Problems in the Behavioural Sciences

Computer Models of Mind

MARGARET A. BODEN

Computer models of mind

Computational approaches in theoretical psychology

Margaret A. Boden, F.B.A.
Professor of Philosophy and Psychology, The University of Sussex

The right of the University of Cambridge to print and sell all manner of books was granted by Henry VIII in 1534. The University has printed and published continuously since 1584.

CAMBRIDGE UNIVERSITY PRESS
Cambridge
New York New Rochelle Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
32 East 57th Street, New York, NY 10022, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1988

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Boden, Margaret A.

Computer models of mind.

(Problems in the behavioural sciences)

Bibliography: p.
Includes index.

1. Psychology – Computer simulation. 2. Psychology – Philosophy. I. Title. II. Series.
BF39.5.B63 1988 150'.724 87-25625

British Library Cataloguing in Publication Data

Boden, Margaret A.

Computer models of mind: computational approaches in theoretical psychology – (Problems in the behavioural sciences; 5).

1. Psychology – Experiments 2. Electronic digital computers
I. Title II. Series
150'.724 BF198.5

ISBN 0 521 24868 X hard covers
ISBN 0 521 27033 2 paperback

Preface

This book asks how computer models have been used, and might be used, to help us formulate psychological theories about the mind.

The models and concepts discussed here were selected for their psychological significance, not their technological promise. This is not a sharp divide, for even technologically motivated work may involve matters of psychological interest; "expert systems", for instance, raise questions about how people store and communicate knowledge, and how it is transformed as expertise grows. However, I have concentrated on computer models whose psychological relevance is comparatively direct. Usually, though not always, these models were intended by their creator as simulations of mental processes or tests of psychological theories.

The program-details of a few of the computer models mentioned in this book – specifically, scene-analysis programs and some very early simulations of language-understanding, problem-solving, and learning – were described in much greater detail in my *Artificial Intelligence and Natural Man* (first published in 1977; second edition, including a new "Postscript" chapter and additional bibliography, 1987). Most of the research discussed in this volume, however, was mentioned there only briefly – if at all. This is partly because some of it is very recent, but is owing also to the different aims of the two books. There, I concentrated on explaining just how AI-programs work, and referred in a very general way to the wider psychological, philosophical, and social implications of artificial intelligence. Here, I focus throughout on particular questions within theoretical psychology, ranging from domain-specific topics to more general methodological issues.

In short, I aim to clarify the ways in which practising psychologists can use computational ideas and computer-modelling to further their research.

M.A.B.

Brighton, England

Acknowledgments

I am grateful to the following friends and colleagues for their helpful comments on various chapters (they should not be blamed for any mistakes which remain): Andy Clark, Gerald Gazdar, Geoffrey Hinton, David Hogg, Rudi Lutz, George Mather, Chris Mellish, Mike Scaife, Aaron Sloman, and Stephanie Thornton. Rickie Dammann drew my attention to the quotation I have used as the motto. Alison Mudd, with her usual good-humoured patience, undertook the chore of coaxing various computer printers to transfer my drafts from disc to paper. I am indebted to the Series Editor, Jeffrey Gray, for his constructive advice and encouragement. And I thank Jacqueline Korn, Jeremy Mynott, and Susan Milmoe, all of whose understanding and help have been invaluable.

I am obliged also to the Economic and Social Research Council for granting me a Personal Research Fellowship, and to the University of Sussex for allowing me a year's leave to take advantage of it.

Contents

1 Introduction

Computational psychology today is rather like the dragon in earlier times: of the people who are seeking it, not all agree on what they expect to find or how they hope to find it–while many others doubt that it exists to be found at all. My aim in this book is to explore the theoretical diversity of work in computational psychology, and the various controversial philosophical assumptions that underlie it. In doing so, I shall examine selected examples of psychological research using the methodology of computer-modelling.

Working models of living creatures are not new. They have existed as chic toys for many centuries: in the palaces of ancient Alexandria, the courts of medieval Islam, and the estates of the eighteenth-century European aristocracy. For over a hundred years they have cavorted in the pages of fiction, focussing cultural fears and fantasies of various kinds. But only very rarely in past centuries was the model-building motivated by the desire for theoretical understanding of men or animals. Even then (as in Vaucanson's nimble-fingered, fast-tongued flute-player), the focus was on bodily rather than mental processes. The attempt to build working models of intelligence that might help us to understand the nature and functioning of the mind is a very recent one.

By the mid twentieth century, speculative sketches of artificial animals could be found in the psychological literature. For example, in 1939 E. C. Tolman analysed trial-and-error learning in terms of an imaginary "schematic sowbug", and in 1943 C. L. Hull fantasized self-maintaining robots whose behaviour would be complex and adaptive like ours. The mathematician A. M. Turing was already discussing the logical properties of various possible types of computing machine in the 1930s [Turing, 1936]. With the rise of cybernetics in the early 1940s even purposive behaviour was characterized (by N. Wiener [1948] and others) as a basically mechanistic phenomenon.

At about the same time, K. J. W. Craik [1943] argued that psychological explanations should refer to cerebral models, conceived of as representational mechanisms (functioning "in the same way" as the phenomena being represented) capable of generating behaviour and thought of various kinds. But even Craik, influential though he was, could provide only the outlines of a new philosophical approach to psychology. He could not suggest any specific information-processing mechanisms which might be concerned, still less how such mechanisms might be built. Detailed theoretical hypotheses about mental representations had to await the development of the digital computer. Only then could mental modelling be seriously considered as a theoretical or practical enterprise.

The first such machines were being designed in the early 1940s, most importantly by J. von Neumann in the USA. (Turing and some others in England were tackling this problem too, and built the first electronic computer; but their work was top-secret for years because of its use in deciphering the ENIGMA code in the Second World War [Hodges, 1983].) In designing the digital computer, von Neumann was influenced not only by Turing's earlier work on the theory of computation, but also by some novel ideas about the logical functions of the brain. These were largely due to the physiologist and psychiatrist W. S. McCulloch and the mathematician W. H. Pitts [1943]. (Unlike McCulloch and Pitts, however, von Neumann did not believe that binary logic could model human thought: as he put it, "the language of the brain is not the language of mathematics"; his suggestion that thermodynamic probability is better suited has recently been revived, as we shall see in Chapter 7 [von Neumann, 1958, p. 80].)

In their paper "A Logical Calculus of the Ideas Immanent in Nervous Activity", McCulloch and Pitts [1943] compared the logical circuitry of digital computers with sets of interconnected neurones. Equating the on–off binary states of the computer to the all-or-none properties of nervous conduction, they claimed that the logical properties of the brain as a whole could be understood in terms of the logical properties of its constituent cells. Focussing on the computational or information-processing potential of neurones rather than their actual physiology, they argued that specific types of neuronal unit, functioning under particular constraints, would have specifiable logical properties.

For example, a certain simple arrangement of threshold-units would function as an "AND-gate", in which one cell in effect detects the simultaneous excitation of two others. Another primitive network embodies what logicians term exclusive OR, since one cell fires if and only if one (but not both) of two other cells is firing. Yet another computes AND-NOT, so that Cell 3 fires if and only if Cell 1 is firing and Cell 2 is not. More complex dispositions of units would compute not logical connectives such as AND, OR, and AND-NOT, but entire logical expressions–such as "A and (B or [C and D] but not both) and/or E and not-F". More generally, McCulloch and Pitts proved that every finite expression of the propositional calculus can in principle be computed by some neural network of the general type they described.

This paper, together with related work of McCulloch and Pitts [McCulloch, 1965], was very influential. The physiological psychologist D. O. Hebb [1949], introducing his seminal theory of *cell-assemblies,* saw "great potential value" in the McCulloch–Pitts project of describing the functional organization of the cerebral cortex in terms of mathematical analyses of the properties of groups of interacting neurones. Other psychologists were prompted to take an interest in computer models of various sorts. Very soon, attempts were made to build working models of mindlike functions, such as pattern-recognition, goal-directed behaviour, and logical thought. Early examples included Grey Walter's [1953] light-seeking tortoises and various computer simulations of neural nets, like F. Rosenblatt's [1958] *perceptrons,* reflecting physiological ideas (including Hebb's) about neural excitation and inhibition.

The late 1950s saw two distinct approaches to computer-modelling. Both owed much to the example set by McCulloch and Pitts. But whereas one stemmed from their (and Hebb's) ideas about the neurophysiological structure of the brain, the other owed more to their work on the embodiment of the propositional calculus in digital computers. On the one hand, there were early computer models of learning and pattern-recognition, inspired by Rosenblatt's work on perceptrons: O. G. Selfridge's [1959] Pandemonium was an influential example. On the other hand, there were early problem-solving computer programs such as the Logic Theorist [Newell, Shaw & Simon, 1957], and its successor the General Problem Solver (GPS) [Newell & Simon, 1961]. Crude though these early systems were, they raised some interesting theoretical problems. For example, it is still a controversial question (as we shall see) whether the essence of psychological processing was better represented by Pandemonium or by GPS.

Some of these pioneering models were described in *The Mechanization of Thought Processes* (1959), a report of a symposium (at the National Physical Laboratory) published by Her Majesty's Stationery Office and not easily accessible to the uninitiated, including most psychologists. But in the early 1960s, two books appeared that established the psychological visibility of this new way of thinking about thinking, and indicated some long-term goals.

One, significantly entitled *Computers and Thought,* was a collection of papers describing most of the then existing computer models of psychological interest [Feigenbaum & Feldman, 1963]. In addition to GPS, these included models of pattern-recognition, learning, concept-formation, geometrical reasoning, social interaction, and memory. It also contained more wide-ranging discussions, such as Turing's [1950] classic paper "Computing Machinery and Intelligence" and M. L. Minsky's [1961] speculations on the steps needed to achieve artificial intelligence in the future.

The other was *Plans and the Structure of Behavior* by G. A. Miller, E. Galanter, and K. H. Pribram [1960]. This book criticized S–R behaviourism, then the psychological orthodoxy, for ignoring what went on at the hyphen: namely, the unobservable mental processes intervening between stimulus and response. The authors looked to computer-modelling for concepts with which to describe these mental processes. They offered a sustained, if highly speculative, argument that the entire psychological spectrum – from instinct and motor control, through memory and language, to personality, psychopathology, and hypnosis – should be painted in computational colours. Strongly influenced by current advances in automatic problem-solving (notably GPS), they claimed that psychological explanations should specify hierarchical *plans* of various types. These plans are goal-directed procedures, and in conscious organisms many goal-states are represented and evaluated in the "image", a data-structure that reflects the individual's motivation and cultural interests. The proper aim of psychology, on this view, is the identification of procedures essentially comparable to computer programs.

Miller and his co-authors acknowledged an important intellectual debt to K. S. Lashley and N. Chomsky. Lashley had argued (in "The Problem of Serial Order in Behavior" [1951]) that all motor-skills – including speech – have a many-levelled hierarchical structure, which cannot be explained by Sherringtonian neurophysiology or behaviourism but requires us to postulate central controlling mechanisms in the brain. Chomsky [1957] too had stressed the hierarchical structure of language, and in a paper jointly written with Miller ("Finite State Languages" [1958]) had presented a mathematical proof of Lashley's insight that this cannot be explained behaviouristically. His *Syntactic Structures* [1957] had even offered a formal account of language in terms of a specific set of generative rules.

Nevertheless, the way in which Miller and his colleagues wished to explain language (and other psychological phenomena) differed significantly from Chomsky's approach. They thought of these phenomena as procedures, much like computer programs. But Chomsky's formal rules were not programs, nor were they primarily intended as specifications of actual psychological processes. Rather, they were abstract descriptions of structural relationships. They were "generative" in the timeless mathematical sense, whereby a formula is said to generate a series of numbers, not in the sense of being descriptions of a temporal process of generation. Similarly, his "transformations" were abstract mappings from one structure to another (as a square-root function transforms 9 into 3), not actual psychological changes or mental events. Likewise, his "finite-state and non-finite *machines*" were mathematical definitions (as are Turing machines), not descriptions of any actual manufactured systems that might conform to those definitions.

So in characterizing psychology as the study of plans, Miller and his co-authors were recommending a rather different sort of enterprise from Chomsky's. Many people would say that this was because they were psychologists whereas Chomsky was a linguist: they were interested in how language happens in the mind, whereas he was concerned only with language in itself. This is true, as far as it goes. But it obscures an important controversy. For some influential voices now argue that Chomsky's approach, rather than Miller's, is the more fundamental to a scientific psychology, that psychologists should focus primarily on *what* is computed, not on *how* it is computed. Like the controversy pitting Pandemonium against GPS, this dispute illustrates our main theme: the diverse theoretical and methodological assumptions informing computational psychology.

Computational psychology was compared, above, to dragons–about whose lineaments few agree and none is certain. Computational psychology is a broad church, whose members differ significantly about general methodology as well as specific detail. However, much as medieval bestiaries depicted their diverse dragons in broadly similar ways, so computational psychologists, despite their many differences, share certain very general philosophical assumptions about what it is they are looking for.

They all focus on the mind, considered as an informational not an energetic system, and take psychology to be the study of *mental* life. But so do many (non-behaviourist) psychologists who cannot be termed computationalists. And very many of them use computers–but again, so do many other psychologists (for calculating statistical significance, for example). The point is that computers–or rather, concepts drawn from computer science–play some central theoretical role in the computational psychologist's claims about what the mind is and how it functions. Accordingly, *computational* (as opposed to computer-using) psychologists share three characteristic ways of theorizing.

First, computational psychologists adopt a functionalist approach to the mind, in which mental states are abstractly defined in terms of their causal role (with respect to other mental states and observable behaviour). Functionalism is a philosophy of the late twentieth century in which the mind is conceived of in terms of the computational properties of universal Turing machines. Every psychological phenomenon (or at least every such phenomenon that is potentially capable of scientific explanation) is assumed to be generated by some *effective procedure,* some precisely specifiable set of instructions defining the succession of mental states within the mind. Since computer science is the study of effective procedures, this psychological approach takes computational concepts seriously.

Second, computational psychologists conceive of the mind as a representational system, and see psychology as the study of the various computational processes whereby mental representations are constructed, organized, interpreted, and transformed. A corollary is that they use *intentional* terminology (often including much of the vocabulary of everyday folk-psychology). That is, they think of (many) mental phenomena as having a meaning, or semantic content, as being directed upon some object or imaginary object outside the mind itself. (So, of course, do humanistic or hermeneutic psychologists, whose stress on intentionality and interpretation is closer in spirit to that of computational psychology than they believe.)

And third, they think about neuroscience (if they think about it at all) in a broadly computational way, asking what sorts of logical operations or functional relations might be embodied in neural networks. What the brain does that enables it to embody the mind is the main question, not what it is in itself as a physical system. As one such theorist has put it: "finding a cell that recognizes one's grandmother does not tell you very much more than you started with; after all, you know you can recognize your grandmother. What is needed is an answer to how you, or a cell, or anything at all, does it. The discovery of the cell tells one what does it, but not how it can be done" [Mayhew, 1983, p. 214].

These three process-oriented characteristics constitute a minimal definition of the family "computational psychology". In subsequent chapters we shall examine some fundamental ways in which the family members differ.

We shall see, for instance, that computational psychologists may mean rather different things by "computation" and by "representation". Whereas some theorists focus on symbol-manipulation defined in terms of formal rules, and take the (von Neumann) general-purpose digital computer as their inspiration, others emphasize parallel-processing computational networks whose behaviour is not defined by such rules. The former use a research methodology based on the ideas of orthodox artificial intelligence (AI), while the latter, no less sympathetic to computer-modelling in general, sometimes argue that these particular ideas have fundamentally misled us for the last quarter-century.

Artificial intelligence developed out of the mid-century experiments with complex programming mentioned above. Its goal is to understand, whether for theoretical or technological purposes, how representational structures can generate behaviour and how intelligent behaviour can emerge out of unintelligent behaviour. So AI-workers attempt to write computer programs (and/or to design machines) capable of performing complex information-processing tasks with a high degree of flexibility and context-sensitivity, like those faced by human and animal minds–such as perception, language, memory, motor-control, and thinking. This requires the rigorous expression of diverse symbolic processes, in terms of a rich variety of computational concepts specifically developed for managing informational complexity.

Computationally inclined psychologists in general see AI as potentially useful, largely because its conceptual focus is on representation and processes of symbolic transformation. (Perhaps one should rather say that these *appear to be* and *are generally taken to be* the conceptual focus of AI, so as not to beg the question, discussed in Chapter 8, whether a computer can properly be said to symbolize or represent anything in the full sense.) Moreover, they recognize that AI's emphasis on rigour encourages psychologists to be more precise, often pointing to unsuspected theoretical lacunae. And they appreciate that computer-modelling offers a manageable way of representing complexity, since the computational power of a computer can be used to infer the implications of a program where the unassisted human mind is unable to do so.

For these reasons, and for historical reasons too, many of the computer models discussed in the following chapters were developed in close relation to, or even within, artificial intelligence.

But this does not mean that all computational psychologists feel that AI has lived up to its early promise, or wish to be closely identified with it. As we shall see, some criticize AI as being "merely empirical", in the sense that it often achieves practical results by methods it does not understand and so cannot responsibly generalize. Such critics complain that there is, as yet, too little theoretical work in AI: too few proofs that a particular class of computation can or cannot work, given certain types of computational machine, and too few principled accounts of the advantages and disadvantages of distinct classes of program.

When these psychologists regard a program as interesting, they do so not because it achieves a particular result, but because the programmer attempts a general proof that results of this class can be computed by computational systems of this form, given certain specific constraints (which may apply to naturally evolved psychological systems). Indeed, there may not even be a program, but only an abstract analysis of the information-processing task in question. Such theorists agree with Chomsky in stressing the *what* of computation, rather than the *how*. Accordingly, they may use the term "computational" to refer not (as is more usual) to computational *processes*, but to the *abstract analysis of the task* facing the psychological system–irrespective of how it is actually performed. (This maverick non-procedural usage is introduced in Chapter 3, in relation to D. Marr's views on the "computational level" of explanation. When it is used in the succeeding chapters, the context or an explicit reference to Marr should prevent confusion with the more familiar sense of the term.)

Still less does the intellectual debt to AI mean that all computational psychologists favour the particular type of computer-modelling typical of orthodox AI-research (what J. Haugeland [1985] has termed GOFAI: Good Old-Fashioned AI). As mentioned above, some of them criticize mainstream AI for ignoring, at least until recently, the potential of types of computation that are not well-suited to the von Neumann machines traditionally used in AI. Instead, they recommend the study of parallel-processing, "connectionist", systems. These do not function by following explicitly represented rules, and concepts or representations are embodied in them as *patterns of activity* across entire networks of computational units. (At present, connectionist models have to be simulated on von Neumann machines; but connectionist hardware is being developed by various groups [e.g., Hillis, 1985].)

Another disagreement concerns the theoretical significance (for psychology) of neuroscience. Most computational psychologists (and AI-researchers) ignore physiology, arguing that computational questions are essentially distinct from questions of physical mechanism and so can be asked in their own terms. Indeed, the irrelevance of neurophysiology has acquired the status of a dogma in some circles. Heresy, however, abounds: there are opposing groups who champion the application of neurophysiological insights in formulating the basic outlines of their computer models.

The additional argument is often given (by the dogmatists) that we know so little about the neurophysiology of mental phenomena that it would be folly to try to constrain psychological theories and computer models by neurophysiological knowledge. The strongest counterexample concerns the physiology of some (low-level) aspects of vision. This is not only known to rest on massively parallel computation, but also involves cells (responsible for the early stages of visual processing) which are relatively accessible to neuroscientific investigations. Accordingly, it has prompted the formation of those connectionist psychological theories and computer models which attempt the most detailed neurophysiological interpretations. Parallelist theories of other mental processes (such as language or memory) are constrained by neurophysiology only in a much more general sense.

Again, some computationalists insist that the only intellectually responsible way to present a psychological theory is to embody it in a functioning computer model, whether von Neumann or not. Others see programming and model-building as largely irrelevant activities, although they plan their experiments and formulate their theories with computational questions in mind. People disagree, moreover, about how empirical data can be used to validate a theory presented in programmed form. And they differ over how we can decide which features of the program are really relevant to the theory concerned, and which are there only because the theorist had to write a program that would actually run.

Further controversy attends the question whether computational psychology can in principle explain the higher mental processes (such as problem-solving and memory). Many believe that it can. But others, who agree that problem-solving and memory are computational phenomena, hope for a scientific explanation only of peripheral processes. This controversy rests partly on varying ideas about what the aims of science are – specifically, whether they must include detailed prediction or detailed post-diction after the fact.

Finally, most computational psychologists trust that their approach will explain how representations function, and many believe it will even help to illuminate what representations are. But a few, together with many critics outside the computationalist camp, argue that computer models (and psychological theories grounded in them) in principle cannot exhibit or explain genuine *representation* (or *meaning*) at all.

Given these myriad disagreements, it is clear that computational psychology is a beast with many different incarnations. These species must be described in giving an account of the genus as a whole. So, if my theme is the nature–and the theoretical diversity–of computational psychology overall, my strategy is to proceed from the particular to the general.

In Chapters 2 to 7, various philosophical and methodological claims are introduced in relation to specific computer models of psychological processes, and to the psychological theories and experimental programmes associated with them. General concepts such as *representation, functional architecture, task-analysis, connectionism,* and *modularity* enter the text in the context of modelling imagery (Chapter 2) or low-level vision (Chapter 3), but appear also in the subsequent discussions of computer models of language (Chapters 4 and 5), problem-solving (Chapter 6), and learning (Chapter 7). Finally, Chapter 8 builds on this groundwork in drawing together some controversial issues in the foundations of computational psychology, and in asking whether it is possible at all.

For the sake of the sceptical reader who suspects at the outset that this dragon-hunt is doomed to failure–that computational psychology is not possible at all–it may be useful to make a Popperian point. Science advances not only by conjectures but also by refutations [Popper, 1963]. Even if we decided that a computational methodology is not appropriate to psychological modelling after all, we still should have learnt something. We should have some reasonably precise ideas about just what is faulty or inadequate in current models, and we might even have arrived at some notion of what sort of theoretical power is lacking.

However, this is an unnecessarily negative way of justifying the computational approach. We shall see that, despite all the difficulties and unsettled controversies, some distinct gains have been made for psychology by attempts to consider the mind as a computational machine.

2 Patterns, polyhedra, imagery

Work on vision includes some of the most fruitful psychological research inspired by computational ideas. It also exemplifies many of the general methodological controversies mentioned in Chapter 1. The discussion of computer vision (in this chapter and the next) will introduce topics such as the psychological relevance of neurophysiology, the need for new kinds of computation, the theoretical importance of programs as such, the role of orthodox AI, and the sense in which a psychological theory should be "computational".

The philosophical background

The key problem for the psychology of vision is how our visual system enables us to gain reliable information about the environment. Implicit in this question are two others: How far does vision depend on high-level inferences as opposed to autonomous low-level processes? And do the low-level processes themselves actively construct a visual representation, or do they merely pick up information presented to the eyes? Insofar as one regards vision as the construction (at whatever level) of representations, one faces also a fourth question: What is the nature of visual representations? This, in turn, prompts enquiry into the nature of visual imagery.

Two theoretical poles from which to approach these questions were established by the seventeenth-century philosophers, and have informed experimental psychology since its inception a hundred years ago.

Empiricists (like Locke) stress low-level, automatic, processes rather than high-level judgments or control, and see these processes as passively reflecting the input information rather than transforming or interpreting it. Moreover, they assume a *tabula rasa* in the newborn organism, any interpretative activity being not only high-level but learnt. Images are merely less vivid copies of sense-impressions, which can be imaginatively combined, as when one pictures a unicorn.

Rationalist accounts, by contrast, stress the active construction of mental representations, and the contribution to vision of (unconscious) conceptual judgments. Descartes, for instance, said that though we think we see people walking down the street, all we really see is coats and hats moving. Although the connection between coats and people has to be learnt, rationalists assume that many structuring principles are innate. Imagination is