. The common-sense knowledge problem has kept AI from even beginning to fulfill Simon's prediction made twenty years ago, that "within twenty years machines will be capable of doing any work a man can do".

Indeed, the common-sense knowledge problem has blocked all progress in theoretical AI for the past decade. Winograd was one of the first to see the limitations of SHRDLU and all script and frame attempts to extend the micro-worlds approach. Having "lost faith" in AI, he now teaches Heidegger in his computer science courses at Stanford, and points out "the difficulty of formalizing the common-sense background that determines which scripts, goals and strategies are relevant and how they interact.

What sustains AI in this impasse is the conviction that the common sense knowledge problem must be solvable since human beings have obviously solved it. But human beings may not normally use common sense *knowledge* at all. As Heidegger and Wittgenstein point out, what common sense *understanding* amounts to might well be *everyday know-how*. By know-how we do not mean procedural rules, but knowing what to do in a vast number of special cases. For example, common sense physics has turned out to be extremely hard to spell out in a set of facts and rules. When one tries, one either requires more common sense to understand the facts and rules one finds or else one produces formulas of such complexity that it seems highly unlikely they are in a child's mind.

Doing theoretical physics also requires background skills which may not be formalizable, but the domain itself can be described by abstract laws that make no reference to these background skills. AI researchers conclude that common sense physics too must be expressible as a set of abstract principles. But it just may be that the problem of finding a *theory* of common sense physics is insoluble because the domain has no theoretical structure. By playing all day with all sorts of liquids and solids for several years the child may simply have learned to discriminate prototypical cases of solids, liquids, etc. and learned typical skilled responses to their typical behavior in typical circumstances. The same might well be the case for the social world. If background understanding is indeed a skill, and skills are based on whole patterns and not on rules, we would expect symbolic representations to fail to capture our commonsense understanding.

In the light of this impasse, classical, symbol-based AI appears more and more to be a perfect example of what Imre Lakatos has called a degenerating research program. As we have seen, AI began auspiciously with Newell and Simon's work at RAND, and by the late 1960s had turned into a flourishing research program. Minsky predicted that "within a generation the problem of creating `artificial intelligence' will be substantially solved. Then, rather suddenly, the field ran into unexpected difficulties. It turned out to be much harder than one expected to formulate a theory of common-sense. It was not, as Minsky had hoped, just a question of cataloguing a few hundred thousand facts. The common-sense knowledge problem became the center of concern. Minsky's mood changed completely in five years. He told a reporter: "the AI problem is one of the hardest science has ever undertaken.

The Rationalist tradition had finally been put to an empirical test and it had failed. The idea of producing a formal, atomistic theory of the everyday common-sense world and representing that theory in a symbol manipulator had run into just the difficulties Heidegger and Wittgenstein discovered. Frank Rosenblatt's intuition that it would be hopelessly hard to formalize the world and thus give a formal specification of intelligent behavior had been vindicated. His repressed research program -- using the computer to instantiate a holistic model of an idealized brain -- which had never really been refuted, became again a live option.

In journalistic accounts of the history of AI Rosenblatt is vilified by anonymous detractors as a snake-oil salesman:

Present-day researchers remember that Rosenblatt was given to steady and extravagant statements about the performance of his machine. "He was a press agent's dream," one scientist says, "a real medicine man. To hear him tell it, the Perceptron was capable of fantastic things. And maybe it was. But you couldn't prove it by the work Frank did".

In fact he was much clearer about the capacities and limitations of the various types of perceptrons than Simon and Minsky were about their symbolic programs. Now he is being rehabilitated. Rumelhart, Hinton and McClelland reflect this new appreciation of his pioneering work:

Rosenblatt's work was very controversial at the time, and the specific models he proposed were not up to all the hopes he had for them. But his vision of the human information processing system as a dynamic, interactive, self-organizing system lies at the core of the PDP approach.

The studies of perceptrons ... clearly anticipated many of the results in use today. The critique of perceptrons by Minsky and Papert was widely misinterpreted as destroying their credibility, whereas the work simply showed limitations on the power of the most limited class of perceptron-like mechanisms, and said nothing about more powerful, multiple layer models.

Frustrated AI researchers, tired of clinging to a research program which Jerry Lettvin characterized in the early 1980s as "the only straw afloat", flocked to the new paradigm. Rumelhart and McClelland's book, *Parallel Distributed Processing*, sold 6000 copies the day it went on the market. 30,000 are now in print. As Paul Smolensky put it:

In the past half-decade the connectionist approach to cognitive modeling has grown from an obscure cult claiming a few true believers to a movement so vigorous that recent meetings of the Cognitive Science Society have begun to look like connectionist pep rallies.

If multilayered networks succeed in fulfilling their promise researchers will have to give up Descartes', Husserl's and early Wittgenstein's conviction that the only way to produce intelligent behavior is to mirror the world with a formal theory in the mind. Worse, one may have to give up the more basic intuition at the source of philosophy that there must be a theory of every aspect of reality, i.e., there must be elements and principles in terms of which one can account for the intelligibility of any domain. Neural networks may show that Heidegger, later Wittgenstein and Rosenblatt were right in thinking that we behave intelligently in the world without having a theory of that world. If a theory is not *necessary* to explain intelligent behavior we have to be prepared to raise the question whether, in everyday domains, such a theoretical explanation is even *possible*.

Neural net modelers, influenced by symbol manipulating AI, are expending considerable effort, once their nets have been trained to perform a task, trying to find the features represented by individual nodes and sets of nodes. Results thus far are equivocal. Consider Geoffrey Hinton's network for learning concepts by means of distributed representations. Hinton's network can be trained to encode relationships in a domain which human beings conceptualize in terms of features, without the network being given the features that human beings use. Hinton produces examples of cases in which in the trained network some nodes can be interpreted as corresponding to the features that human beings pick out, although they only roughly correspond to these features. Most nodes, however, cannot be interpreted semantically at all. A feature used in a symbolic representation is either present or not. In the net, however, although certain nodes are more active when a certain feature is present in the domain, the amount of activity varies not just with the presence or absence of this feature, but is affected by the presence or absence of other features as well.

Hinton has picked a domain, family relationships, which is constructed by human beings precisely in terms of the features, such as generation and nationality, which human beings normally notice. Hinton then analyzes those cases in which, starting with certain random initial connection strengths, some nodes after learning can be interpreted as representing these features. Calculations using Hinton's model show, however, that even his net seems, for some random initial connection strengths, to learn its associations without any obvious use of these everyday features.

In one very limited sense, any successfully trained multilayer net has an interpretation in terms of features -- not everyday features but what we shall call highly abstract features. Consider the particularly simple case of layers of binary units activated by feedforward, but not lateral or feedback, connections. To construct an account from a network that has learned certain associations, each node one level above the input nodes could, on the basis of connections to it, be interpreted as detecting when one of a certain set of input patterns is present. (Some of the patterns will be the ones

used in training and some will never have been used.) If the set of input patterns which a particular node detects is given an invented name (it almost certainly won't have a name in our vocabulary), the node could be interpreted as detecting the highly abstract feature so named. Hence, every node one level above the input level can be characterized as a feature detector. Similarly, every node a level above these nodes can be interpreted as detecting a higher-order feature which is defined as the presence of one of a specified set of patterns among the first level features detectors. And so on up the hierarchy.

The fact that intelligence, defined as the knowledge of a certain set of associations appropriate to a domain, can always be accounted for in terms of relations among a number of highly abstract features of a skill domain does not, however, preserve the rationalist intuition that these explanatory features must capture the essential structure of the domain, i.e., that one could base a theory on them. If the net is taught one more association of an input/output pair (where the input prior to training produces an output different from the one to be learned), the interpretation of at least some of the nodes will have to be changed. So the features which some of the nodes picked out before the last instance of training would turn out not to have been invariant structural features of the domain.

Once one has abandoned the philosophical approach of classical AI and accepted the atheoretical claim of neural netmodeling, one question remains: How much of everyday intelligence can such a network be expected to capture? Classical AI researchers are quick to point out -- as Rosenblatt already noted -- that neural net modelers have so far had difficulty dealing with step-wise problem solving. Connectionists respond that they are confident that they will solve that problem in time. This response, however, reminds one too much of the way that the symbol manipulators in the sixties responded to the criticism that their programs were poor at the perception of patterns. The old struggle between intellectualists who, because they can do contextfree logic think they have a handle on everyday cognition but are poor at understanding perception, and gestaltists who have the rudiments of an account of perception, but none of everyday cognition, goes on. One might think, using the metaphor of the right and left brain, that perhaps the brain/mind uses each strategy when appropriate. The problem would then be how to combine them. One cannot just switch back and forth for, as Heidegger and the gestaltists saw, the pragmatic background plays a crucial role in determining relevance even in everyday logic and problem solving, and experts in any field, even logic, grasp operations in terms of their functional similarities.

It is even premature to consider combining the two approaches, since so far neither has accomplished enough to be on solid ground. Neural network modeling may simply be getting a deserved chance to fail as did the symbolic approach.

Still there is an important difference to remember as each research program struggles on. The physical symbol system approach seems to be failing because it is simply false to assume that there must be a theory of every domain. Neural network modeling, however, is not committed to this or any other philosophical assumption. However, simply building an interactive net sufficiently similar to the one our brain has evolved may be just too hard. Indeed, the common sense knowledge problem, which has blocked the progress of symbolic representation techniques for fifteen years, may be looming on the neural net horizon, although connectionists may not yet recognize it. All neural net modelers agree that for a net to be intelligent it must be able to generalize, that is, given sufficient examples of inputs associated with one particular output, it should associate further inputs of the same type with that same output. The questions arises, however: What counts as the same type? The designer of the net has a specific definition in mind of the type required for a reasonable generalization, and counts it a success if the net generalizes to other instances of this type. But when the net produces an unexpected association can one say it has failed to generalize? One could equally well say that the net has all along been acting on a different definition of the type in question and that that difference has just been revealed. (All the "continue this sequence" questions found on intelligence tests really have more than one possible answer but most humans share a sense of what is simple and reasonable and therefore acceptable.)

Neural network modelers attempt to avoid this ambiguity and make the net produce "reasonable" generalizations by considering only a pre-specified allowable family of generalizations, i.e., allowable transformations which will count as acceptable generalizations (the hypothesis space). They then attempt to design the architecture of their nets so that the net transforms inputs into outputs only in ways which are in the hypothesis space. Generalization will then be possible only on the designer's terms. While a few examples will be insufficient to identify uniquely the appropriate member of the hypothesis space, after enough examples only one hypothesis will account for all the examples. The net will then have learned the appropriate generalization principle, i.e., all further input will produce what, from the designer's point of view, is the appropriate output.

The problem here is that the designer has determined by means of the architecture of the net that certain possible generalizations will never be found. All this is well and good for toy problems in which there is no question of what constitutes a reasonable generalization, but in real-world situations a large part of human intelligence consists in generalizing in ways appropriate to the context. If the designer restricts the net to a pre-defined class of appropriate responses, the net will be exhibiting the intelligence built into it by the designer for that context but will not have the common sense that would enable it to adapt to other contexts as would a truly human intelligence.

Perhaps a net must share size, architecture and initial connection configuration with the human brain if it is to share our sense of appropriate generalizations. If it is to learn from its own "experiences" to make associations that are human-like rather than be taught to make associations which have been specified by its trainer, it must also share our sense of appropriateness of outputs, and this means it must share our needs, desires, and emotions and have a human-like body with the same physical movements, abilities and possible injuries. If Heidegger and Wittgenstein are right, human beings are much more holistic than neural nets. Intelligence has to be motivated by purposes in the organism and other goals picked up by the organism from an on-going culture. If the minimum unit of analysis is that of a whole organism geared into a whole cultural world, neural nets as well as symbolically programmed computers, still have a very long way to go.

Hubert L. Dreyfus Stuart E. Dreyfus University of California, Berkeley