Also published in this series

The Concept of Evidence, edited by Peter Achinstein
Perceptual Knowledge, edited by Jonathan Dancy
The Philosophy of Law, edited by Ronald M. Dworkin
Theories of Ethics, edited by Philippa Foot
The Philosophy of History, edited by Patrick Gardiner
The Philosophy of Mind, edited by Jonathan Glover
Scientific Revolutions, edited by Ian Hacking
Divine Commands and Morality, edited by Paul Helm
Hegel, edited by Michael Inwood
The Philosophy of Linguistics, edited by Jerrold J. Katz
Reference and Modality, edited by Leonard Linsky
The Philosophy of Religion, edited by Basil Mitchell
The Concept of God, edited by Thomas V. Morris
A Priori Knowledge, edited by Paul K. Moser
Aesthetics, edited by Harold Osborne
The Theory of Meaning, edited by G. H. R. Parkinson
The Philosophy of Education, edited by R. S. Peters
Political Philosophy, edited by Anthony Quinton
Practical Reasoning, edited by Joseph Raz
The Philosophy of Social Explanation, edited by Alan Ryan
Propositions and Attitudes, edited by Nathan Salmon and Scott Soames
Consequentialism and its Critics, edited by Samuel Scheffler
The Philosophy of Language, edited by J. R. Searle
Semantic Syntax, edited by Pieter A. M. Seuren
Applied Ethics, edited by Peter Singer
Philosophical Logic, edited by P. F. Strawson
Locke on Human Understanding, edited by I. C. Tipton
Free Will, edited by Gary Watson
The Philosophy of Action, edited by Alan R. White
Leibniz: Metaphysics and Philosophy of Science, edited by R. S. Woolhouse

Other volumes are in preparation

THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE

Edited by

MARGARET A. BODEN

OXFORD UNIVERSITY PRESS

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Petaling Jaya Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press, New York

First published 1990
Paperback reprinted 1990

British Library Cataloguing in Publication Data
The Philosophy of artificial intelligence.—(Oxford readings in philosophy)
1. Artificial intelligence. Related to human cognition
2. Man. Cognition. Related artificial intelligence
I. Boden, Margaret A. (Margaret Ann), 1936–
006.3

ISBN 0–19–824855–5
ISBN 0–19–824854–7 (Pbk.)

Library of Congress Cataloging in Publication Data
The philosophy of artificial intelligence/edited by Margaret A. Boden.
p. cm.–(Oxford readings in philosophy)
Bibliography. Includes index.
1. Artificial intelligence—Philosophy. 2. Philosophy.
I. Boden, Margaret A. II. Series.
Q335.P48 1990 006.3'01–dc20 89–34075

ISBN 0–19–824855–5
ISBN 0–19–824854–7 (Pbk.)

Set by Litho Link Limited, Welshpool, Powys
Printed in Great Britain by
Courier International Ltd,
Tiptree, Essex.

CONTENTS

ABBREVIATIONS

AI Review	*Artifical Intelligence Review*
AISB Bulletin	*AISB Bulletin* (of the Society for the Study of Artificial Intelligence and Simulation of Behaviour)
Biol. Cybernetics	*Biological Cybernetics*
Brit. J. Phil. Science	*British Journal for the Philosophy of Science*
Commun. ACM	*Communications of the Association for Computing Machinery*
J. Experimental Psychol: General	*Journal of Experimental Psychology: General*
J. Philosophy	*Journal of Philosophy*
J. Theory of Social Behaviour	*Journal for the Theory of Social Behaviour*
Math. Comp.	*Mathematical Computing*
Phil. Trans. Roy. Soc. B	*Philosophical Transactions of the Royal Society, B*
Proc. & Addresses of Amer. Philos. Assoc.	*Proceedings and Addresses of the American Philosophical Association*
Proc. Aristotelian Soc.	*Proceedings of the Aristotelian Society*
Proc. 5th IJCAI Conference	*Proceedings of the Fifth International Joint Conference on Artificial Intelligence*
Proc. 1st AISB Conference	*Proceedings of the First AISB Conference* (of the Society for the Study of Artificial Intelligence and Simulation of Behaviour)
Proc. London Math. Society	*Proceedings of the London Mathematical Society*
Proc. Nat. Acad. Sci.	*Proceedings of the National Academy of Science*
Proc. Nat. Conf. AI	*Proceedings of the National Conference on Artificial Intelligence*
Proc. Roy. Soc. B	*Proceedings of the Royal Society, B*
Q. J. Exp. Psychol.	*Quarterly Journal of Experimental Psychology*
SIGART Newsletter	*SIGART Newsletter* (of the Special Interest Group on Artificial Intelligence, Association for Computing Machinery)
SIGSAM Bull., ACM	*SIGSAM Bulletin* (Association for Computing Machinery, Special Interest Group on Symbolic and Algebraic Manipulation)

Smolensky, P. (1986). 'Information Processing in Dymanical Systems: Foundations of Harmony Theory.' In McClelland, Rumelhart, and the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, pp. 194–281. Cambridge, Mass.: MIT/Bradford Books.

—— (1988). 'On the Proper Treatment of Connectionism.' *Behavioral and Brain Sciences* 11: 1–74.

13

MAKING A MIND VERSUS MODELLING THE BRAIN: ARTIFICIAL INTELLIGENCE BACK AT A BRANCH-POINT

HUBERT L. DREYFUS and STUART E. DREYFUS

> [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. 504(66e))

> [I]nformation is not stored anywhere in particular. Rather it is stored everywhere. Information is better thought of as 'evoked' than 'found'.
>
> David Rumelhart and Donald Norman (1981: 3)

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-programme, emerged and struggled for recognition. One faction saw computers as a system for manipulating mental symbols; the other, as a medium for modelling the brain. One sought to use computers to instantiate a formal representation of the world; the other, to simulate the interactions of neurones. 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 neuroscience.

The rallying cry of the first group was that both minds and digital computers are physical-symbol systems. By 1955 Allen Newell and

Herbert L. Dreyfus and Stuart E. Dreyfus, 'Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint' from *Artificial Intelligence* 117, no. 1 (Winter 1988), Cambridge, Mass. Reprinted by permission of *Daedalus*, Journal of the American Academy of Arts and Sciences.

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 (Newell 1983: 196).

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 a certain 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—a device that generated intelligent behaviour by manipulating symbols by means of formal rules. Newell and Simon stated their view as a hypothesis:

> *The Physical Symbol System Hypothesis.* 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 1981: 41).

Newell and Simon trace the roots of their hypothesis back to Gottlob Frege, Bertrand Russell, and Alfred North Whitehead (1981: 42), but Frege and company were of course themselves heirs to a long, atomistic, rationalist tradition. Descartes had already assumed that all understanding consisted of forming and manipulating appropriate representations, that these representations could be analysed into primitive elements (*naturas simplices*), and that all phenomena could be understood as complex combinations of these simple elements. Moreover, at the same time, Hobbes had implicitly assumed that the elements were formal components 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 . . . ' (1958: 45). 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 (1951: 18). According to Leibniz, in understanding we analyse concepts into more simple elements. In order to avoid a regress to 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 that these elements represent. Leibniz envisaged 'a kind of alphabet of human thoughts' (1951: 20) whose 'characters must show, when they are used in demonstrations, some kind of connection, grouping and order which are also found in the objects' (1951: 10).

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

> 1.1 The world is the totality of facts, not of things.

Facts in turn, he held, could be exhaustively analysed 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, Wittgenstein claimed, 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 (1960).

AI can be thought of as the attempt to find the primitive elements and logical relations in the subject (man or computer) that mirror the primitive objects and their relations that 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-programme on it.

The opposed intuition, that we should set about creating artificial intelligence by modelling the brain rather than the mind's symbolic representation of the world, drew its inspiration not from philosophy but from what was soon to be called neuroscience. It was directly inspired by the work of D. O. Hebb, who in 1949 suggested that a mass of neurones could learn if

when neurone A and neurone B were simultaneously excited, that excitation increased the strength of the connection between them.

This lead was followed by Frank Rosenblatt, who reasoned that since intelligent behaviour based on our representation of the world was likely to be hard to formalize. AI should instead attempt to automate the procedures by which a network of neurones 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 (1962*b*: 386).

Another way to put the difference between the two research-programmes 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, that could then 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 (1962*b*: 387).

Both approaches met with immediate and startling success. By 1956 Newell and Simon had succeeded 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 as if the physical-symbol systems 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 (1958: 6).

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 (1958: 6).[1]

Rosenblatt put his ideas to work in a type of device that he called a perceptron.[2] By 1956 Rosenblatt was able to train a perceptron to classify certain types of patterns as similar and to separate these from other patterns that 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 (1958: i. 449).

In the early sixties both approaches looked equally promising, and both made themselves equally vulnerable by making exaggerated claims. Yet

[1] Heuristic rules are rules that when used by human beings are said to be based on experience or judgement. Such rules frequently lead to plausible solutions to problems or increase the efficiency of a problem-solving procedure. Whereas algorithms guarantee a correct solution (if there is one) in a finite time, heuristics only increase the likelihood of finding a plausible solution.

[2] Rumelhart and McClelland (1986) describe the perceptron as follows: 'Such machines consist of what is generally called a *retina*, an array of binary inputs sometimes taken to be arranged in a two-dimensional spatial layout; a set of *predicates*, a set of binary threshold units with fixed connections to a subset of units in the retina such that each predicate computes some local function over the subset of units to which it is connected; and one or more decision units, with modifiable connections to the predicates' (i. 111). They contrast the way a parallel distributed processing (PDP) model like the perceptron stores information with the way information is stored by symbolic representation: 'In most models, knowledge is stored as a static copy of a pattern. Retrieval amounts to finding the pattern in long-term memory and copying it into a buffer or working memory. There is no real difference between the stored representation in long-term memory and the active representation in working memory. In PDP models, though, this is not the case. In these models, the patterns themselves are not stored. Rather, what is stored is the *connection strengths* between units that allow these patterns to be re-created (i. 31). . . . [K]nowledge about any individual pattern is not stored in the connections of a special unit reserved for that pattern, but is distributed over the connections among a large number of processing units' (i. 33). This new notion of representation led directly to Rosenblatt's idea that such machines should be able to acquire their ability through learning rather than by being programmed with features and rules: [I]f the knowledge is [in] the strengths of the connections, learning must be a matter of finding the right connection strengths so that the right patterns of activation will be produced under the right circumstances. This is an extremely important property of this class of models, for it opens up the possibility that an information processing mechanism could learn, as a result of tuning its connections, to capture the interdependencies between activations that it is exposed to in the course of processing' (i. 32).

the results of the internal war between the two research-programmes were 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 programmes, journals, and symposia that constitute a flourishing research-programme.

Reconstructing how this change came about is complicated by the myth of manifest destiny that any ongoing research-programme generates. Thus, it looks to the victors as if symbolic information processing won out because it was on the right track, while the neural network or connectionist approach lost because it simply didn't work. But this account of the history of the field is a retrospective illusion. Both research-programmes 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 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 would soon become intractable. In 1969 Marvin Minsky and Seymour Papert said 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 (19).

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 and David Zipser have succinctly summed it up, 'Combinatorial explosion catches you sooner or later, although sometimes in different ways in parallel than in serial (Rumelhart and McClelland 1986: i. 158). Both sides had, as Jerry Fodor once put it, walked into a game of 3-dimensional chess, thinking it was tick-tack-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 branch-point, 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 attacking the idea of the perceptron. 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 (1969: 4).

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 (1969: 19).

They were quite right. Artificial neural nets may, but need not, allow an interpretation of their hidden nodes[3] in terms of features a human being could recognize and use to solve the problem. While neural network modelling 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 their book suggests much more than it actually

[3] Hidden nodes are nodes that neither directly detect the input to the net nor constitute its output. They are, however, either directly or indirectly linked by connections with adjustable strengths to the nodes detecting the input and those constituting the output.

demonstrates. They set out to analyse the capacity of a one-layer perceptron,[4] while completely ignoring in the mathematical portion of their book Rosenblatt's chapters on multilayer machines and his proof of the convergence of a probabilistic learning algorithm based on back-propagation[5] of errors (1962*a*: 292).[6] 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 analysing computational systems (1986: i. 111).

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 (1986: i. 112).

Yet in the conclusion to *Perceptrons*, 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 (1969: 231–2).

Their attack on gestalt thinking in AI succeeded beyond their wildest dreams. Only an unappreciated few, among them Stephen Grossberg,

[4] A one-layer network has no hidden nodes, while multilayer networks do contain hidden nodes.

[5] Back-propagation of errors requires recursively computing, starting with the output nodes, the effects of changing the strengths of connections on the difference between the desired output and the output produced by an input. The weights are then adjusted during learning to reduce the difference.

[6] See also: 'The addition of a fourth layer of signal transmission units, or cross-coupling the A-units of a three-layer perceptron, permits the solution of generalization problems, over arbitrary transformation groups. . . . In back-coupled perceptrons, selective attention to familiar objects in a complex field can occur. It is also possible for such a perceptron to attend selectively to objects which move differentially relative to their background' (Rosenblatt 1962*a*: 576).

James A. Anderson, and Teuvo 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 (1986: i. 112).

But why was it enough? Both approaches had produced some promising work and some unfounded promises.[7] 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 that 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' (1981: 64).

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, and he had trouble getting his work published. By 1970, as far as AI was concerned, neural nets were dead. In his history of AI, Newell says the issue of symbols versus numbers 'is certainly not alive now and has not been for a long time' (1983: 10). Rosenblatt is not even mentioned in John Haugeland's (1985) or Margaret Boden's (1977) histories of the AI field.[8]

[7] For an evaluation of the symbolic representation approach's actual successes up to 1978, see Dreyfus (1979).

[8] Work on neural nets was continued in a marginal way in psychology and neuroscience. James A. Anderson at Brown University continued to defend a net model in psychology, although he had to live off other researchers' grants, and Stephen Grossberg worked out an elegant mathematical implementation of elementary cognitive capacities. For Anderson's position see Anderson (1978). For examples of Grossberg's work during the dark ages, see his (1982) book. Kohonen's early work is reported in *Associative Memory—A System-Theoretical Approach* (Berlin: Springer-Verlag, 1977). At MIT Minsky continued to lecture on neural nets and to assign theses investigating their logical properties. But according to Papert, Minsky did so only because nets had interesting mathematical properties, whereas nothing interesting could be proved concerning the properties of symbol systems. Moreover, many AI researchers assumed that since Turing machines were symbol manipulators and Turing had proved that Turing machines could compute anything, he had proved that all intelligibility could be captured by logic. On this view a holistic (and in those days statistical) approach needed justification, while the symbolic AI approach did not. This confidence, however, was based on confusing the uninterpreted symbols of a Turing machine (zeros and ones) with the semantically interpreted symbols of AI.

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 combinational puzzles. Even more important, it seemed that given the computing power available at the time, the neural-net researchers could do only speculative neuroscience 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 asumption 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, to look at things this way is to ignore both the pre-eminent role of pattern discrimination in human expertise and also the background of common-sense understanding that is presupposed in everyday real-world thinking. Taking account of this background may well require pattern recognition.

This thought brings us back to the philosophical tradition. It was not just Descartes and his descendants who 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 (Heidegger 1962: §14–21; Dreyfus 1988). This means that philosophy has from the start systematically ignored or distorted the everyday context of human activity.[9] The branch of the philosophical tradition that descends from Socrates through Plato, Descartes, Leibniz, and 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 among 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 of theories they have learned in another life, outside the everyday world. Once learned, such theories function in this world by controlling the thinker's mind, whether

[9] According to Heidegger, Aristotle came closer than any other philosopher to understanding the importance of everyday activity, but even he succumbed to the distortion of the phenomenon of the everyday world implicit in common sense.

he or she 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 among those elements that account 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* . . . [italics added] (1951: 48).

The symbolic information-processing approach gains its assurance from this transfer to all domains of methods that have been developed by philosophers and that are successful in the natural sciences. Since, in 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 to base a formal, symbolic representation on this theoretical analysis. In this vein Terry Winograd 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 (1976: 9).

No doubt theories about the universe are often built up gradually by modelling 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 lawlike relations between what Papert and Minsky call 'structural primitives'. Since no one *argues* for atomistic reduction in AI, it seems that AI workers just 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 assumption may well account for the way the physical-symbol system hypothesis so quickly turned into a revelation and for the ease with which Papert and Minsky's book triumphed over the holism of the perceptron.

Teaching philosophy at MIT in the mid-sixties, one of us—Hubert—was soon drawn into the debate over the possibility of AI. It was obvious that researchers such as Newell, Simon, and Minsky were the heirs to the philosophical tradition. But given the conclusions of the later Wittgenstein and the early Heidegger, that did not seem to be a good

omen for the reductionist research-programme. 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 *Tractatus*, his *Philosophical Investigations*, was published in 1953, just as AI took over the abstract, atomistic tradition he was attacking. After writing the *Tractatus*, Wittgenstein spent years doing what he called phenomenology (1975)—looking in vain for the atomic facts and basic objects his theory required. He ended by abandoning his *Tractatus* 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 intelligently in coping with everyday problems.

In the *Philosophical Investigations* Wittgenstein directly criticized the logical atomism of the *Tractatus*:

> 'What lies behind the idea that names really signify simples'?—Socrates says in the *Theaetetus*: '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' (*Tractatus Logico-Philosophicus*) 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' (1953: 21).

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 (Dreyfus 1982). 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.[10]

[10] 'Der Sinn . . . so wie wir ihn bestimmt haben, ist nicht ein konkretes Wesen im Gesamtbestande des Noema, sondern eine Art ihm einwohnender abstrackter Form.' See Husserl (1950). For evidence that Husserl held that the noema accounts for the intentionality of mental activity, see Hubert Dreyfus, 'Husserl's Perceptual Noema', in Dreyfus (1982).

This meaning, or symbolic representation, as conceived by Husserl, is a complex entity that has a difficult job to perform. In *Ideas Pertaining to a Pure Phenomenology* (1982), Husserl bravely tried 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 was 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 a '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' (1960: 45). The noema must contain a rule describing all the features that can be expected with certainty in exploring a certain *type* of object—features that remain 'inviolably the same: as long as the objectivity remains intended as *this* one and of this kind' (1960: 53). 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*. . . .' (1960: 51).

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* (1981: 96).

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 that 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. 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 by pointing 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 said, 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 part of our socialization, which 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 (1962: 121–1).

This defines the splitting of the ways between Husserl and AI on the one hand and Heidegger and the 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, discrimination, and so on, which are not intentional states and so, *a fortiori*, do not have any representational content to be explicated in terms of elements and rules?

By making a move that was soon to become familiar in AI circles, Husserl tried to avoid the problem Heidegger posed. Husserl 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. One could, in principle, he held, 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 behaviour. 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 (1970: 149).

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 that a formal axiomatic system might describe 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 left open the question whether these everyday concepts could none the less be formalized. (This was like raising and leaving open the AI question whether one can axiomatize common-sense physics.) Taking up Leibniz's dream of a mathesis of all experience, Husserl added:

> 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 (1952: v. 134).

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 the subject's common-sense 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 psycho-physical being and humanity as such, culture as such (1960: 54–5).

He spoke of the noema's 'huge concreteness' (1969: 244) and of its 'tremendous complication' (1969: 246), and he sadly concluded at the age of seventy-five that he was a perpetual beginner and that phenomenology was an 'infinite task' (1970: 291).

There are hints in his paper 'A Framework for Representing Knowledge' 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 (1981: 124).

To a student of contemporary philosophy, Minsky's naïveté and faith are astonishing. Husserl's phenomenology *was* just such a research effort. Indeed, philosophers from Socrates through Leibniz to early Wittgenstein 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, one of us—Hubert—predicted trouble for symbolic information-processing. As Newell notes in his history of AI, this 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 (1983: 222–3).

The trouble was, indeed, not long in coming to the fore, as the everyday world took its revenge on AI as it had on traditional philosophy. As we see it, the research-programme launched by Newell and Simon has gone through three ten-year stages. From 1955 to 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 a class of problems with the general heuristic search principle known as means–end analysis—namely, to 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 (1965–75), led by Marvin Minsky and Seymour Papert at MIT, 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 MIT 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 could learn 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 theatre, 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 theatre 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, subworlds are not related like isolable physical systems to the larger systems they *compose* but rather are local elaborations of a whole that they *presuppose*. Micro-worlds are not worlds but isolated meaningless domains, and it has gradually become clear that there is no way they could be combined and extended to arrive at the world of everyday life.

In its third 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-structures, 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 of twenty years ago that 'within twenty years machines will be capable of doing any work a man can do' (1965: 96).

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 course 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' (1984: 142).

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 pointed 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.[11] 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 that may not be formalizable, but the domain itself can be described by abstract laws that make no reference to these background skills. AI researchers mistakenly conclude that common-sense physics too must be expressible as a

[11] This account of skill is spelled out and defended in Dreyfus and Dreyfus (1986).

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 with all sorts of liquids and solids every day for several years, a child may simply learn to discriminate prototypical cases of solids, liquids, and so on and learn typical skilled responses to their typical behaviour in typical circumstances. The same might well be the case for the social world. If background understanding is indeed a skill and if skills are based on whole patterns and not on rules, we would expect symbolic representations to fail to capture our common-sense understanding.

In the light of this impasse, classical, symbol-based AI appears more and more to be a perfect example of what Imre Lakatos (1978) has called a degenerating research-programme. As we have seen, AI began auspiciously with Newell and Simon's work at Rand and by the late 1960s turned into a flourishing research-programme. Minsky predicted that 'within a generation the problem of creating "artificial intelligence" will be substantially solved' (1977: 2). 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 centre of concern. Minsky's mood changed completely in five years. He told a reporter that 'the AI problem is one of the hardest science has ever undertaken' (Kolata 1982: 1237).

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 of representing that theory in a symbol manipulator had run into just the difficulties Heidegger and Wittgenstein had discovered. Frank Rosenblatt's intuition that it would be hopelessly difficult to formalize the world and thus to give a formal specification of intelligent behaviour had been vindicated. His repressed research-programme (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' (McCorduck 1979: 87).

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.[12] Now he is being rehabilitated. David Rumelhart, Geoffrey Hinton, and James 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 (1986: i. 45).
>
> 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 (1986: ii. 535).

Frustrated AI researchers, tired of clinging to a research-programme that Jerry Lettvin characterized in the early 1980s as 'the only straw afloat', flocked to the new paradigm. Rumelhart and McClelland's book

[12] Some typical quotations from Rosenblatt's *Principles of Neurodynamics*: 'In a learning experiment, a perceptron is typically exposed to a sequence of patterns containing representatives of each type or class which is to be distinguished, and the appropriate choice of a response is "reinforced" according to some rule for memory modification. The perceptron is then presented with a test stimulus, and the probability of giving the appropriate response for the class of the stimulus is ascertained. . . . If the test stimulus activates a set of sensory elements which are entirely distinct from those which were activated in previous exposures to stimuli of the same class, the experiment is a test of "pure generalization." The simplest of perceptrons . . . have no capability for pure generalization, but can be shown to perform quite respectably in discrimination experiments particularly if the test stimulus is nearly identical to one of the patterns previously experienced (p. 68). . . . Perceptrons considered to date show little resemblance to human subjects in their figure-detection capabilities, and gestalt-organizing tendencies (p. 71). . . . The recognition of sequences in rudimentary form is well within the capability of suitably organized perceptrons, but the problem of figural organization and segmentation presents problems which are just as serious here as in the case of static pattern perception (p. 72). . . . In a simple perception, patterns are recognized before "relations"; indeed, abstract relations, such as "A is above B" or "the triangle is inside the circle" are never abstracted as such, but can only be acquired by means of a sort of exhaustive rote-learning procedure, in which every case in which the relation holds is taught to the perceptron individually (p. 73). . . . A network consisting of less than three layers of signal transmission units, or a network consisting exclusively of linear elements connected in series, is incapable of learning to discriminate classes of patterns in an isotropic environment (where any pattern can occur in all possible retinal locations, without boundary effects) (p. 575). . . . A number of speculative models which are likely to be capable of learning sequential programs, analysis of speech into phonemes, and learning substantive "meanings" for nouns and verbs with simple sensory referents have been presented in the preceding chapters. Such systems represent the upper limits of abstract behaviour in perceptrons considered to date. They are handicapped by a lack of satisfactory 'temporary memory,' by an inability to perceive abstract topological relations in a simple fashion, and by an inability to isolate meaningful figural entities, or objects, except under special conditions (p. 577). . . . The applications most likely to be realizable with the kinds of perceptrons described in this volume include character recognition and "reading machines", speech recognition (for distinct, clearly separated words), and extremely limited capabilities for pictorial recognition, or the recognition of objects against simple backgrounds. "Perception" in a broader sense may be potentially within the grasp of the descendants of our present models, but a great deal of fundamental knowledge must be obtained before a sufficiently sophisticated design can be prescribed to permit a perceptron to compete with a man under normal environmental conditions' (p. 583).

Parallel Distributed Processing sold six thousand copies the day it went onto the market, and thirty thousand 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 (forthcoming).

If multilayered networks succeed in fulfilling their promise, researchers will have to give up the conviction of Descartes, Husserl, and early Wittgenstein that the only way to produce intelligent behaviour 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—that is, 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 behaviour, we have to be prepared to raise the question whether in everyday domains such a theoretical explanation is even *possible*.

Neural net modellers, influenced by symbol-manipulating AI, are expending considerable effort, once their nets have been trained to perform a task, in trying to find the features represented by individual nodes and sets of nodes. Results thus far are equivocal. Consider Hinton's (1986) network for learning concepts by means of distributed representations. The network can be trained to encode relationships in a domain that 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 some nodes in the trained network can be interpreted as corresponding to the features that human beings pick out, although these nodes only roughly correspond to those 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 not only varies 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—that is constructed by human beings precisely in terms of the features that human beings normally notice, such as generation and nationality. Hinton then analyses those cases in which, starting with certain random initial-

connection strengths, some nodes can, after learning, be interpreted as representing those features. Calculations using Hinton's model show, however, that even his net seems to learn its associations for some random initial-connection strengths without any obvious use of these everyday features.

In one very limited sense, any successfully trained multilayer net can be interpreted in terms of features—not everyday features but what we shall call highly abstract features. Consider the simple case of layers of binary units activated by feed-forward, but not lateral or feedback, connections. To construct such an account from a network that has learned certain associations, each node one level above the input nodes could, on the basis of the 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 that 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 could be characterized as a feature detector. Similarly, every node a level above those nodes could be interpreted as detecting a higher-order feature, defined as the presence of one of a specified set of patterns among the first level of feature 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 so that one could base a theory on them. If the net were taught one more association of an input–output pair (where the input prior to training produced an output different from the one to be learned), the interpretation of at least some of the nodes would have to be changed. So the features that 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 net modelling, 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 modellers have so far had difficulty dealing with stepwide 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 continues between intellectualists, who think that because they can do context-free logic 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 no account of everyday cognition.[13] One might think, using the metaphor of the right and the left brain, that perhaps the brain or the mind uses each strategy when appropriate. The problem would then be how to combine the strategies. 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 modelling may simply be getting a deserved chance to fail, as did the symbolic approach.

Still, there is an important difference to bear in mind as each research-programme 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 modelling, however, is not committed to this or any other philosophical assumption. Nevertheless, 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 researchers may not yet recognize it. All neural net modellers 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 question arises, however: what counts as the same type? The designer of the net has in mind a specific definition 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

[13] For a recent influential account of perception that denies the need for mental representation, see Gibson (1979). Gibson and Rosenblatt collaborated on a research paper for the US Air Force in 1955; see Gibson, Olum, and Rosenblatt (1955).

'continue this sequence' questions found on intelligence tests really have more than one possible answer, but most human beings share a sense of what is simple and reasonable and therefore acceptable.)

Neural network modellers attempt to avoid this ambiguity and make the net produce 'reasonable' generalizations by considering only a prespecified allowable family of generalizations—that is, allowable transformations that will count as acceptable generalizations (the hypothesis space). These modellers then attempt to design the architecture of their nets so that they transform inputs into outputs only in ways that 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. That is, 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 that are appropriate to a context. If the designer restricts the net to a predefined 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 a truly human intelligence would.

Perhaps a net must share size, architecture, and initial-connection configuration with the human brain if it is to share our sense of appropriate generalization. If it is to learn from its own 'experiences' to make associations that are humanlike rather than be taught to make associations that have been specified by its trainer, a net must also share our sense of appropriateness of output, and this means it must share our needs, desires, and emotions and have a humanlike body with appropriate physical movements, abilities, and vulnerability to injury.

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 goals picked up by the organism from an ongoing 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.

REFERENCES

Anderson, J. A. (1978). 'Neural Models with Cognitive Implications.' In D. LaBerse and S. J. Samuels (eds.), *Basic Processing in Reading*, Hillsdale, NJ: Erlbaum.

Boden, M. (1977). *Artificial Intelligence and Natural Man*. New York: Basic Books.

Dreyfus, H. (1979). *What Computers Can't Do*, 2nd edn. New York: Harper & Row.

—— (1988). *Being-in-the-World: A Commentary on Division I of 'Being and Time'*. Cambridge, Mass.: MIT Press.

—— (ed.) (1982). *Husserl, Intentionality and Cognitive Science*. Cambridge, Mass.: MIT Press.

—— and Dreyfus, S. (1986). *Mind Over Machine*. New York: Macmillan.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin.

—— Olum, P., and Rosenblatt, F. (1955). 'Parallax and Perspective During Aircraft Landing.' *American Journal of Psychology* 68: 372–85.

Grossberg, S. (1982). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition and Motor Control*. Boston: Reidel Press.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, Mass.: MIT Press.

Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley.

Heidegger, M. (1962). *Being and Time*. New York: Harper & Row.

Hinton, G. (1986). 'Learning Distributed Representations of Concepts.' In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Amherst, Mass.: Cognitive Science Society.

Hobbes, T. (1958). *Leviathan*. New York: Library of Liberal Arts.

Husserl, E. (1950). *Ideen Zu Einer Reinen Phänomenologie und Phänomenologischen Philosophie*. The Hague: Nijhoff.

—— (1952). *Ideen Zu Einer Reinen Phänomenologie und Phenomenologischen Philosophie*, bk. 3 in vol. 5, *Husserliana*. The Hague: Nijhoff.

—— (1960). *Cartesian Meditations*, trans. D. Cairns. The Hague: Nijhoff.

—— (1969). *Formal and Transcendental Logic*, trans. D. Cairns. The Hague: Nijhoff.

—— (1970). *Crisis of European Sciences and Transcendental Phenomenology*, trans. D. Carr. Evanston: Northwestern University Press.

—— (1982). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*, trans. F. Kersten. The Hague: Nijhoff.

Kohonen, T. (1977). *Associative Memory: A System-Theoretical Approach*. Berlin: Springer-Verlag.

Kolata, G. (1982). 'How Can Computers Get Common Sense?' *Science* 217 (24 Sept.): 1237.

Lakatos, I. (1978). *Philosophical Papers*, ed. J. Worrall. Cambridge: Cambridge University Press.

Leibniz, G. (1951). *Selections*, ed. P. Wiener. New York: Scribner.

Lighthill, Sir James (1973). 'Artificial Intelligence: A General Survey.' In *Artificial Intelligence: A Paper Symposium*. London: Science Research Council.

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

Minsky, M. (1977). *Computation: Finite and Infinite Machines*. New York: Prentice-Hall.

—— (1981). 'A Framework for Representing Knowledge.' In J. Haugeland (ed.), *Mind Design*, pp. 95–128. Cambridge, Mass.: MIT Press.

—— and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Mass.: MIT Press.

Newell, A. (1983). 'Intellectual Issues in the History of Artificial Intelligence.' In F. Machlup and U. Mansfield (eds.), *The Study of Information: Interdisciplinary Messages*, pp. 196–227. New York: Wiley.

—— and Simon, H. (1958). 'Heuristic Problem Solving: The Next Advance in Operations Research.' *Operations Research* 6 (Jan.–Feb.): 6.

—— (1981). 'Computer Science as Empirical Inquiry: Symbols and Search.' In J. Haugeland (ed.), *Mind Design*, pp. 35–66. Cambridge, Mass.: MIT Press.

Rosenblatt, F. (1958). *Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory*, Vol. 1. London: HMS Office.

—— (1962*a*). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books.

—— (1962*b*). 'Strategic Approaches to the Study of Brain Models.' In H. von Foerster (ed.), *Principles of Self-Organization*, Elmsford, NY: Pergamon Press.

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2 vols. Cambridge, Mass.: MIT Press.

—— and Norman, D. A (1981). 'A Comparison of Models.' In G. Hinton and J. Anderson (eds.), *Parallel Models of Associative Memory*, pp. 3–6. Hillsdale, NJ: Erlbaum.

Simon, H. (1965). *The Shape of Automation for Men and Management*. New York: Harper & Row.

Smolensky, P. [1988]. 'On the Proper Treatment of Connectionism.' *Behavioral and Brain Sciences* [11: 1–74].

Winograd, T. (1976). 'Artificial Intelligence and Language Comprehension.' In *Artificial Intelligence and Language Comprehension*, Washington, DC: National Institute of Education.

—— (1984). 'Computer Software for Working with Language. *Scientific American* (Sept.): 142ff.

Wittgenstein, L. (1948). *Last Writings on the Philosophy of Psychology*, Vol. 1, trans. corrected. Chicago: University of Chicago Press, 1982.

—— (1953). *Philosophical Investigations*. Oxford: Basil Blackwell.

—— (1960). *Tractatus Logico-Philosophicus*. London: Routledge & Kegan Paul.

—— (1975). *Philosophical Remarks*. Chicago: University of Chicago.