THE COMPUTER REVOLUTION IN PHILOSOPHY:

Philosophy Science and Models of Mind

AARON SLOMAN

Reader in Philosophy and Artificial Intelligence,
Cognitive Studies Programme, The University of Sussex

THE HARVESTER PRESS

Since 1991 the author has been Professor of Artificial Intelligence and Cognitive Science in the School of Computer Science at the University of Birmingham, UK.

(Slightly edited versions of original title pages.)

THE COMPUTER REVOLUTION IN PHILOSOPHY:

Philosophy Science and Models of Mind

HARVESTER STUDIES IN COGNITIVE SCIENCE
General Editor: Margaret A. Boden

Harvester Studies in Cognitive Science is a new series which will explore the nature of knowledge by way of a distinctive theoretical approach one that takes account of the complex structures and interacting processes that make thought and action possible.

Intelligence can be studied from the point of view of psychology, philosophy, linguistics, pedagogy and artificial intelligence, and all these different emphases will be represented within the series.

Other titles in this series:

ARTIFICIAL INTELLIGENCE AND NATURAL MAN:
Margaret A. Boden

INFERENTIAL SEMANTICS:
Frederick Parker-Rhodes

Other titles in preparation:
THE FORMAL MECHANICS OF MIND:
Stephen N. Thomas

THE COGNITIVE PARADIGM:
Marc De Mey

ANALOGICAL THINKING: MYTHS ANDMECHANISMS:
Robin Anderson

EDUCATION AND ARTIFICIAL INTELLIGENCE:
Tim O'Shea

First published in Great Britain in 1978 by
THE HARVESTER PRESS LIMITED
Publisher: John Spiers
2 Stanford Terrace, Hassocks, Sussex

(Also published in the USA by Humanities Press, 1978)

British Library Cataloguing in Publication Data

Sloman, Aaron

The computer revolution in philosophy.
(Harvester studies in cognitive science).
1. Intellect. 2. Artificial intelligence
1. Title
128'.2 BF431
ISBN 0-85527-389-5
ISBN 0-85527-542-1 Pbk.

Printed in England by Redwood Burn Limited,

Trowbridge & Esher

CONTENTS
(WARNING: Page numbers may not be accurate)

PART TWO *Mechanisms*

Footnotes are at the end of each chapter.

Online Contents Page
Next: Preface.
Updated: 31 Jan 2007

Index page from web site

THE COMPUTER REVOLUTION IN PHILOSOPHY:

Philosophy, science and models of mind.

http://www.cs.bham.ac.uk/research/projects/cogaff/crp/

Aaron Sloman
School of Computer Science
The University of Birmingham.

This book, published in 1978 by Harvester Press and Humanities Press, has been out of print for many years, and is now online. This online version was produced from a scanned in copy of the original, digitised by OCR software and made available in September 2001. Since then a number of notes and corrections have been added.

The whole book can be downloaded in a single PDF file (about 1.7MB) from http://www.cs.bham.ac.uk/research/projects/cogaff/crp/crp.pdf It is also available in Eprints repository of ASSC (The Association for the Scientific Study of Consciousness.) The separate chapters are also available, as follows.

ONLINE CONTENTS

Remaining contents of this section (from online version)

Some Reviews and Other Comments on this Book

Comments on the historical significance (or non-significance!) of this book can be found in the introduction to Luciano Floridi's textbook "Philosophy of information" referenced on Blackwell's site.

Several of the reviews published in response to the original book are now available online, e.g. Donald Mackay's review in the British Journal for the Philosophy of Science Vol 30 No 3 (1979), which castigated me for not reviewing previous relevant work by Craik, Wiener and McCulloch. An excellent survey of their work and others is now available in Margaret Boden's two volume *Mind as Machine: A History of Cognitive Science* published by Oxford University Press 29th June 2006.

An excellent survey of their work and others is now available in Margaret Boden's two volume *Mind as Machine: A History of Cognitive Science* published by Oxford University Press 29th June 2006

(see also http://www.cs.bham.ac.uk/research/projects/cogaff/misc/boden-mindasmachine.html)

Perhaps the earliest published reference to this book is

Shallice, T., & Evans, M. E. (1978). The involvement of the frontal lobes in cognitive estimation. Cortex, 14, 294-303, available at:

http://www-personal.umich.edu/~evansem/shallice-evans.doc

Philosophical relevance

Some parts of the book are dated whereas others are still relevant both to the scientific study of mind and to philosophical questions about the aims of science, the nature of theories and explanations, varieties of concept formation, and to questions about the nature of mind.

In particular, Chapter 2 analyses the variety of scientific advances ranging from shallow discoveries of new laws and correlations to deep science which extends our ontology, i.e. our understanding of what is possible, rather than just our understanding of what happens when.

Insofar as AI explores designs for possible mental mechanisms, possible mental architectures, and possible minds using those mechanisms and architectures, it is primarily a contribution to deep science, in contrast with most empirical psychology which is shallow science, exploring correlations.

This "design stance" approach to the study of mind was very different from the "intentional stance" being developed by Dan Dennett at the same time, expounded in his 1978 book *Brainstorms*, and later partly re-invented by Alan Newell as the study of "The knowledge Level" (see his 1990 book *Unified Theories of Cognition*). Both Dennett and Newell based their methodologies on a presumption of rationality, whereas the design-stance considers functionality, which is possible without rationality, as insects and microbes demonstrate well, Functional mechanisms may provide limited rationality, as Herb Simon noted in his 1969 book *The Sciences of the Artificial*.

Relevance to AI and Cognitive Science

In some ways the AI portions of the book are not as out of date as the publication date might suggest because it recommends approaches that have not yet been explored fully (e.g. the study of human-like mental architectures in Chapter 6); and some of the alternatives that have been explored have not made huge amounts of progress (e.g. there has been much vision research in directions that are different from those recommended in Chapter 9).

I believe that ideas about "Representational Redescription" presented in Anette Karmiloff-Smith's book *Beyond Modularity* summarised in her BBS 2004 article with pre-print here are illustrated by my discussion of some of what goes on when a child learns about numbers in Chapter 8. That chapter suggests mechanisms and processes involved in learning about numbers that could be important for developmental psychology, philosophy and AI, but have never been properly developed.

Some chapters have short notes commenting on developments since the time the book was published. I may add more such notes from time to time.

More recent work by the author

A draft sequel to this book was partly written around 1985, but never published because I was dissatisfied with many of the ideas, especially because I did not think the notion of "computation" was well defined. More recent work developing themes from the book is available in the

Cognition and Affect Project directory

and also in the slides for recent conference and seminar presentations here:

http://www.cs.bham.ac.uk/research/cogaff/talks/

and in the papers, discussion notes and presentations related to the CoSy robotic project here:

http://www.cs.bham.ac.uk/research/projects/cosy/papers/

A particularly relevant discussion note is my answer to the question 'what is information?' -- in the context of the notion of an information-processing system (not Shannon's notion of information):

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/whats-information.html

A more complete list of things I have done, many of which which grew out of the ideas in this book, can be found in http://www.cs.bham.ac.uk/~axs/my-doings.html

Information about the online version

The book has been scanned and converted to HTML. This was completed on 29 Sep 2001. I am very grateful to Manuela Viezzer for photocopying the book and to Sammy Snow for giving up so much time to scanning it in. Thanks also to Chris Glur for reporting bits of the text that still needed cleaning up after scanning and conversion to html.

The OCR package used had a hard task and very many errors had to be corrected in the digitised version. It is likely that many still remain. Please report any to me at A.Sloman@cs.bham.ac.uk.

It proved necessary to redo all the figures, for which I used the TGIF package, freely available for Linux and Unix systems from these sites:

http://bourbon.cs.umd.edu:8001/tgif/
ftp://ftp.cs.ucla.edu/pub/tgif/

The HTML version has several minor corrections and additions, and a number of recently added notes and comments.

NOTE About PDF versions

PDF versions were produced by reading the html files into odt format in OpenOffice, then making minor formatting changes and exporting to PDF. OpenOffice is freely available for a variety of platforms from http://www.openoffice.org

Download everything at once

In HTML and PDF format

The files may be accessed online via the table of contents, in HTML and PDF or the whole book fetched as one PDF file (about 1.7MBytes).

Alternatively, the complete set of HTML and PDF chapters can be downloaded for local use packaged in a zip file:

http://www.cs.bham.ac.uk/research/cogaff/crp.zip

or a gzipped tar file:

http://www.cs.bham.ac.uk/research/cogaff/crp.tar.gz

NOTE on educational predictions

The world has changed a lot since the book was published, but not enough, in one important respect.

In the Preface and in Chapter 1 comments were made about how the invention of computing was analogous to the combination of the invention of writing and of the printing press, and predictions were made about the power of computing to transform our educational system to stretch minds.

Alas the predictions have not yet come true: instead computers are used in schools for lots of shallow activities. Instead of teaching cooking, as used to happen in 'domestic science' courses we teaching them 'information cooking' using word processors, browsers, an the like. We don't teach them to design, debug, test, analyse, explain new machines and tools, merely to use existing ones as black boxes. That's like teaching cooking instead of teaching chemistry.

In 2004 a paper on that topic, accepted for a UK conference on grand challenges in computing education referred back to the predictions in the book and how the opportunities still remain. The paper, entitled 'Education Grand Challenge: A New Kind of Liberal Education --- Making People Want a Computing Education For Its Own Sake' is available in HTML and PDF formats here
http://www.cs.bham.ac.uk/research/cogaff/misc/gced.html
Additional comments were made in 2006 in this document Why Computing Education has Failed and How to Fix it

---

Last updated: 1 March 2007

PREFACE

Another book on how computers are going to change our lives? Yes, but this is more about computing than about computers, and it is more about how our thoughts may be changed than about how housework and factory chores will be taken over by a new breed of slaves.

Thoughts can be changed in many ways. The invention of painting and drawing permitted new thoughts in the processes of creating and interpreting pictures. The invention of speaking and writing also permitted profound extensions of our abilities to think and communicate. Computing is a bit like the invention of paper (a new medium of expression) and the invention of writing (new symbolisms to be embedded in the medium) combined. But the writing is more important than the paper. And computing is more important than computers: programming languages, computational theories and concepts -- these are what computing is about, not transistors, logic gates or flashing lights. Computers are pieces of machinery which permit the development of computing as pencil and paper permit the development of writing. In both cases the physical form of the medium used is not very important, provided that it can perform the required functions.

Computing can change our ways of thinking about many things, mathematics, biology, engineering, administrative procedures, and many more. But my main concern is that it can change our thinking about ourselves: giving us new models, metaphors, and other thinking tools to aid our efforts to fathom the mysteries of the human mind and heart. The new discipline of Artificial Intelligence is the branch of computing most directly concerned with this revolution. By giving us new, deeper, insights into some of our inner processes, it changes our thinking about ourselves. It therefore changes some of our inner processes, and so changes what we are, like all social, technological and intellectual revolutions.

I cannot predict all these changes, and certainly shall not try. The book is mainly about philosophical thinking, and its transformation in the light of computing. But one of my themes is that philosophy is not as limited an activity as you might think. The boundaries between philosophy and other theoretical and practical activities, notably education, software engineering, therapy and the scientific study of man, cannot be drawn as neatly as academic syllabuses might suggest. This blurring of disciplinary boundaries helps to substantiate a claim that a revolution in philosophy is intimately bound up with a revolution in the scientific study of man and its practical applications. Methodological excursions into the nature of science and philosophy therefore take up rather more of this book than I would have liked. But the issues are generally misunderstood, and I felt something needed to be done about that.

I think the revolution is also relevant to several branches of science and engineering not directly concerned with the study of man. Biology, for example, seems to be ripe for a computational revolution. And I don't mean that biologists should use computers to juggle numbers -- number crunching is not what this book is about. Nor is it what computing is essentially about. Further, it may be useful to try to understand the relationship between chemistry and physics by thinking of physical structures as providing a computer on which chemical programs are executed. But I am not so sure about that one, and will not pursue it.

Though fascinated by the intellectual problems discussed in the book, I would find it hard to justify spending public money working on them if it were not for the possibility of important consequences, including applications to education. But perhaps I should not worry: so much public money is wasted

on futile research and teaching, to say nothing of incompetent public administration, ridiculous defence preparations, profits for manufacturers and purveyors of shoddy, useless or harmful goods (like cigarettes), that a little innocent academic study is marginal.

Early drafts of this book included lots of nasty comments on the current state of philosophy, psychology, social science, and education. I have tried to remove them or tone them down, since many were based on my ignorance and prejudice. In particular, my knowledge of psychology at the time of writing was dominated by lectures, seminars, textbooks and journal articles from the 1960s. Nowadays many psychologists are as critical as I could be of such psychology (which does not mean they will agree with my criticisms and proposed remedies). And Andreski's *Social Science as Sorcery* makes many of my criticisms of social science redundant.

I expect I shall be treading on many toes in my bridge-building comments. The fact that I have not read everything relevant will no doubt lead me into howlers. Well, that's life. Criticisms and corrections, published or private will be welcomed. (Except for arguments about whether I am doing philosophy or psychology or some kind of engineering. Demarcation disputes are usually a waste of time. Instead ask: are the problems interesting or important, and is some real progress made towards dealing with them?)

Since the book is aimed at a wide variety of readers with different backgrounds, it will be found by each of them to vary in clarity and interest from section to section. One person's banal oversimplification is another's mind-stretching novelty. Partly for this reason, the different chapters vary in style and overlap in content. The importance of the topic, and the shortage of informed discussion seemed to justify offering the book for publication despite its many flaws.

One thing that will infuriate some readers is my refusal to pay close attention to published arguments in the literature about whether machines can think, or whether people are machines of some sort. People who argue about this sort of thing are usually ignorant of developments in artificial intelligence, and their grasp of the real problems and possibilities in designing intelligent machines is therefore inadequate. Alternatively, they know about machines, but are ignorant of many old philosophical problems for mechanist theories of mind.

Most of the discussions (on both sides) contain more prejudice and rhetoric than analysis or argument. I think this is because in the end there is not much scope for rational discussion on this issue. It is ultimately an ethical question whether you should treat robots like people, or at least like cats, dogs or chimpanzees; not a question of fact. And that ethical question is the real meat behind the question whether artefacts could ever think or feel, at any rate when the question is discussed without any attempt to actually *design* a thinking or feeling machine.

When intelligent robots are made (with the help of philosophers), in a few hundred or a few thousand years time, some people will respond by accepting them as communicants and friends, whereas others will use all the old racialist arguments for depriving them of the status of persons. Did you know that you were a racialist?

But perhaps when it comes to living and working with robots, some people will be surprised how hard it is to retain the old disbelief in their consciousness, just as people have been surprised to find that someone of a different colour may actually be good to relate to as a person. For an unusually informative and well-informed statement of the racialist position concerning machines see Weizenbaum 1976. I admire his book, despite profound disagreements with it.

So, this book is an attempt to publicise an important, but largely unnoticed, facet of the computer revolution: its potential for transforming our ways of thinking about ourselves. Perhaps it will lead someone else, knowledgeable about developments in computing and Artificial Intelligence, to do a better job, and substantiate my claim that within a few years philosophers, psychologists,

educationalists, psychiatrists, and others will be professionally incompetent if they are not well-informed about these developments.

ACKNOWLEDGEMENTS

I have not always attributed ideas or arguments derived from others. I tend to remember content, not sources. Equally I'll not mind if others use my ideas without acknowledgement. The property-ethic dominates too much academic writing. It will be obvious to some readers that besides recent work in artificial intelligence the central ideas of Kant's *Critique of Pure Reason* have had an enormous influence on this book. Writings of Frege, Wittgenstein, Ryle, Austin, Popper, Chomsky, and indirectly Piaget have also played an important role. Many colleagues and students have helped me in a variety of ways: by provoking me to disagreement, by discussing issues with me, or by reading and commenting on earlier drafts of one or more chapters. This has been going on for a long time, so I am not sure that the following list includes everyone who has refined or revised my ideas, or given me new ones:

Frank Birch, Margaret Boden, Mike Brady, Alan Bundy, Max Clowes, Steve Draper, Gerald Gazdar, Roger Goodwin, Steven Hardy, Pat Hayes, Geoffrey Hinton, Laurie Hollings, Nechama Inbar, Robert Kowalski, John Krige, Tony Leggett, Barbara Lloyd, Christopher Longuet-Higgins, Alan Mackworth, Frank O'Gorman, David Owen, Richard Power, Julie Rutkowska, Alison Sloman, Jim Stansfield, Robin Stanton, Sylvia Weir, Alan White, Peter Williams.

Pru Heron, Jane Blackett, Judith Dennison, Maryanne McGinn and Pat Norton helped with typing and editing. Jane Blackett also helped with the diagrams.

The U.K. Science Research Council helped, first of all by enabling me to visit the Department of Artificial Intelligence in Edinburgh University for a year in 19723, and secondly by providing me with equipment and research staff for a three year project on computer vision at Sussex.

Bernard Meltzer was a very helpful host for my visit to Edinburgh, and several members of the department kindly spent hours helping me learn programming, and discussing computing concepts, especially Bob Boyer, J. Moore, Julian Davies and Danny Bobrow. Steve Hardy and Frank O'Gorman continued my computing education when I returned from Edinburgh. Several of my main themes concerning the status of mind can be traced back to interactions with Stuart Sutherland (e.g. see his 1970) and Margaret Boden. Her book *Artificial Intelligence and Natural Man,* like other things she has written, adopts a standpoint very similar to mine, and we have been talking about these issues over many years. So I have probably cribbed more from her than I know.

She also helped by encouraging me to put together various privately circulated papers when I had despaired of being able to produce a coherent, readable book. By writing her book she removed the need for me to give a detailed survey of current work in the field of A.I. Instead I urge readers to study her survey to get a good overview.

I owe my conversion to Artificial Intelligence, towards the end of 1969, to Max Clowes. I learnt a great deal by attending his lectures for undergraduates. He first pointed out to me that things I was trying to do in philosophical papers I was writing were being done better in A.I., and urged me to take up programming. I resisted for some time, arguing that I should first finish various draft papers and a book. Fortunately, I eventually realised that the best plan was to scrapthem.

(I have not been so successful at convincing others that their intellectual investments are not as valuable as the new ideas and techniques waiting to be learnt. I suspect, in some cases, this is partly because they were allowed by the British educational system to abandon scientific and mathematical

subjects and rigorous thinking at a fairly early age to specialise in arts and humanities subjects. I believe that the knowledge-explosion, and the needs of our complex modern societies, make it essential that we completely re-think the structure of formal education, from primary schools upwards: indefinitely continued teaching and learning at all ages in sciences, arts, humanities, crafts (including programming) must be encouraged. Perhaps that will be the best way to cope with unemployment produced byautomation, and the like. But I'm digressing!).

Alison, Benjamin and Jonathan tolerated (most of the time) my withdrawal from family life for the sake of this book and other work. I did not wish to have children, but as will appear frequently in this book (e.g., in the chapter on learning about numbers), observing them and interacting with them has taught me a great deal. In return, my excursions into artificial intelligence and the topics of the book have changed my way of relating to children. I think I now understand their problems better, and have acquired a deeper respect for their intellectual powers.

The University of Sussex provided a fertile environment for the development of the ideas reported here, by permitting a small group of almost fanatical enthusiasts to set up a 'Cognitive Studies Programme' for interdisciplinary teaching and research, and providing us with an excellent though miniscule computing laboratory. But for the willingness of the computer to sit up with me into the early hours helping me edit, format, and print out draft chapters (and keeping me warm when the heating was off), the book would not have been ready for a long time to come.

I hope that, one day, even better computing facilities will be commonplace in primary schools, for kids to play with. After all, primary schools are more important than universities, aren't they?

NOTE ADDED APRIL 2001

I am grateful to Manuela Viezzer, a PhD student at the University of Birmingham, for offering to photocopy the pages of this book, and to Sammy Snow, a member of clerical staff, for scanning them in her spare time.

Book contents page
Next: Chapter One

Last updated: 14 Jan 2002

CHAPTER 1

INTRODUCTION AND OVERVIEW

1.1. Computers as toys to stretch our minds

Developments in science andtechnology are responsible for some of the best and some of the worst features of our lives. The computer is no exception. There are plenty of reasons for being pessimistic about its effects in the short run, in a society where the lust for power, profit, status and material possessions are dominant motives, and where those with knowledge -- for instance scientists, doctors and programmers -- can so easily manipulate and mislead those without.

Nevertheless I am convinced that the ill effects of computers can eventually be outweighed by their benefits. I am not thinking of the obvious benefits, like liberation from drudgery and the development of new kinds of information services. Rather, I have in mind the role of the computer, and the processes which run on it, as a new medium of self-expression, perhaps comparable in importance to the invention of writing.

Think of it like this. From early childhood onwards we all need to play with toys, be they bricks, dolls, construction kits, paint and brushes, words, nursery rhymes, stories, pencil and paper, mathematical problems, crossword puzzles, games like chess, musical instruments, theatres, scientific laboratories, scientific theories, or other people. We need to interact with all these playthings and playmates in order to develop our understanding of ourselves and our environment that is, in order to develop our concepts, our thinking strategies, our means of expression and even our tastes, desires and aims in life. The fruitfulness of such play depends in part on how complex the toy and the processes it generates, and how rich the interaction between player and toy are.

A modern digital computer is perhaps the most complex toy ever created by man. It can also be as richly interactive as a musical instrument. And it is certainly the most flexible: the very same computer may simultaneously be helping an eight year old child to generate pictures on a screen and helping a professional programmer to understand the unexpected behaviour of a very complex program he has designed. Meanwhile other users may be attempting to create electronic music, designing a program to translate English into French, testing a program which analyses and describes pictures, or simply treating the computer as an interactive diary. A few old-fashioned scientists may even be doing some numerical computations.

Unlike pet animals and other people (also rich, flexible and interactive), computers are toys designed by people. So people can understand how they work. Moreover the designs of the programs which run on them can be and are being extended by people, and this can go on indefinitely. As we extend these designs, our ability to think and talk about complex structures and processes is extended. We develop new concepts, new languages, new ways of thinking. So we acquire powerful new tools with which to try to understand other complex systems which we have not designed, including systems which have so far largely resisted our attempts at comprehension: for instance human minds and social systems. Despite the existence of university departments of psychology, sociology, education, politics, anthropology, economics and international relations, it is clear that understanding of these domains is currently at a pathetically inadequate level: current theories don't yet provide a basis for designing satisfactory educational procedures, psychological therapies, or government policies.

But apart from the professionals, ordinary people need concepts, symbolisms, metaphors and models to help them understand the world, and in particular to help them understand themselves and other

people. At present much of our informal thinking about people uses unsatisfactory mechanistic models and metaphors, which we are often not even aware of using. For instance even people who strongly oppose the application of computing metaphors to mental processes, on the grounds that computers are mere mechanisms, often unthinkingly use much cruder mechanistic metaphors, for instance 'He needed to let off steam', I was pulled in two directions at once, but the desire to help my family was stronger', 'His thinking is stuck in a rut', 'The atmosphere in the room was highly charged'. Opponents of the spread of computational metaphors are in effect unwittingly condemning people to go on living with hydraulic, clock-work, and electrical metaphors derived from previous advances in science and technology.

To summarise so far: it can be argued that computers, or, to be more precise, combinations of computers and programs, constitute profoundly important new toys which can give us new means of expression and communication and help us create an ever-increasing new stock of concepts and metaphors for thinking about all sorts of complex systems, including ourselves.

I believe that not only psychology and social sciences but also biology and even chemistry and physics can be transformed by attempting to view complex processes as computational processes, including rich information flow between sub-processes and the construction and manipulating of symbolic structures within processes. This should supersede older paradigms, such as the paradigm which represents processes in termsof equations or correlations between numerical variables.

This paradigm worked well for a while in physics but now seems to dominate, and perhaps to strangle, other disciplines for which it is irrelevant. Apart from computing science, linguistics and logic seem to be the only sciences which have sharply and successfully broken away from the paradigm of 'variables, equations and correlations'. But perhaps it is significant that the last two pretend not to be concerned with processes, only with structures. This is a serious limitation, as I shall try to show in later chapters.

1.2. The Revolution in Philosophy

Well, suppose it is true that developments in computing can lead to major advances in the scientific study of man and society: what have these scientific advances to do with philosophy?

The very question presupposes a view of philosophy as something separate from science, a view which I shall attempt to challenge and undermine later, since it is based both on a misconception of the aims and methods of science and on the arrogant assumption by many philosophers that they are the privileged guardians of a method of discovering important non-empirical truths.

But there is a more direct answer to the question, which is that very many of the problems and concepts discussed by philosophers over the centuries have been concerned with *processes,* whereas philosophers, like everybody else, have been crippled in their thinking about processes by too limited a collection of concepts and formalisms. Here are some age-old philosophical problems explicitly or implicitly concerned with processes. How can sensory experience provide a rational basis for beliefs about physical objects? How can concepts be acquired through experience, and what other methods of concept formation are there? Are there rational procedures for generating theories or hypotheses? What is the relation between mind and body? How can non-empirical knowledge, such as logical or mathematical knowledge, be acquired? How can the utterance of a sentence relate to the world in such a way as to say something true or false? How can a one-dimensional string of words be understood as describing a three-dimensional or multi-dimensional portion of the world? What forms of rational inference are there? How can motives generate decisions, intentions and actions? How do non-verbal representations work? Are there rational procedures for resolving social conflicts?

There are many more problems in all branches of philosophy concerned with processes, such as perceiving, inferring, remembering, recognising, understanding, learning, proving, explaining,

communicating, referring, describing, interpreting, imagining, creating, deliberating, choosing, acting, testing, verifying, and so on. Philosophers, like most scientists, have an inadequate set of tools for theorising about such matters, being restricted to something like common sense plus the concepts of logic and physics. A few have clutched at more recent technical developments, such as concepts from control theory (e.g. feedback) and the mathematical theory of games (e.g. payoff matrix), but these are hopelessly deficient for the tasks of philosophy, just as they are for the task of psychology.

The new discipline of artificial intelligence explores ways of enabling computers to do things which previously could be done only by people and the higher mammals (like seeing things, solving problems, making and testing plans, forming hypotheses, proving theorems, and understanding English). It is rapidly extending our ability to think about processes of the kinds which are of interest to philosophy. So it is important for philosophers to investigate whether these new ideas can be used to clarify and perhaps helpfully reformulate old philosophical problems, re-evaluate old philosophical theories, and, above all, to construct important new answers to old questions. As in any healthy discipline, this is bound to generate a host of new problems, and maybe some of them can be solved too.

I am prepared to go so far as to say that within a few years, if there remain any philosophers who are not familiar with some of the main developments in artificial intelligence, it will be fair to accuse them of professional incompetence, and that to teach courses in philosophy of mind, epistemology, aesthetics, philosophy of science, philosophy of language, ethics, metaphysics, and other main areas of philosophy, without discussing the relevant aspects of artificial intelligence will be as irresponsible as giving a degree course in physics which includes no quantum theory. Later in this book I shall elucidate some of the connections. Chapter 4, for example, will show how concepts and techniques of philosophy are relevant to AI and cognitive science.

Philosophy can make progress, despite appearances. Perhaps in future the major advances will be made by people who do not call themselves philosophers.

After that build-up you might expect a report on some of the major achievements in artificial intelligence to follow. But that is not the purpose of this book: an excellent survey can be found in Margaret Boden's book. *Artificial Intelligence and Natural Man,* and other works mentioned in the bibliography will take the interested reader into the depths of particular problem areas. (Textbooks on AI will be especially useful for readers wishing to get involved in *doing* artificial intelligence.)

My main aim in this book is to re-interpret some age-old philosophical problems, in the light of developments in computing. These developments are also relevant to current issues in psychology and education. Most of the topics are closely related to frontier research in artificial intelligence, including my own research into giving a computer visual experiences, and analysing motivational and emotional processes in computational terms.

Some of the philosophical topics in Part One of the book are included not only because I think I have learnt important things by relating them to computational ideas, but also because I think misconceptions about them are among the obstacles preventing philosophers from accepting the relevance of computing. Similar misconceptions may confuse workers in AI and cognitive science about the nature of their discipline.

For instance, the chapters on the aims of science and the relations between science and philosophy attempt to undermine the wide-spread assumption that philosophers are doing something so different from scientists that they need not bother with scientific developments and *vice versa*. Those chapters are also based on the idea that developments in science and philosophy form a computational process not unlike the one we call human learning.

The remaining chapters, in Part Two, contain attempts to use computational ideas in discussing some

problems in metaphysics, philosophy of mind, epistemology, philosophy of language and philosophy of mathematics. I believe that further analysis of the nature of number concepts and arithmetical knowledge in terms of symbol-manipulating processes could lead to profound developments in primary school teaching, as well as solving old problems in philosophy of mathematics.

In the remainder of this chapter I shall attempt to present, in bold outline, some of the main themes of the computer revolution, followed by a brief definition of ``Artificial Intelligence''. This will help to set the stage for what follows. Some of the themes will be developed in detail in later chapters. Others will simply have to be taken for granted as far as this book is concerned. Margaret Boden's book and more recent textbooks on AI fill most of the gaps.

1.3. Themes from the Computer Revolution

1. Computers are commonly viewed as elaborate numerical calculators or at best as devices for blindly storing and retrieving information or blindly following sequences of instructions programmed into them. However, they can be more accurately viewed as an extension of human means of expression and communication, comparable in importance to the invention of writing. Programs running on a computer provide us with a medium for thinking new thoughts, trying them out, and gradually extending, deepening and clarifying them. This is because, when suitably programmed, computers are devices for constructing, manipulating, analysing, interpreting and transforming symbolic structures of all kinds, including their own programs.

2. Concepts of 'cause', law', and 'mechanism', discussed by philosophers, and used by scientists, are seriously impoverished by comparison with the newly emerging concepts.

The old concepts suffice for relatively simple physical mechanisms, like clocks, typewriters, steam engines and unprogrammed computers, whose limitations can be illustrated by their inability to support a notion of *purpose*.

By contrast, a programmed computer may include representations of itself, its actions, possible futures, reasons for choosing, and methods of inference, and can therefore sometimes contain purposes which generate behaviour, as opposed to merely containing physical structures and processes which generate behaviour. So biologists and psychologists who aim to banish talk of purposes from science, thereby ignore some of the most important new developments in science. So do philosophers and psychologists who use the existence of purposive human behaviour to 'disprove' the possibility of a scientific study of man.

3. Learning that a computer contains a certain sub-program enables you to explain some of the things it can do, but provides no basis for predicting what it *always* or *frequently* does, since that will depend on a large number of other factors which determine when this sub-program is executed and the environment in which it is executed. So a scientific investigation of computational processes need not be primarily a search for *laws* so much as an attempt to describe and explain what sorts of things are and are not *possible*. A central form of question in science and philosophy is 'How is so and so possible?' Many scientists, especially those studying people and social systems, mislead themselves and their students into thinking that science is essentially a search for laws and correlations, so that they overlook the study of possibilities. Linguists (especially since Chomsky) have grasped this point, however. (This topic is developed at length in chapter 2.)

4. Similarly there is a wide-spread myth that the scientific study of complex systems requires the use of numerical measurements, equations, calculus, and the other mathematical paraphernalia of physics. These things are useless for describing or explaining the important aspects of the behaviour of complex programs (e.g. a computer, operating system, or Winograd's program described in his book *Understanding Natural Language*).

Instead of equations and the like, quite new non-numerical formalisms have evolved in the form of programming languages, along with a host of informal concepts relating the languages, the programs expressed therein, and the processes they generate. Many of these concepts (e.g. *parsing, compiling, interpreting, pointer, mutual recursion, side-effect, pattern matching)* are very general, and it is quite likely that they could be of much more use to students of biology, psychology and social science than the kinds of numerical mathematics they are normally taught, which are of limited use for theorising about complex interacting structures. Unfortunately although many scientists dimly grasp this point (e.g. when they compare the DNA molecule with a computer program) they are often unable to *use* the relationship: their conception of a computer program is limited to the sorts of data-processing programs written in low-level languages like Fortran or Basic.

5. It is important to distinguish cybernetics and so-called 'systems theory' from this broader science of computation, for the former are mostly concerned with processes involving relatively fixed structures in which something quantifiable (e.g. money, energy, electric current, the total population of a species) flows between or characterises substructures. Their formalisms and theories are too simple to say anything precise about the communication of a sentence, plan or problem, or to represent the process of construction or modification of a symbolic structure which stores information or abilities.

Similarly, the mathematical theory of information, of Shannon and Weaver, is mostly irrelevant, although computer programs are often said to be information-processing mechanisms. The use of the word 'information' in the mathematical theory has proved to be utterly misleading. It is not concerned with meaning or content or sense or connotation or denotation, but with probability and redundancy in signals. If more suitable terminology had been chosen, then perhaps a horde of artists, composers, linguists, anthropologists, and even philosophers would not have been misled.

I am not denying the importance of the theory to electronic engineering and physics. In some contexts it is useful to think of communication as sending a signal down a noisy line, and understanding as involving some process of decoding signals. But human communication is quite different: we do not decode, we interpret, using enormous amounts of background knowledge and problem-solving abilities. That is, we map one class of structures (e.g. 2-D images), into another class (e.g. 3-D scenes). Chapter 9 elaborates on this, in describing work in computer vision. The same is true of artificial intelligence programs which understand language. Information theory is not concerned with such mappings.

6. One of the major new insights is that computational processes may be markedly decoupled from the physical processes of the underlying computer. Computers with quite different basic components and architecture may be equivalent in an important sense: a program which runs on one of them can be made to run on any other either by means of a second program which simulates the first computer on the second, or by means of a suitable compiler or interpreter program which *translates* the first program into a formalism which the second computer can execute. So a program may run on a *virtual* machine.

Differences in size can be got round by attaching peripheral storage devices such as magnetic discs or tapes, leaving only differences in speed.

So all modern digital computers are theoretically equivalent, and the detailed physical structure and properties of a computer need not constrain or determine the symbol-manipulating and problem-solving processes which can run on it: any constraints, except for speed, can be overcome by providing more storage and feeding in new programs. Similarly, the programs do not determine the computers on which they can run.

7. Thus reductionism is refuted. For instance, if biological processes are computational processes running on a physico-chemical computer, then essentially the same processes could, with suitable re-

programming, run on a different sort of computer. Equally, the same computer could permit quite different computations: so the nature of the physical world need not determine biological processes. Just as the electronic engineers who build and maintain a computer may be quite unable to describe or understand some of the programs which run on it, so may physicists and chemists lack the resources to describe, explain or predict biological processes. Similarly psychology need not be reducible to physiology, nor social processes to psychological ones. To say that wholes may be more than the sum of their parts, and that qualitatively new processes may 'emerge' from old ones, now becomes an acceptable part of the science of computation, rather than old-fashioned mysticism. Many anti-reductionists have had this thought prior to the development of computing, but have been unable to give it a clear and indisputable foundation.

8. There need not be only *two* layers: programs and physical machine. A suitably programmed computer (e.g. a computer with a compiler program in it[2]), is itself a new computer a new 'virtual machine' which in turn may be programmed so as to support new kinds of processes. Thus a single process may involve many layers of computations, each using the next lower layer as its underlying machine. But that is not all. The relations may sometimes not even be hierarchically organised, for instance if process A forms part of the underlying machine for process B and process B forms part of the underlying machine for process A. Social and psychological, psychological and physiological processes, seem to be related in this mutually supportive way. Chapters 6 and 9 present some examples. The development of good tools for thinking about a system composed of multiple interlocking processes is only just beginning. Systems of differential equations and the other tools of mathematical physics are worse than useless, for the attempt to use them can yield quite distorted descriptions of processes invdving intelligent systems, and encourage us to ask unfruitful questions.

9. Philosophers sometimes claim that it is the business of philosophy only to analyse concepts, not to criticise them. But constructive criticism is often needed and in many cases the task will not be performed if philosophers shirk it. An important new task for philosophers is constructively critical analysis of the concepts and underlying presuppositions emerging from computer science and especially artificial intelligence. Further, by carefully analysing the mismatch between some of our very complicated ordinary concepts like *goal, decide, infer, perceive, emotion, believe, understand,* and the models being developed in artificial intelligence, philosophers may help to counteract unproductive exaggerated claims and pave the way for further developments. They will be rewarded by being helped with some of their philosophical problems.

10. For example, the computational metaphor, paradoxically, provides support for a claim that human decisions are not physically or physiologically determined, since, as explained above, if the mind is a computational process using the brain as a computer then it follows that the brain does not constrain the range of mental processes, any more than a computer constrains the set of algorithms that can run on it. It can be more illuminating to think of the program (or mind) as constraining the physical processes than vice versa.

Moreover, since the state of a computation can be frozen, and stored in some non-material medium such as a radio signal transmitted to a distant planet, and then restarted on a different computer, we see that the hitherto non-scientific hypothesis that people can survive bodily death, and be resurrected later on, acquires a new lease of life. Not that this version is likely to please theologians, since it no longer requires a god.

11. Recent attempts to give computers perceptual abilities seem to have settled the empiricist/rationalist debate by supporting Immanuel Kant's claim that no experiencing is possible without information-processing (analysis, comparison, interpretation of data) and that no information-processing is possible without pre-existing knowledge in the form of symbol-manipulating procedures, data-structures, and quite specific descriptive abilities. (This topic is elaborated in chapter

9.)

Shallow philosophical, linguistic and psychological disputes about innate or non-empirical knowledge are being replaced by much harder and deeper explorations of exactly what pre-existing knowledge is required, or sufficient, for particular types of empirical and non-empirical learning. What knowledge of two- and three-dimensional geometry and of physics does a robot need in order to be able to interpret its visual images in terms of tables, chairs and dishes to be carried to the sink? What kind of knowledge about its own symbolisms and symbol-manipulating procedures will a baby robot need in order to stumble upon and understand the discovery that counting a row of buttons from left to right necessarily produces the same result as counting from right to left, if no mistakes occur? (More on this sort of thing in the chapter on learning about numbers.)

Similarly, philosophical debates about the possibility of 'synthetic apriori' knowledge dissolve in the light of new insights into the enormous variety of ways in which a computational system (including a human society?) may make inferences, and perhaps discover necessary truths about the capabilities and limitations of its current stock of programs. For an example see the book by Sussman about a program which learns to build better programs for stacking blocks by analysing why initial versions go wrong. (G.J. Sussman, *A Computational Model of Skill Acquisition,* American Elsevier, 1975.)

Epistemology, developmental psychology, and the history of ideas (including science and art) may be integrated in a single computational framework. The chapters on the aims of science and on number concepts are intended as a small step in this direction.

12. One of the bigger obstacles to progress in science and philosophy is often our inability to tell when we lack an explanation of something. Before Newton, people thought they understood why unsupported objects fell. Similarly, we think practice explains learning, familiarity explains recognition, desire explains action. Philosophers often assume that if you have experienced instances and non-instances of some concept, then this 'ostensive definition' suffices to explain how you could have learnt this concept. So our experience of seeing blue things and straight lines is supposed to explain how we acquire the concepts *blue* and *straight.* As for *how* the relevant aspects of instances and non-instances are noticed, related to one another and to previous experiences, and how the irrelevant aspects are left out of consideration the question isn't even asked. (Winston asked it, and gave some answers to it in the form of a primitive learning program: see his 1975.) Psychologists don't normally ask these questions either: having been indoctrinated with the paradigm of dependent and independent variables, they fail to distinguish a study of the *circumstances* in which some behaviour does and does not occur, from a search for an *explanation* of that behaviour.

People assume that if a person or animal wants something, then this, together with relevant beliefs, suffices to explain the resulting actions. But no decent theory is offered to explain *how* desires and beliefs are capable of generating action, and in particular no theory of how an individual finds relevant beliefs in his huge store of information, or how conflicting motives enter into the process, or how beliefs, purposes, skills, etc. are combined in the design of an action (e.g. an utterance) suited to the current situation. The closest thing to a theory in the minds of most people is the model of desires as physical forces pushing us in different directions, with the strongest force winning. The mathematical theory of games and decisions is a first crude attempt to improve on this, but is based on the false assumptions that people start with a well-defined set of alternative actions when they take decisions.

Work in artificial intelligence on programs which formulate and execute plans is beginning to unravel some of the intricacies of such processes. My chapter on aspects of the mechanism of mind will discuss some of the problems. (Chapter 6).

By trying to turn our explanations and theories into designs for working systems, we soon discover

their poverty. The computer, unlike academic colleagues, is not convinced by fine prose, impressive looking diagrams or jargon, or even mathematical equations. If your theory doesn't work then the *behaviour* of the system you have designed will soon reveal the need for improvement. often errors in your design will prevent it behaving at all.

Books don't behave. We have long needed a medium for expressing theories about behaving systems. Now we have one, and a few years of programming explorations can resolve or clarify some issues which have survived centuries of disputation.

Progress in philosophy (and psychology) will now come from those who take seriously the attempt to *design a person.* I propose a new criterion for evaluating philosophical writings: could they help someone designing a mind, a language, a society or a world?

The same criterion is relevant to theorising in psychology. The difference is that philosophy is not so much concerned with finding the correct explanation of *actual* human behaviour. Its aims are more general. For more on the difference see chapters 2 and 3.

13. A frequently repeated discovery, using the new methodology, is that what seemed simple and easy to explain turns out to be very complex, requiring sophisticated computational resources, for instance: seeing a dot, remembering a word, learning from an example, improving through practice, recognising a familiar shape, associating two ideas, picking up a pencil. Of course, it may be that for all these achievements there are simple explanations, of kinds hitherto quite unknown. But at least we have learnt that we don't know them, and that is real progress. This also teaches a new respect for the intellects of infants and other animals. How does a bee manage to alight on a flower without crashing into it?

14. There are some interesting implications of the points made in 7 and 8 above. I mentioned that two computational processes may be mutually supportive. Similarly, two procedures may contain each other as parts, two information structures may contain each other as parts. More generally, a whole system may be built up from large numbers of *mutually recursive* procedures and data-structures, which interlock so tightly that no element can be properly defined except in terms of the whole system. (Recursive rules in formal grammars illustrate the same idea.) Since the system cannot be broken down hierarchically into parts, then parts of those parts, until relatively simple concepts and facts are reached, it follows that anyone learning about the system has to learn many different interrelated things in parallel, tolerating confusion, oversimplifications, inaccuracies, and constantly altering what has previously been learnt in the light of what comes later.[3]

So the process of *learning* a complex interlocking network of circular concepts, theories and procedures may have much in common with the task of *designing* one.

If all this is correct it not only undermines philosophical attempts to perform a logical analysis of our concepts in terms of ever more primitive ones (as Wittgenstein, for example, assumed possible in his *Tractatus Logico Philosophicus)*, it also has profound implications for the psychology of learning and for educational practice. It seems to imply that learning may be a highly creative process, that cumulative educational programmes may be misguided, and that teachers should not expect pupils to get things right while they are in the midst of learning a collection of mutually recursive concepts. This theme will be illustrated in more detail in the chapter on learning about numbers.

(One implication is that this book cannot be written in such a way as to introduce readers to the main ideas one at a time in a clear and accurate way. Readers who are new to the system of concepts will have to revisit different portions of the book frequently. No author has the right to expect this. The book is therefore quite likely to fail to communicate.)

15. Much of what is said in this book simply reports *common sense.* That is, it attempts to articulate

much of the sound intuitive knowledge we have picked up over years of interacting with the physical world and with other people.

Making common sense explicit is the goal of much philosophising. Common sense should not be confused with *common opinions,* namely the beliefs we can readily formulate when asked: these are often false over-generalisations or merely the result of prejudice. Common sense is a rich and profound store of information, not about laws, but about what people are capable of doing, thinking or experiencing.

But common sense, like our knowledge of the grammar of our native language, is hard to get at and articulate, which is one reason why so much of philosophy, psychology and social science is vapid, or simply false.

Philosophers have been struggling for centuries to develop techniques for articulating common sense and unacknowledged presuppositions, such as the techniques of conceptual analysis and the exploration of paradoxes. Artificial intelligence provides an important new tool for doing this. It helps us find our mistakes quickly. One reason for this is that attempts to make computers understand what we say soon break down if we haven't learnt to articulate in the programs the presuppositions and rich conceptual structures which we use in understanding such things. (See Abelson, 'The structure of belief systems', and Schank & Abelson, 1977.)

Further, when you've designed a program whose behaviour is meant to exemplify some familiar concept, such as learning, perceiving, conversing, or achieving a goal, then in trying to interact with the program and in experiencing its behaviour it often happens that you come to realise that it does not really exemplify your concept after all, and this may help you to pin down features of the concept, essential to its use, which you had not previously noticed. So artificial intelligence contributes to conceptual analysis. (The interaction is two-way.)

16. Of course, merely *imagining* the program's behaviour would often suffice: doing the program isn't necessary in principle. But one of the sad and yet exhilarating facts most programmers soon learn is that it is hard to be sufficiently imaginative to anticipate the kinds of behaviour one's program can produce, especially when it is a complex system capable of generating millions of different kinds of processes depending on what you do with it. It is a myth that programs do just what the programmer intended them to do, especially when they are interacting with compilers, operating systems and hardware designed by someone else. The result is often behaviour that nobody planned and nobody can understand.

Thus new possibilities are discovered. Such discoveries may serve the same role as thought-experiments have often done in physics. So computational experiments may help to extend common sense as well as helping us to analyse it.

17. One of the things I have been trying to do is undermine the conflict between those who claim that a scientific study of man is possible and those who claim it isn't. Both sides are usually adopting a quite mistaken view of the essence of science. Bad philosophical ideas seem to have a habit of pervading a whole culture (like the supposed dichotomy between the emotional, intuitive aspects of people and the cognitive, intellectual, or rational aspects -- a dichotomy I have tried to undermine elsewhere).

The chapter on the aims of science attempts to correct widespread but mistaken views about the nature of science. I first became aware of the mistakes under the influence of linguistics and artificial intelligence.

18. One of the main themes of the revolution is that the pure scientist needs to behave like an engineer: designing and testing *working* theories. The more complex the processes studied, the closer

the two must become. Pure and applied science merge. And philosophers need to join in.

19. I'll end with one more wildly speculative remark. Social systems are among the most complex computational processes created by man (whether intentionally or not). Most of the people currently charged with designing, maintaining, improving or even studying such processes are almost completely ignorant of the concepts, and untrained in the skills, required for thinking about very complex interacting processes. Instead they mess about with *variables* (on ordinal, interval or ratio scales), looking for *correlations* between them, convinced that *measurement* and *laws* are the stuff of science, without recognizing that such techniques are merely useful stop-gaps for dealing with phenomena you don't yet understand. In years to come, our willingness to trust these politicians, civil servants, economists, educationalists and the like with the task of managing our social system will look rather laughable. I am not suggesting that programmers should govern us. Rather, I venture to suggest that if everyone were allowed to play with computers from childhood, not only would education become much more fun and stretch our minds muchfurther, but people might be a lot better equipped to face many of the tasks which currently defeat us because we don't know how to think about them. Computer 'experts' would find it harder to exploit us.

1.4. What is Artificial Intelligence?

The best way to answer this question is to look at the aims of A.I., and some of the methods for achieving those aims, and to show how the subject is decomposable into sub-domains and related to other disciplines. This would require a whole book, which is not my current purpose. So I'll give an incomplete answer by describing and commenting on some of the aims. AI is not just the attempt to make machines do things which when done by people are called ``intelligent''. It is much broader and deeper than this. For it includes the scientific and philosophical aims of *understanding* as well as the engineering aim of *making*.

The aims of Artificial Intelligence

1. **Theoretical analysis of possible effective explanations of intelligent behaviour.**
2. **Explaining human abilities.**
3. **Construction of intelligent artefacts.**

Comments on the aims:

a. The first aim is very close to the aims of Philosophy. The main difference is the requirement that explanations be 'effective'. That is they should form part of, or be capable of contributing usefully to the design of, a working system, i.e. one which generates the behaviour to be explained.

b. The second aim is often formulated, by people working in A.I., as the aim of designing machines which 'simulate' human behaviour, i.e. behave like people. There are many problems about this, e.g. which people? People differ enormously. Also what does like' mean? Programs, mechanisms, and people may be compared at many different levels.

c. The programming of computers is not an essential part of the first two aims: rather it is a research method. It imposes a discipline, and provides a tool for finding out what your explanations are theoretically capable of explaining. Sometimes they can do more than you intended usually less.

d. People doing A.I. do not usually bother much about experiments or surveys of the kinds psychologists and social scientists do, because the main current need is not for more *data* but for better theories and theory-building concepts and formalisms, so that we can begin to explain the masses of data we already have. (In fact a typical strategy for getting theory-

building off the ground, in A.I. as in other sciences, is to try to explain idealised and simplified situations, in which much of the available data are ignored: e.g. A.I. programs concerned with 'toy' worlds (like the world of overlapping letters described in chapter 9), and physicists treating moving objects as point masses.)

e. An issue which bothers psychologists is how we can tell whether a particular program really does explain some human ability, as opposed to merely mimicking it. The short answer is that there is never any way of establishing that a scientific explanation is correct. However, it is possible to compare rival explanations, and to tell whether we are making progress. Criteria for doing this are formulated in chapter 2.

f. The notion of 'intelligent behaviour' in the first aim is easy to illustrate but hard to define. It includes behaviour based on the ability to cope in a systematic fashion with a range of problems of varying structures, and the ability (consciously or unconsciously) to build, describe, interpret, compare, modify and use complex structures, including symbolic structures like sentences, pictures, maps and plans for action. A.I. is not specially concerned with unusual or meritorious forms of intelligence: ordinary human beings and other animals display the kinds of intelligence whose possibility A.I. seeks to explain.

g. It turns out that there is not just one thing called 'intelligence', but an enormous variety of kinds of expertise the ability to see various kinds of things, the ability to understand a language, the ability to learn different kinds of things, the ability to make plans, to test plans, to solve problems, to monitor our actions, etc. It also includes the ability to have motives, emotions, and attitudes, e.g. to feel lonely, embarrassed, proud, disgusted, elated, and so on. Each of these abilities involves domain-specific knowledge (factual and procedural knowing that and knowing how). So, much current work in A.I. is exploration of the *knowledge* underlying competence in a variety of specialised domains seeing blocks, understanding children's stories, making plans for building things out of blocks, assembling bits of machinery, reading handwriting, synthesising or checking computer programs, solving puzzles, playing chess and other games, solving geometrical problems, proving logical and mathematical theorems, etc.

   I.e. a great deal of A.I. research is highly 'domain-specific', and amounts to an attempt to explicitly formulate knowledge people already use unconsciously in ordinary life or specialised activities. This is closely related to conceptual analysis as practised by linguists and philosophers. (See Chapter 4.)

h. Alongside all this, there is the search for generality. So research is in progress on possible computing mechanisms and concepts which are not necessarily relevant only to one domain, but may be useful, or necessary, for explaining many different varieties of intelligence, e.g. mechanisms concerned with good ways of storing and retrieving information, making inferences, controlling processes, allowing sub-processes to interact and influence one another, allowing factual knowledge to be translated into procedural forms as required, etc. However, the role of general mechanisms seems to be much less important in explaining intelligent abilities than the role of domain specific knowledge.

i. As pointed out below, much of the domain-specific research overlaps with research in other disciplines, e.g. Linguistics, Psychology, Education, Philosophy, Anthropology, and perhaps Physiology. For example, you can't make a computer understand English without studying syntactic, semantic and pragmatic rules of English, that is, without doing Linguistics.

j. A major effect of A.I. research as already mentioned is to establish that apparently simple tasks, like seeing a line, may involve very complex cognitive processes, using substantial prior

knowledge.

k. One side-effect of attempts to understand human abilities well enough to give them to computers, has been the introduction of some new approaches to teaching those abilities to children, for instance LOGO projects (see papers by Papert). These projects use a programming language based on programming languages developed for A.I. research, and they teach children and other beginners programming using such a language. These languages are much more suitable for teaching beginners than BASIC or FORTRAN, the most commonly used languages, because (a) they are very much more powerful, making it relatively easy to get the computer to do complex things and (b) they are not restricted to numerical computations. For example, LOGO, used at MIT and Edinburgh University, and POP-2, which we use at Sussex University, provide facilities suitable for manipulating words and sentences, drawing pictures, etc. (See Burstall et al. 1971.)

l. A.I. gives people much more respect for the achievements of children, and more insight into the problems they have to solve in learning what they do. This leads to a better understanding of possible reasons for not learning so well.

Note

The remaining chapters, apart from chapter 10 should be readable in any order. On the whole, people knowledgeable about philosophy and ignorant of computing will probably find chapters 2 to 5 easier than the following chapters. People interested in trying to understand how people work, and not so concerned with abstract methodological issues, may find chapters 2 to 5 tedious (or difficult?), and should start with Part Two, though they'll not be able to follow all the methodological asides, which refer back to earlier chapters.

1.5. Conclusion

The primary aim of my research is to understand aspects of the human mind. Different people will be interested in different aspects, and many will not be interested in the aspects I have chosen: scientific creativity, decision making, visual perception, the use of verbal and non-verbal symbolisms, and learning of elementary mathematics. At present I can only report fragmentary progress. Whether it is called philosophy, psychology, computing science, or anything else doesn't really interest me. The methods of all these disciplines are needed if progress is to be made. It may be that the human mind is too complex to be understood by the human mind. But the desire to attempt the impossible seems to be one of its persistent features.

Endnotes

(1) I write 'program' not 'programme' since the former is a technical term referring to a collection of definitions, instructions and information expressed in a precise language capable of being interpreted by a computer. For more details see J. Weizenbaum, *Computer Power and Human Reason.* There is much in this book that I disagree with, but it is well worth reading, and may be a useful antidote to some of my excesses.

(2) A compiler is a program which translates programs from one programming language into another. E.g. an ALGOL compiler may translate ALGOL programs into the 'machine code' of a particular computer.

(3) Apparently Hegel anticipated some of these ideas. His admirers might advance their understanding of his problems by turning to the study of computation.

Last updated: 28 Jan 2007 (Minor reformatting).

PART ONE: METHODOLOGICAL PRELIMINARIES

CHAPTER 2

WHAT ARE THE AIMS OF SCIENCE?[1]

Part One: Overview

2.1.1. Introduction

Very many persons and institutions are engaged in what they call scientific research. Do their activities have anything in common? They seem to ask very different sorts of questions, about very different sorts of objects, events and processes, and they use very different methods for finding answers.

If we ask scientists what science is and what its aims are, we get a confusing variety of answers.

Whom should we believe? Do scientists really know what they are doing, or are they perhaps as confused about their aims and methods as the rest of us? I suggest that it is as hard for a scientist to characterise the aims and methods of science in general as it is for normal persons to characterise the grammatical rules governing their own use of language. But I am going to stick my neck out and try.

If we are to understand the nature of science, we must see it as an activity and achievement of the human mind alongside others, such as the achievements of children in learning to talk and to cope with people and other objects in their environment, and the achievements of non-scientists living in a rich and complex world which constantly poses problems to be solved. Looking at scientific knowledge as one form of human knowledge, scientific understanding as one form of human understanding, scientific investigation as one form of human problem-solving activity, we can begin to see more clearly what science is, and also what kind of mechanism the human mind is.

I suggest that no *simple* slogan or definition, such as can be found in textbooks of science or philosophy can capture its aims. For instance, I shall try to show that it is grossly misleading to characterise science as a search for laws. Science is a complex network of different interlocking activities with multiple practical and theoretical aims and a great variety of methods. I shall try to describe some of the aims and their relationships. Oversimple characterisations, by both scientists and philosophers, have led to unnecessary and crippling restrictions on the activities of some would-be scientists, especially in the social and behavioural sciences, and to harmfully rigid barriers between science and philosophy.

By undermining the slogan that science is the search for laws, and subsidiary slogans such as that quantification is essential, that scientific theories must be empirically refutable, and that the methods of philosophers cannot *serve* the aims of scientists, I shall try to liberate some scientists from the dogmas indoctrinated in universities and colleges. I shall also try in later chapters to show philosophers how they can contribute to the scientific study of man, thereby escaping from the

barrenness and triviality complained of so often by non-philosophers and philosophy students.

An important reason for studying the aims and methods of science is that it may give us insights into the learning processes of children, and help us design machines which can learn. Equally, the latter project should help us understand science. A side-effect of my argument is to undermine some old philosophical distinctions and pour cold water on battles which rage around them like the distinction between subjectivity and objectivity, the distinction between science and philosophy and the battles between empiricists and rationalists.

My views have been powerfully influenced by the writings of Karl Popper. However, several major points of disagreement with him will emerge.

2.1.2. First crude subdivision of aims of science

Science has not just one aim but several. The aims of scientific investigation can be crudely subdivided as follows:

1. To extend man's knowledge and understanding of the form and contents of the universe (*factual aims*),
2. To extend man's control over the universe, and to use this to improve the world (*technological or practical aims*),
3. To discover how things ought to be, what sorts of things are good or bad and how best to further the purposes of nature or (in the case of religious scientists) God (*normative aims*).

Whether the third aim makes sense (and many scientists and philosophers would dispute this) depends on whether it is possible to derive values and norms from facts. I shall not discuss it as it is not relevant to the main purposes of this book. The second kind of aim will not be given much attention either, except when relevant to discussions of the first kind of aim, on which I shall concentrate.

These aims are not restricted to science. We all, including infants and children, aim to extend our knowledge and understanding: science is unique only in the degree of rigour, system and co-operation between individuals involved in its methods. For the present, however, I shall not explore the *peculiarities* of science, since what it has in *common* with other forms of acquisition of knowledge has been too long neglected, and it is the common features I want to describe.

In particular, notice that one cannot have the aim of *extending* one's knowledge unless one presupposes that one's knowledge is incomplete, or perhaps even includes mistakes. This means that pursuing science requires systematic self-criticism in order to find the gaps and errors. This distinguishes both science and perhaps the curiosity of young children from some other belief systems, such as dogmatic theological systems and political ideologies. (See chapter 6 for the role of self-criticism in intelligence.) But it does not distinguish science from philosophy. Let us now examine the factual aims of science more.closely.

2.1.3. A further subdivision of the factual aims: form and content

The aims of extending knowledge and understanding can be subdivided as follows:

(1.a) Extending knowledge of the form of the world:

> Extending knowledge of what sorts of things are possible and impossible in the world, and how or why they are (the aim of interpreting the world, or learning about its *form).* (This will be further subdivided below.)

NOTE: I would now express the aim of 'extending knowledge of what sorts of things are possible' in terms of 'extending the ontology' we use. This is also part of the process of child development, e.g. as illustrated in this presentation.
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0604
'Ontology extension' in evolution and in development, in animals and machines.

(l.b) Extending knowledge of the content of the world:

> Extending knowledge of what particular objects, events, processes, or states of affairs exist or existed in particular places at particular times (the aim of acquiring 'historical' knowledge, or learning about the *contents* of the world).

A similar distinction pervades the writings of Karl Popper, though he would disagree with some of the things I say below about (1.a). Different branches of science tend to stress one or other of these aims, though both aims are usually present to some extent. For instance, physics is more concerned with aim (1.a), studying the form of the world, whereas astronomy is perhaps more concerned with (1.b), studying the contents.

Geology, geography, biology, anthropology, human history, sociology, and some kinds of linguistics tend to be more concerned with (1.b), i.e. with learning about the particular contents of particular parts of the universe. Chemistry, some branches of biology, economics and psychology attempt to investigate truths not so restricted in scope. In the jargon of philosophers, (1.a) is concerned with universals, (l.b) with particulars.

However, the two scientific aims are very closely linked. One cannot discover what sorts of things are *possible,* nor test explanatory theories, except by discovering particular facts about what *actually* exists or occurs. Conversely, one cannot really understand *particular* objects, events, processes, etc., except insofar as one classifies and explains them in the light of more general knowledge about what *kinds* of things there can be and how or why. These two aims are closely linked in all forms of learning about the world, not only in science. The study of form and the study of content go hand in hand. (This must be an important factor in the design of intelligent machines.)

I have characterised these aims in a *dynamic* form: the aim is to extend knowledge, to go on learning. Some might say that the aim is to arrive at some terminal state when everything is known about the form and content of the world, or at least the form. There are serious problems about whether this suggestion makes sense: for example how could one tell that this goal had been reached? But I do not wish to pursue the matter. For the present, it is sufficient to note that it makes sense to talk of extending knowledge, that is removing errors and filling gaps, whether or not any final state of complete knowledge is possible. Some of the criteria for deciding what is an extension or improvement will be mentioned later.

Many philosophers of science have found it hard to explain the sense in which science makes progress, or is cumulative. (E.g. Kuhn (1962), last chapter.) This is because they tend to think of science as being mainly concerned with laws; and supposed laws are constantly being refuted or replaced by others. Very little seems to survive. But if we see science as being also concerned with knowledge of what is possible, then it is obviously cumulative. For a single instance demonstrates a new possibility and, unlike a law, this cannot be refuted by new occurrences, even if the possibility is re-described from time to time as the language of scientists evolves.

Hypotheses about the *limits* of possibilities (laws) lack this security, for they are constantly subject to revision as the boundaries are pushed further out, by newly discovered (or created) possibilities. Explanations of possibilities and their limits frequently need to be refined or replaced, for the same reason. But this is all a necessary part of the process of learning and understanding more about what is

possible in the world. (This is true of child development too.) It is an organic, principled growth. Let us now look more closely at aim (1.a), the aim of extending knowledge of *the form* of the world.

Part Two: Interpreting the world

2.2.1. The interpretative aims of science subdivided

The aim (l.a) of interpreting the world, or learning about its form, can be subdivided into several subgoals listed below. They are all closely related. To call some of them 'scientific' and others 'metaphysical' or 'philosophical', as empiricists and Popperians tend to do, is to ignore their inter-dependence. Rather, they are all aspects of the attempt to discover what is and what is not possible in the world and to understand why.

All the following types of learning will ultimately have to be catered for in intelligent machines.

a. Development of *new concepts and symbolisms* making it possible to conceive of, represent, think about and ask questions about new kinds or ranges of possibilities (e.g. new kinds of physical substances, events, processes, animals, mental states, human behaviour, languages, social systems, etc.). This aim includes the construction of taxonomies, typologies, scales of measurement and notations for structural descriptions of chemical compounds or sentences, or processes. This extension of our conceptual and symbolic powers is one of the major functions of mathematics in science. A major boost has recently come from computing studies.

b. Extending knowledge of what kinds of things (including events and processes) *are possible in the world',* i.e. what kinds of things are not merely conceivable or representable *but really can exist or occur.* Finding our what actually exists, and trying to make new things exist, are often means to this end. We can distinguish knowledge of absolute possibility concerning a phenomenon X (X can exist) from knowledge of relative possibility (X can exist in conditions C). Extending knowledge of relative possibilities for X is an important way of extending knowledge of what is possible. All this should be distinguished from (e) below, the goal of finding out what kinds of things are most likely, common or frequent, either absolutely or in specified conditions. The latter is a concern with *probabilities* not *possibilities.* Subgoal (b) clearly presupposes (a), for one can only acknowledge possibilities that one can conceive of, describe or represent.

c. Constructing *theories to explain known possibilities:* i.e. theories about the underlying structures, mechanisms, and processes capable of generating such possibilities. For instance, a theory of the constituents of atoms may explain the possibility of chemical elements with different properties. Generative grammars are offered by linguists as explanations of how it is possible for us to understand an indefinitely large set of sentences. 'How is this possible?' is the typical form of a request for this kind of explanatory theory, and should be contrasted with the question 'Why is this so?' or 'Why is this impossible?', discussed in (f), below. Artificial intelligence models provide a major new species of explanations of possibilities. E.g., they explain the possibility of various kinds of mental processes, including learning, perceiving, solving problems, and understanding language. Clearly (c) presupposes (b), and therefore (a).

d. Finding limitations on combinations of known possibilities. These are often called laws of nature: for instance to say that it is a law of nature that all X's are Y's is to say that it is *impossible* for something to be both an X and not a Y. It is these laws, limitations or impossibilities which make the world relatively stable and predictable. This goal, like (c), presupposes (b), since one can only discover limitations of possibilities if one already knows about those possibilities. (This subgoal of science is the one most commonly stressed in the

writings of scientists and philosophers. It subsumes the goal of discovering causal connections, since 'X causes Y' means, roughly 'the occurrence of X makes the non-occurrence of Y impossible.')

e. Finding *regular or statistical correlations* between different possibilities, for instance correlations of the form In conditions C, 90% of all X's are Y's'. This is a search for probabilities. It presupposes (b) for the same reason as (d) does. Except in quantum physics, the search for such statistical correlations is really only a stopgap or means towards acquiring a deeper understanding of the sort described in (d), above. Alternatively, it may be an aim of a historical science: facts about relative frequencies and proportions of various kinds of objects, events or processes are often important facts about the *content* of a particular part of the world. For instance, most of the correlations unearthed by social scientists are culture-relative. Such information may have practical value despite its theoretical poverty.

f. Constructing *theories to explain known impossibilities, laws and correlations*. Such theories answer 'Why?' questions, and are generally refinements of the theories described in (c). That is, explaining limits of possibilities (i.e. explaining laws) presupposes or refines an explanation of the possibilities limited. The theory of molecules composed of atoms which can recombine explains the *possibility* of chemical change. Further refinements concerning weights and valencies of atoms explain the observed *limitations:* the laws of constant and multiple proportions.

g. Detecting and eliminating inadequate concepts, symbolisms, beliefs about what is and is not possible, and inadequate explanations of possibilities and laws. That this is a subgoal of science is, as already remarked, implied by saying that an aim of science is to *extend* knowledge. As many philosophers of science have pointed out, it is not generally possible to *prove* explanatory theories in science; at most they can only be refuted or shown to be inadequate in some way. Moreover, when several candidates survive refutation, the most that can be done is to compare their relative merits and faults, without necessarily establishing the absolute superiority of one over the other. It is often assumed that the only kinds of proper tests are empirical (i.e. observations of new facts, in experiments or in nature). However, we shall see that many important tests are not empirical.

If forced to summarise all this in a single slogan, one could say: *A major aim of science is to find out what sorts of things are and are not possible in the world, and to explain how and why.*

A similar aim must motivate intelligent learning machines.

Though too short to be clear, this may be a useful antidote to more common slogans stressing the discovery and explanation of laws and regularities. Such slogans lead to an excessive concern with prediction, control and testing, topics mainly relevant to subgoals (d) to (g), while insufficient attention is paid to the more fundamental aims (a) to (c), especially in psychology and social science. The result is often misguided research, theorising and teaching.

I shall say more about these three fundamental aims later. The next two sections contain further general discussion of the relations between these seven interpretative aims, and the previously mentioned historical and technological aims of science.

2.2.2. More on the interpretative and historicalaims of science

Unlike the historical scientist, the interpretative scientist is interested in actual objects, events or

situations only insofar as they are *specimens* of *what is possible.* The research chemist is not interested in the fact that *this* particular sample of water was, on a certain day, decomposed into hydrogen and oxygen in *that* laboratory, except insofar as this illustrates something universal, such as the *possibility* of decomposing water.

This possibility refutes the theory that water is a chemical element and corroborates the alternative hypothesis that all water is composed of hydrogen and oxygen, and also more general theories about possible kinds of transformations of matter. Similarly, although an 'historical' biologist may be interested in recording, for a fascinated public, the flora and fauna of a foreign isle, or the antics of a particularly intelligent chimpanzee, the 'interpretative' biologist is interested only insofar as they illustrate something, such as what *kinds* of plants and animals can exist (or can exist in certain conditions), or what *kinds* of behaviour are possible for a chimpanzee, or for some other class containing the animal in question.

In short, the interpretative scientist studies *the form* of the world, using the *contents* only as evidence, whereas the historical scientist simply studies the contents. There is no reason why any one science, or scientist, should be classified entirely as interpretative, or entirely as historical. Different elements may intermingle in one branch of science. For instance, a linguist studying a particular dialect is an interpretative scientist insofar as he is not concerned merely to record the actual set of sentences uttered by certain speakers of that dialect, but to characterise the full range of sentences that *would* or *could* be intelligible to an ordinary speaker of that dialect, namely, a range of possibilities.

However, insofar as he is interested merely in finding out exactly what dialect is intelligible to a certain spatio-temporally restricted group of persons, he is an historical linguist, as contrasted with a linguist who is interested in this dialect primarily as a *sample* of the kinds of language which human societies *can* develop: the attempt to characterise this set of possible languages is often called the search for linguistic universals.

Thus a richer terminology would be required for a precise description of hybrid historical and interpretative aims. This is not relevant to our present concerns and will not be pursued further.

Like the interpretative aim, the "historical" aim of finding out about the contents of particular bits of the world must also be built into intelligent machines. Moreover, the pursuit of these two aims by a machine will interact, as in science.

2.2.3. Interpreting the world and changing it

It is often said that the utility of science is to be explained in terms of the discovery of laws and regularities with predictive content. This is how the factual aims (1) subserve the technological aims (2), distinguished previously. For instance, a law which states that whenever A occurs, in situations of type S, B will occur, can be used not only to explain and predict particular occurrences of B, but also as a basis for making B occur, if either of A or S occurs and one can make the other occur. Similarly, knowledge of laws may provide a basis for *preventing* unwanted events. This pragmatic value of laws is not here disputed. However, the discovery, representation, and explanation of absolute or relative *possibilities* is also of great practical importance, even in cases where it is not known how to predict, produce or prevent their realisation.

For example, knowing that rain is possible and wanting to stay dry, one can take a waterproof covering whenever one goes out. More generally, one can take precautions to prevent the effects of an unwanted possibility, even if one cannot predict or prevent it.

Similarly, one can take steps to get the best out of possibilities one knows about but cannot predict or

produce, like building tanks to catch water in case it rains, which might be worth doing even if one had no idea how often rain fell, provided one needed the water enough and had time and materials to spare.

The discovery of possibilities may have technological significance in less direct ways. Knowing that something is possible can provide a boost to research into an understanding of how and why, so that its occurrence may be predicted or brought about, or new variants produced. Knowledge that it was possible for things heavier than air to fly, namely birds, provoked research into ways of enabling men and machines to do so. That was a case of a possibility demonstrated by actual instances, then extended to a wider range of instances.

Sometimes a possibility is explained by a theory before instances are known, and this again can have great technological importance, as in the case of Einstein's discovery of the possibility of converting mass into kinetic energy, or the theoretical discovery of the possibility of lasers before they were made. Much of engineering design consists of demonstrating that some new phenomenon is possible and showing how, or that some possibility can be produced in new ways or in new conditions. An intelligent planning system may also need to be able to generate types of possibilities before instances are known actually to exist. This is commonplace in engineering design.

Formally this technological activity has much in common with the supposedly purer or more theoretical activity of inventing a new theory to explain some previously known possibility, or using the ideas of one science to explain possibilities observed in another, for instance using physics to explain chemical possibilities, and using chemistry to explain the very complicated possibility of sexual reproduction. (See J. Watson, 1968.) 'Pure' science first discovers instances of possibilities then creates explanations of those possibilities whereas 'applied' science uses explanations of possibilities to create instances. The kinds of creativity and modes of reasoning involved are often similar. More generally, any form of intelligent action requires an understanding of possibilities. One cannot change the world sensibly without first interpreting it, even though attempting to change things is often indispensable for correcting mistaken interpretations and deepening one's understanding. Acting intelligently in a situation requires a survey of possibilities, which requires an understanding of the potential for change in the situation. For example, opening a window requires a grasp of the possibilities for movement in the window and its catch. But this requires interpreting what is actual, i.e. relating it to general knowledge of what sorts of things are possible in what circumstances: so action requires knowledge of the form of the world. Grasping new possibilities often involves inventing new concepts, new languages in which to represent them, a topic discussed later.

Much more could be said about relations between the interpretative aims of science, and the historical and technological aims. Instead, let's take a closer look at some of the interpretative aims of science, the aims concerned with learning about and understanding possibilities. We shall attempt to clarify the similarities and differences between these aims, and then proceed to formulate criteria for assessing some of the achievements of scientists.

Part Three: Elucidation of subgoal (a)

2.3.1. More on the interpretative aims of science

Earlier I distinguished factual aims of science from technological and normative aims, then divided factual aims into interpretative and historical aims. The interpretative aims were further subdivided into seven components, of which the first three were:

a. Developing new concepts and symbolisms making it possible to conceive of, think about and ask questions about new types of possibilities;

b. Extending knowledge of what kinds of things really are possible, and not merely conceivable;

c. Constructing explanations of how such things are possible.

The three aims are very tightly interconnected. It is very hard to describe the distinctions between them accurately, and I am sure I do not yet understand these matters aright. Moreover, each of them could be further subdivided. Detailed historical analysis is required here, so that similarities and differences between cases can be described accurately and a more satisfactory typology developed: a contribution to the scientific study of science. Alas, this will require the help of persons more scholarly than I. Let's take a closer look at (a).

2.3.2. The role of concepts and symbolisms

Individuals (and cultural groups) can differ not only in the things they know or believe, but also in the possibilities they can grasp, the concepts they use, the generative power of their language, the questions they can ask.

As new concepts and symbolisms are developed, and the language extended, new questions become askable. For instance, people who grasp the concepts 'hotter' and longer' can understand the question whether metal rods get longer when they are made hotter. And they may even be able to grasp crude distinctions between metals according to which grows longer *faster* when heated. But in order to learn to think about whether the change in length is *proportional* to the change in temperature, so that they can then use the constant of proportionality (divided by the length of the rod) to define a numerical 'coefficient of expansion' for each metal, they need to grasp numerical representation of differences in temperature and length ('hotter by how much?', longer by how much?').

Similarly, although people may have a crude grasp of distinctions between velocity and acceleration, and be able to detect gross changes in either, on the basis of their own experiences of moving things, being moved, and perceiving moving objects, nevertheless, until they have learnt how to relate concepts of distance and time to numerical interval scales, they cannot easily make precise distinctions between different velocities, or between acceleration and rate of change of acceleration, nor think of precise relations between these concepts. These familiar examples show the power of extending scientific language by introducing numerical concepts and notations corresponding to old non-numerical concepts. This sort of thing has been so important in physics that many have been deluded into thinking it part of the definition of a scientist that he uses numbers!

The replacement of Roman numerals with the Arabic system is an example of a powerful notational advance. Another was the Cartesian method of using arithmetic to represent geometry and vice versa. Both involved numbers.

2.3.3. Non numerical concepts and symbolisms

Non-numerical conceptual and notational devices have also been important. Examples are concepts used in describing structures of plants and animals, concepts used for describing structures of mechanical systems and electrical circuits (geometrical and topological concepts), taxonomies or typologies, and grammatical concepts (see N. Chomsky (1957).) Non-numerical computing concepts and formalisms are the newest example.

All sorts of notations besides numerical and algebraic ones have played an important role in extending the abilities of scientists to express what they know and want to find out.

Pictures, diagrams, maps, models, graphs, flow charts, and computer programs, have all been used. Examples include: the diagrams used in the study of levers, pulleys, bending beams, and other

mechanical systems; the 'pictures' of molecules used by chemists, for instance, in the following representation of the formation of water from hydrogen and oxygen

```
(H-H, H-H, 0=0) -----> (H-O-H, H-O-H)
```

circuit diagrams used in electronics; optical drawings showing the paths of light rays; plates showing tracks of subatomic particles; and the 'trees' used by linguists to represent structures or sentences. I shall argue later that these non-verbal forms of representation play a part in valid reasoning, scientific and non-scientific, conscious and unconscious.

2.3.4. Unverbalised concepts

Concepts may also be used without being represented explicitly by any external symbol. There are philosophers who dispute that these are cases of the use of concepts, but in the face of well known facts I can only regard this as verbal quibbling. We know that young children and other animals can discriminate, recognise and react intelligently to things which they cannot name or describe. The consistency, creativity and appropriateness of their behaviour shows that they act on the basis of reasons, even if they cannot articulate them or are unaware of them.

The same is true of an adult who cannot describe the features of musical compositions which enable him to recognise styles of composers and appreciate their music, or the cues which enable him to judge another's mood. Non-logicians can often distinguish valid from invalid arguments without being able to say how. They have not learnt the overt language of logicians.

No doubt this is true also of many scientists, especially when they are in the early phases of some kind of conceptual development. They may then, like children and chimpanzees, be unable to articulate fully the reasons they have for some of the decisions they take about interpreting evidence and assessing hypotheses.

Even after going a stage further and learning how to articulate their reasons, scientists may not yet have learned how to *teach* their new concepts to colleagues and rival theorists. So attempts at rational persuasion break down. This has misled some philosophers and historians of science (e.g. Kuhn) into thinking that there are no reasons, and inferring that the decisions of scientists are irrational or non-rational. This is as silly as assuming that a mathematician is irrational simply because he cannot explain a theorem to a four year old child. The child may have much to learn before he can understand the problem, let alone the reasoning, and the mathematician may be a poor teacher.

Concepts are not simple things which you either grasp or don't grasp, or which can be completely conveyed by an explicit definition or axiomatic characterisation. For instance, as work of Piaget has shown so clearly, and Wittgenstein less clearly, very many of our familiar concepts, like 'number', 'more', 'cause', 'moral' and language', are very complex structures of which different fragments may be grasped at different times. In a later chapter I shall illustrate this by analysing some of the complexities children master when they learn to count.

2.3.5. The power of explicit symbolisation

The more of one's concepts and associated procedures one is able to represent explicitly in symbols of some sort, the greater one's power to explore possibilities systematically by manipulating those symbols. For instance, by explicitly characterising aspects of our intuitive grasp of spatial structures in the form of axioms and definitions, one becomes able to experiment with alterations in the axioms and definitions, and thereby invent concepts of non-euclidean or other new sorts of geometries. This kind of "reflective abstraction" should play a role in learning machines one day.

In this way one can learn to think about new sorts of possibilities without waiting to be confronted with them. (This kind of thing may also happen below the level of consciousness, in children and scientists, as part of the process of learning and discovery.) Of course, one may also extrapolate too far, and construct representations of things which are not *really* possible in the world, so empirical investigation of some sort is required to discover whether things which are conceivable or representable can also exist. For instance, merely analysing the concept of an element with atomic number 325 will not decide whether such a thing can occur. This is the reason for distinguishing the first aim of interpretative science, namely extending concepts and symbolisms, from the second aim, namely extending knowledge of what is really possible.

2.3.6. Two phases in knowledge-acquisition: understanding and knowing

It is not always noted in epistemological discussions that there are two important phases or steps in the acquisition of knowledge. Discovering that p is true first of all requires the ability to understand the possibility that p might be true and might be false, which requires grasping the concepts used in the proposition p. The second phase is finding out that p is true, for instance by empirical observation, use of testimony, inference from what is already known, or some combination of these. In the first phase one is able to ask a question, in the second one has an answer. (There may be primitive kinds of knowledge-acquisition, in people and other animals, in which questions are never understood, only information acquired and used. But science is not like this.)

Usually philosophers plunge into discussions of such questions as whether we can know anything about the future, or rationally believe anything about the future, without first asking how a rational being can even *think* about the future or *think* about alternative possible future states of affairs. (Work in artificial intelligence is beginning to explore these problems.)

Philosophers are therefore attempting to assess the rationality of certain decisions on the basis of a drastically incomplete account of the resources that might enter into the decision-making process. The reason why a study of our ability to think of things has been shirked is partly because it is so hard to do, partly because of an unwarranted restriction of rationality to relations between evidence and belief-contents, and partly because many philosophers think that the investigation of conceptual mechanisms is a task for psychologists not philosophers. However, most psychologists never even think of the important questions, and those who do usually lack the techniques of conceptual analysis required for tackling them: so the job does not get done. (Piaget seems to be an exception.)

There is a need for a tremendous amount of research into what it is to understand various sorts of concepts, and what makes it possible. There is also a need for some kind of taxonomy of types of conceptual change, whether in individuals or in cultures.

2.3.7. Examples of conceptual change

Here are some examples of possibilities of conceptual change which still require adequate explanations:

- The child's invention of a new procedure for using his existing counting procedures in order to answer questions of the form 'What number comes before N?'.
- Going from being able to use numbers in counting procedures to being able to use numbers as *objects* which can themselves be counted, sorted, etc.
- Going from being able to use the decimal representation of integers greater than 9 to *understanding the principles* on which it is based.

- Grasping that a procedure so far used on small sets can be extended indefinitely like counting or matching.
- Going from being able to apply some procedure to objects to thinking of the result as a *property* of the object.
- Going from grasping a relation like 'hotter' or longer' to grasping that it can be used to define equivalence classes of objects of the same temperature or length.
- Going from this to grasping the possibility of comparing *differences* in temperature or length (i.e. understanding an interval scale).
- Going from grasping some general concept defined in terms of a structure, or a function, or some combination of structure and function, to grasping systematic principles for subdividing that concept into different categories.
- Learning to separate the structural and functional aspects of a hybrid concept, like 'knife', or 'experiment'.
- Changing a concept by changing the theories in which it is embedded, in the way that the concept of mass was changed by going from Newtonian mechanics to Einstein's mechanics.
- Developing a more powerful symbolism for an old set of concepts: e.g. inventing differential calculus notation for representing changes, inventing co-ordinate representations of geometrical concepts, inventing the use of variables to express generality as in logic or mathematics, or using the concept of a mathematical function to generalise earlier concepts of regularity or correlation.
- Making explicit the principles previously used implicitly in applying a set of concepts as Einstein did for some old concepts of spatial and temporal relations.
- Coming to see something in common between things one has never previously classified together, like mass and energy, particles and waves,straight lines and geodesics on a sphere.
- Going from knowing a set of formulae and how to manipulate them to being able to see their relevance to a variety of new concrete problems e.g. going from understanding algebra to being able to apply it in real life.
- Grasping a relation between an abstract body of mathematics, and a set of unsolved scientific problems.
- Learning to use the concept of 'recursion' in logic, grammar, or programming.

Until these and other conceptual changes are better understood, discussion of 'incommensurability' of scientific theories and of the role of rationality in science is premature. Meanwhile education will continue to be largely a hit and miss affair, with teachers not knowing what they are doing or how it works, When we really can model conceptual development,things will be very different.

To sum up so far. We have been discussing subgoal (a), namely *developing new concepts and symbolisms making it possible to conceive of, think about and ask questions about new types of possibilities* A system of concepts and symbols with procedures for using them constitutes a language. A language which is used to formulate one theory, will usually also contain resources for formulating alternatives, including the negation of the theory and versions of the theory in which some predicate, relational expression or numerical constant is replaced by another.

So concepts and symbols are tools for *generating* possibilities or questions for investigation. They have greater generative power than theories. The scientist who usefully extends the *language* of science, unlike one who simply proposes a new *theory* using existing concepts and symbols, extends the hypothesis-forming powers of the scientists who understand him. In this sense conceptual advances are more profound.

So the important differences between modern scientists and those of the distant past include not merely the statements and theories thought to be true or false, but also which statements and theories could be thought of at all. Not only are more answers known now, but more questions are intelligible. The same applies to development of an individual.

2.3.8. Criticising conceptual systems

Sometimes old questions become unaskable as a result of conceptual change, like questions about phlogiston or absolute velocity, or perhaps 'medical' questions like 'What did he do to deserve this affliction?' Modern medical science contains no means of generating possibilities constituting answers to this question, though both laymen and some medical men (on Sundays?) may still formulate them. (Incompatible systems of concepts and theories may coexist in one mind but that's another story.)

So science is served not only by extending and differentiating existing concepts: rejection of a concept or typology or mode of representation may also serve the aims of science by reducing the variety of dead-end questions and theories. Concepts, typologies, taxonomies, and symbolisms can, like theories, be rationally criticised, and rejected or modified. Any intelligent learning system will need to have procedures for rationally criticising its current conceptual and symbolic resources. (See Winston (1975) for a simple example of a computer program that modifies its own concepts.)

There are several ways in which a typology and associated notation can be rationally criticised. For instance one may be able to make one or more of these criticism:

(a) That there are some possibilities it doesn't allow for,

(b) That it represents as possible some cases which are not *really* possible,

(c) That some of the subdivisions it makes are of no theoretical importance,

(d) That some category within it should be subdivided into two or more categories, because their instances have different relations to the other categories,

(e) That a principle of subdivision fails to decide all known cases, e.g. because of inapplicable tests,

(f) That the classification procedure generates inconsistent classifications for some instances,

(g) That the notation used does not adequately reflect the structural properties of the typology, or of the instances, e.g. when people use diagrams with bogus detail,

(h) That the concepts used generate questions which apparently cannot be answered by empirical investigation (like the question 'How fast is the Earth moving through the aether?'),

(i) That more powerful explanatory theories can be developed using other tools for representing possibilities.

I suspect that some or all of these criteria are used, unconsciously of course, not only by scientists, but also by young children in developing their conceptual systems. They could also play an important role in an intelligent learning machine.

Several of these criteria will remain rather obscure until later. In particular, the first two can only be understood on the basis of a distinction between what is conceivable or representable and what is really possible in the world. We now examine this, in order to explain the difference between the first two interpretative subgoals of science, namely (a) extending what is conceivable or representable and (b) extending knowledge of what is really possible.

Part Four: Elucidating subgoal (b)

2.4.1. Conceivable or representable versus really possible

The second interpretative aim of science is to find out what kinds of things really are possible in the world and not merely conceivable. This includes such aims as finding out what sorts of physical substances, what kinds of transformations of energy, what kinds of chemical reactions, what kinds of astronomical objects and processes, what kinds of plants and animals, what kinds of animal behaviour, what kinds of mental development, what kinds of mental abnormality, what kinds of language and what kinds of social changes can exist or occur.

This aim is indefinitely extensible: having found out that X's can exist or occur, one can then try to find out whether X's can exist or occur in specified conditions C1, C2, C3, ... Similarly, having found that objects can have one range of properties which can change (e.g. length) and can also have another range of properties which can change (e.g. temperature) one can then try to find out whether these properties can change independently of each other in the same object, such as a bar of metal, or a particular object in specified circumstances, such as a bar of metal under constant pressure or tension. *Such further exploration of the limits of combinations of known possibilities merges into the search for laws* and regularities, as explained previously.

We can conceive of, or describe, a lump of wood turning spontaneously into gold, or a human living unclothed in a vacuum, but it does not follow that these things really can exist. What is the difference? First we look at what it is for something to be conceivable, representable, or describable.

2.4.2. Conceivability as consistent representability

As philosophers well know, the subjective feeling of intelligibility, the feeling of having understood or imagined something, is no guarantee that anything consistent was understood, imagined or conceived of. If someone claims to be able to conceive of the set of all sets which do not contain themselves, then provided he is using words in the normal way we can show, by Russell's well known argument, using steps that he will accept if he is reasonable, that he was wrong, or that his 'conceiving' amounted to nothing more than repeating the phrase, or some equivalent, to himself.[2]

A sentence, phrase, picture, diagram, or other complex symbol will, if intelligible, be part of a language which includes syntactic and semantic rules in accordance with which the symbol is to be interpreted. The mere fact that the symbol is syntactically well-formed does not guarantee that it can be interpreted, though it may mislead us into thinking it can. More precisely, it may have a *sense* but necessarily fail to have any *denotation.* Thus the question 'Does the table exist more slowly than the chair?' is syntactically perfect but we can show that so long as the words are used according to normal semantic rules there can be no answer to the question. For, 'more slowly' when qualifying a verb requires that verb to denote a process or sequence involving changes other than the change of time, so that the rate of change or succession can be measured against time. Existence is not such a process, so

rates of existence cannot be compared. (For more on the connection between sense and failure of reference see Sloman (1971b).)

We can use the notion of what is or is not coherently describable or representable in some well defined language or representational system, as an objective semantic notion. What is conceivable to a person, will be what is coherently representable in some symbolic system which he uses, not necessarily fully consciously. It may be very hard, even for him, to articulate the system he uses, but that does not disprove its existence. These notions are as objective as the notion of logical consistency, which is a special case.

However the mere fact that something is, in this sense, representable or conceivable does not mean that it really can exist. Conversely, what can exist need not be representable or conceivable using the symbolic resources available to scientists (or others) at any particular time: their language may need to be extended. Scientists (like children) may be confronted with an instance of some possibility, like inertial motion, diffraction, or curvature of space-time, without seeing it as such because they lack the concepts. (Kuhn, 1962, chapter X, has over-dramatised this by saying they inhabit a different world.)

The word 'possible' as I have used it, and as others use it, tends to slide between the two cases (a) used as a synonym for 'consistently representable or describable using some representational system', as in logically possible', and (b) used to refer to what can occur or exist in the world. This is why the first two interpretative aims of science are not always clearly distinguished. But what is the difference between (a) and (b)?

This is not an easy question to answer. The main difference is that conceivability or representability can be established simply by analysing the sentence or other symbol used and checking that the syntactic and semantic rules of the language in question do not rule out a consistent interpretation (which is not always easy), whereas checking whether something really is or is not possible requires empirical investigation of some sort. The former involves conceptual analysis (see chapter 4), the latter perception, experiments or surveys.

2.4.3. Proving real possibility or impossibility

If an actual example is found, that conclusively establishes its real possibility. To establish real *impossibility* is very much harder, and perhaps it can never be conclusively established. However one can sometimes be fairly sure that something is not possible in the world either because of extensive and varied attempts to realise it, or on the basis of inference from some well established theory. (For instance, I am convinced by physical and biological arguments that it is impossible for a human being to live without clothing in a vacuum.)

However, possibility is not the same as actual existence. To say that it is possible for ten drugged alligators to be painted with red and yellow stripes and then piled into my bath is not to say that this ever has happened or will happen. Similarly, to say that several courses of action are possible for me, is not to say that I shall actually follow all of them. So, in saying that one of the aims of interpretative science is to find out which kinds of things are possible in the world, I do not mean that the aim is to find out which kinds actually exist, as in historical science. The latter is just a means to the former.

What other means are there of deciding that something is really possible, besides finding an instance? Alas, the only answer I can give to this is that we can reasonably, though only tentatively, infer that something is possible if we have an explanation of its possibility. What this amounts to is roughly the following: (a) we can consistently represent it using symbolic resources which have already been shown to be useful in representing what is actual, and (b) it is not ruled out by any well established law or theory specifying limitations on possibilities.

It is clear that these conditions do not conclusively prove something to be possible, for they rest on current theories of the limitations of what is possible and such theories, being empirical, are bound to include errors and omissions, at any stage in the advance of science. Further, these conditions do not yield clear decisions in all cases. For instance, is it reasonable to believe that it is possible for a normal human being to be trained (perhaps starting from birth) to run a mile in three minutes? It may not be clear whether we already know enough to settle such a question.

2.4.4. Further analysis of 'possible' is required

These conditions for proving unrealised possibilities need to be further defined and illustrated. For the present, however, my aim is simply to indicate roughly how something can be shown to be possible without producing an instance. So I have demonstrated that possibility is a different concept from conceivability (or coherent representability), and also different from existence.

But I still have not given anything approximating to a complete analysis: this would require very much more than describing the criteria for deciding whether something is possible or not. It would also require analysis of the role of the concept of possibility in our thinking, problem-solving, deliberating, regretting, blaming, praising, etc., and its relations to a whole family of modal words, such as 'may', 'can', 'might', 'could', 'would', etc. A mammoth task. (For some useful beginnings see Gibbs, 1970 and White, 1975.) A good analysis would be part of a design for a mind.

At any rate, we cannot analyse 'Things of type X are possible' as *synonymous* with 'Either things of type X already exist, or else they are consistently representable in our symbolic system without being ruled out by known laws', since this would define real possibility in terms of the *current* system of concepts and beliefs. We could try a formula like 'Things of type X are possible if and only if they either exist or are consistently representable in some useful representational system and are not ruled out by any true laws'. But this has the disadvantage of presupposing that there exists some complete set of true laws formulated in some unspecified language which correctly defines all the limitations on what is possible in the world. It is by no means clear that such a presupposition is intelligible. Moreover as a definition it introduces a circularity, since it is notoriously hard to define the concept of a law without presupposing the concept of possibility or some related concept.

Despite the remaining obscurities, I hope I have done enough to indicate both that the first two aims of interpretative science are different, and also that they are very closely related. Now for a closer look at the third aim the aim of explaining possibilities.

Part Five: Elucidating subgoal (c)

2.5.1. Explanations of possibilities

A request for an explanation of a possibility or range of possibilities is characteristically expressed in the form 'How is X possible?' Unfortunately, the role of such explanations in our thought is obscured by the fact that not everyone who requires, seeks or finds such an explanation, or who learns one from other people, asks this sort of question explicitly, or fully articulates the explanation when he has understood it. This partially explains why the role of possibilities and their explanations in science has not been widely acknowledged.

Roughly, an explanation of a possibility or range of possibilities can be defined to be some theory or system of representation which *generates* the possibility or set of possibilities, or representations or descriptions thereof. An explanation of a range of possibilities may be/a grammar for those possibilities. A computer program is a good illustration: it explains the possibility of the behaviours it can generate (which may depend on the environment in which it is executed). In this way Artificial

Intelligence provides explanations of intelligent behaviour. There is much to be clarified in these formulations, but first some examples from the history of science.

2.5.2. Examples of theories purporting to explain possibilities

The examples which follow are not all correct explanations. Some have already been superseded and others probably will be.

- The ancient theory of epicycles explained how it was possible for the apparent paths of planets to exhibit irregularities while the actual paths were constructed out of regular circular motions. Known forms of motion were compounded in a representation of new ones.
- The principle of the lever explained how it was possible for a small force to be transformed into a larger force or *vice versa,* in a wide range of situations.
- Newton's gravitational theory explained how it was possible for the moon to produce tides on earth. His theory of the relation between force and acceleration explained how it was possible for water to remain in a bucket swung overhead.
- The atomic theory after Dalton explained how various kinds of chemical transformations were possible without any change in basic substances. (It also explained why the range of possibilities was restricted according to the laws of constant and multiple proportions, so that it was vastly superior to previous atomic theories.)
- The kinetic theory of heat explained, among other things, how it was possible for heating to produce expansion, and how heat energy and mechanical energy could be interconvertible.
- The theory of natural selection explained how it was possible for undirected ('random') mutations to lead to apparently purposive or goal-directed changes in biological species. The theory of genes explained how it was possible for offspring to inherit some but not all of the characteristics of each parent, and for different siblings to inherit different combinations.
- The theory of 'the selfish gene' has been used to explain the possibility of the evolution of altruistic behaviour (Dawkins, 1977.)
- The theory that atoms were composed of protons, neutrons and electrons explained many of the possibilities summarised in the periodic table of the elements, and explained how it was possible for one element to be transformed into another.
- The wave theory of light explained how it was possible for refraction, diffraction and polarisation effects to occur.
- Quantum theory explains how it is possible for particles to produce interference effects, how it is possible for the photo-electric effect (release of electrons from a metal by light) to have a frequency threshold rather than an intensity threshold, and how it is possible for complex molecules to be stable despite thermal buffeting.
- Einstein's theory of general relativity explained how it is possible for mass and energy to be interconvertible, and for light rays to be curved even in a vacuum. Other possibilities explained before specimens were produced include lasers and super-conductivity.

Some of the theories listed so far not only explained possibilities, but also contained enough detail to make prediction, and in some cases control, possible. This is fairly common in physics, though more difficult in biology. In the case of the human sciences (and philosophy) the ability to predict and

control is rare.

- Marx's social theories explained how it was possible for large numbers of people to collaborate peacefully in social and economic practices against their own interest. He also explained how it was possible for such systems to generate forces tending to their own overthrow.
- Popper has tried to explain how it is possible for the growth of scientific knowledge to be based on rational comparisons and assessment of theories, even though no theory can ever be proved to be right or even probable.
- Chomsky's theory that human minds contain representations of generative grammars explains how it is possible for sentences never before heard or uttered nevertheless to be part of a person's language. The theory (see T. Winograd (1973)) that human minds contain certain sorts of procedures or programs explains how it is possible for new sentences to be produced or understood.
- Freud's theories attempted to explain how it is possible for apparently meaningless slips and aberrations of behaviour to be significant actions. Piaget's theories about the structure of many familiar concepts attempt to explain how it is possible for a child to show in some behaviour that he has grasped the concept andin others that he has not.
- In a later chapter I shall sketch a computational mechanism which explains how it is possible for many kinds of knowledge, skills and other resources to be used in a flexible and integrated way by a single person.
- Work in artificial intelligence explains how certain kinds of perception are possible. (E.g. see Chapter 9)
- Emotivist and prescriptivist theories in moral philosophy explain how it is possible for moral language to be meaningful and to perform a useful function without being a sub-species of descriptive language. Frege, Russell and Whitehead, showed how it was possible for a great deal of mathematical knowledge to be based on logical knowledge. (Some of these examples support the view that aims andmethods of philosophy overlap with those of science.)

2.5.3. Some unexplained possibilities

Known possibilities for which explanations are still lacking abound. Consider the possibility of the growth of an oak from an acorn or a chicken from an egg. Fragments of the mechanism are of course understood already, but there is as yet no explanation of how such an apparently simple structure as a seed or fertilised ovum can *control* its own development in such a way as to produce such a complex structure as a plant or animal. In the terminology introduced below, we can say that as yet the/we *structure* of these known possibilities is unexplained, despite the optimism which followed the discovery of the structure ofDNA.

Another unexplained possibility is the evolution of animals with specific intelligent abilities (like the ability to learn to use tools, or to learn to use language) from species lacking these abilities, and in particular the evolution of human beings.

In the case of human psychology, there are very many possibilities taken for granted as part of common sense, yet still without even fragmentary explanations, for instance the possibility of a newborn infant learning whatever human producing a work of art, the possibility of extending an art form or language, the possibility of using knowledge acquired in one context to solve a problem of a

quite different sort, the possibility of relating one's actions to tastes, preferences, principles, hopes, fears, knowledge, abilities, and social commitments, and the possibility of changing one's moral attitudes through personal experience.

There are missing explanations of possibilities in physics and chemistry also. As far as I know, the possibility of mechanical utilisation of fuel energy at levels of efficiency achieved in animals is still not explained.

2.5.4. Formal requirements for explanations of possibilities

The explanations listed earlier may not be correct explanations, but they at least meet formal conditions for explaining certain possibilities, or perhaps would do if precisely formulated. These conditions will be described below. They are generalisations and elaborations of the basic idea, familiar from writings of philosophers like Popper, Hempel and Nagel, that to explain something by means of a theory is to deduce it from the theory, perhaps with some additional premisses.

Such philosophers normally assume that both the theory and what it explains are expressed in the form of sentences, using natural language supplemented by the technical language of the science concerned. It is also assumed that the deduction is *logical,* that is the inference from theory to what it explains can be shown to be valid according to the rules of inference codified by logicians. (This is sometimes generalised to permit cases where the inference is only probabilistic.)

This concept of deduction and the related notion of explanation needs to be generalised in two ways. First of all, other means of representation besides sentences may be used, such as maps, diagrams, three-dimensional models or computer programs. Secondly, *the forms of inference* include not only the *logical* forms (like 'All A's are B's, All B's are C's. Therefore All A's are C's'), but also the manipulation of other representations. An example is the manipulation of diagrams representing molecular structures, in order to explain the possibility of chemical reactions, like the production of water from hydrogen and oxygen.

I shall explain in chapter 7 exactly what 'valid' means and why this generalisation to non-verbal forms of valid inference should be permitted. Just as the semantic rules of verbal languages guarantee that certain transformations of sentences preserve truth, so can semantic rules of non-verbal representations guarantee that certain manipulations preserve denotation. (This generalisation of the concept of a valid inference is central to the analysis of the elusive concepts of 'cause' and 'mechanistic explanation' but that is another story.)

Typical examples of such non-verbal inference methods are: the use of Venn diagrams in set theory, the 'parallelogram' representation of addition of forces, velocities and other vectors, the use of circuit-diagrams in electronics, the use of a map to select a route, the use of a diagram to show how a machine works. On this view the use of models and so-called 'analogies' in science is simply a change of language: one configuration is used to represent another. All the usual talk about isomorphism of models in this context is as misconceived as the theory that sentences in natural language must be isomorphic with things they describe: there are many more kinds of non-verbal representations than isomorphic models. (See Goodman, 1968, Clowes, 1971, and Toulmin, 1953). I was helped to see all this by an unpublished paper by Max Clowes, called 'Paradigms and syntactic models'.)

We now have a minimal requirement for a theory **T** formulated in sentences or other symbolic apparatus to be an explanation of some range of possibilities, namely:

1. Statements or other representations of the range of possibilities should be validly derivable from **T**, according to whatever criteria for validity are generated by the semantics of the

language' used for **T**.

An illustration of this is the use of the theory of bonds between atoms (the theory of valencies) to explain the possibility of a very large number of chemical compounds and transformations. Knowing the kinds of bonds into which the various atoms can enter, one can generate representations of large numbers of chemical compounds, and chemical reactions, using diagrams or models of molecular structures. Here one range of (relatively primitive) possibilities is used to explain another range.

This simple chemical theory had to be revised and refined of course, but that does not affect the point that at least part of its scientific function while it survived was to explain a range of possibilities according to criterion (1). (In AI research, a program can explain a range of possible behaviours. A derivation consists of running the program, or, preferably, reasoning about the program's capabilities.)

2.5.5. Criteria for comparing explanations of possibilities

However, there are additional requirements if **T** is to be a *good* explanation of the possibilities in question, or at least better than its rivals. Rival theories are assessed according to how well they meet these additional requirements, namely:

2. The theory **T** should be as *definite* as possible: that is, there should be a clear demarcation between what it does and what it does not explain. For instance, although early theories of sub-atomic structure definitely permitted an atom with one proton (hydrogen) to have zero or one neutrons, I doubt that they definitely permitted or ruled out the possibility of an isotope of hydrogen with one proton and, say, twenty neutrons, as more modern theories do.

3. **T** should be *general,* that is, it should explain many significantly different possibilities, preferably including some possibilities not known about before the theory was invented. This criterion should be used with caution. Insofar as a theory generates some possibilities not yet established by actual instances, efforts should be made to find or create instances. If repeated efforts to find actual instances fail, this does not disprove the theory, but it does reduce its credit. So a theory should not explain too many things.

4. **T** should account *for fine structure:* i.e. the descriptions or representations of possibilities generated by **T** should be rich and detailed. Thus a theory merely explaining the possibility of different chemical elements in terms of different possible constituents of their atoms will not be as good as one which also explains how it is possible for the elements listed on the periodic table to have exactly the similarities and differences of properties implied in the table.

5. **T** should be *non-circular,* i.e. the possibilities assumed in **T** should not be of essentially the same character as the possibilities **T** purports to explain. Many philosophical and psychological theories fail this test; computer-based models of human competence pass it, since assuming the possibility of a computer is quite different from assuming the possibility of a mind! However, notice that a kind of circularity, namely recursion, is possible *within* such an explanation. Behaviourist psychology is based on a failure to see this. (See chapter 1, section 3.)

6. The derivations from **T** should be *rigorous',* i.e. within the range of possibilities explained by **T**, the procedures by which those possibilities are deduced or derived should be explicitly specified so that they can be publicly assessed, and not left to the intuitions of individuals. If the theory is very complex, the only way to find out exactly what it does and does not imply (or explain) may be to express it in a computer program and observe the output in a range of test situations. (This takes the place of logical or mathematical deduction.) In fact rigour is

very rarely achieved, even in the physical sciences.

7. The theory **T** should be *plausible:* that is, insofar as it makes any assertions or has any presuppositions about what is the case or what is possible, these should not contradict any known facts. However, sometimes the development of a new theory may lead to the refutation of previously widely held beliefs, so this criterion has to be used with great discretion.

8. The theory should be *economical:* i.e. it should not include assumptions or concepts which are not required to explain the possibilities it is used to explain. Sometimes economy is taken to mean the use of relatively few concepts or assumptions, from which others can be derived as necessary. The latter is not always a good thing to stress, since great economy in primitive concepts can go along with uneconomical derivations and great difficulty of doing anything with the theory, that is, with *heuristic poverty*. For instance, the logicist basis for mathematics proposed by Frege, Russell and Whitehead is very economical in terms of primitive concepts, axioms, and inference rules, yet it is very difficult for a practising mathematician to think about deep mathematical problems if he expresses everything in terms of that basis, using no other concepts. Replacing numerical expressions by equivalents in the basic logical notation produces unmanageably complex formulae, and excessively long and unintelligible proofs. The main points get buried in a mass of detail, and so cannot easily be extracted for use in other contexts. More usual methods have greater heuristic power. So economy is not always a virtue. This is also true of Artificial Intelligence models.

9. The theory should be rich in *heuristic power:* i.e. the concepts, assumptions, symbolisms, and transformation procedures of the theory should be such as to make the detection of gaps and errors, the design of problem-solving strategies, the recognition of relevant evidence, and so on, easily manageable. This is a very difficult concept to define precisely, but it is not a subjective concept. The heuristic power of a theory may be a consequence of its logical structure, as people working in artificial intelligence have been forced to notice. (See chapter 7 and McCarthy and Hayes, 1969, for more on this.)

10. The theory should be *extendable* (compare Lakatos 1970). That is, it should be possible to embed the theory in an improved enlarged theory explaining more possibilities or more of the fine-structure of previously explained possibilities. For instance a theory explaining how people understand language, which cannot be combined with a perceptual theory to explain how people can talk about what they see, or use their eyes to check what they are told, is inferior to a linguistic theory which can be so extended. Extendability is a major criterion for assessing artificial intelligence models of human abilities. However, it is a criterion which can only be applied in retrospect, after further research attempting to extend the model or theory.

So a good explanation of a range of possibilities should be definite, general (but not too general), able to explain fine structure, non-circular, rigorous, plausible, economical, rich in heuristic power, and extendable.

2.5.6. Rational criticism of explanations of possibilities

These criteria indicate ways in which theories explaining possibilities may be criticised rationally. For instance, one may be able to show (by a logical or mathematical argument or by 'running' it on a computer) that the theory does not in fact generate the range of possibilities it is said to explain. (Nearly all psychological theories put forward to explain known human possibilities, such as perception, fail on this point: the theories generate the required range of possibilities only in the mind of a sympathetic audience supplying a large and unspecified set of additional assumptions.)

A theory explaining a range of possibilities may be criticised by showing that it explains too much, including things which so far appear to be impossible. The theory may not explain enough of the known fine structure of the possibilities (like theories of speech understanding which do not explain how hearers can cope with complex syntactic ambiguities, or developmental theories in biology which don't explain how a chicken's egg can grow into something like its mother or father in so many detailed ways).

The explanation may be circular, like theories which attempt to explain human mental functioning by assuming the existence of a spirit or soul with essentially all the abilities it is intended to explain.

The theory may be so indefinite that it is not clear what it does and what it does not explain.

A theory may also be criticised less directly by criticising the specification of the range of possibilities which it is meant to explain (e.g. criticising the typology on.which it is based). For instance the specification may describe a set of structures in ways which are not related to their functions, like describing sentences in terms of transition probabilities between successive words.

Or the set of possibilities explained may be shown to be only a sub-range of some wider set of possibilities which the theory cannot cope with. For instance, a theory which explains how *statements* are constructed and understood can be criticised if it cannot be extended to account for *questions, commands, threats, requests, promises, bets, contracts,* and other types of verbal communication which are clearly functionally related to statements in that they use related syntactic structures and almost the same vocabulary.

If it turns out that a physical theory of the interactions of atoms and their components can only explain the possibility of chemical reactions involving relatively simple molecules, then that will show an inadequacy in the theory.

Similarly, if an economic theory can explain only the possibility of economic processes occurring when there is a very restricted amount of information flow in a community, then that theory is not good enough.

Finally, if a philosophical theory of the function of moral language accounts only for abusive and exhortative uses of that kind of language, then it is clearly inadequate since moral language can be used in a much wider range of ways.

In some cases, whether a theory explaining some specified range of possibilities satisfies these criteria or not, or whether it satisfies them better than a rival theory, is not an empirical question. It is a question to be settled by conceptual, logical and mathematical investigations of the structure of the theory and of what can be derived from it.

Sometimes the theory is too complex for its properties to be exhaustively surveyed. If so, one can only try out various derivations or manipulations in test cases. This is partly analogous to an empirical investigation in that the results are always partial and cannot be worked out in advance by normal human reasoning. Similarly testing a complex computer program may feel like conducting some kind of experiment. Nevertheless, as already remarked, the connections so discovered are not empirical, but logical or mathematical in nature. (Compare Pylyshyn 1978, Sloman 1978.)

These criteria for assessing explanations of possibilities could be justified by showing how their use contributes to the interpretative and practical aims of science. They would also have to play a role in the design of an intelligent learning machine, along with the previously listed criteria for assessing concepts and symbolisms. So these criteria are relevant to developmental psychology and AI, as well as to the methodology of the physical sciences.

2.5.7. Prediction and control

A theory may meet the conditions listed above without being of any use in predicting or explaining particular events or in enabling events or processes to be controlled. This is why I have stressed the explanation *of possibilities*

Although it explains how certain sorts of phenomena are possible, the underlying mechanism or structure postulated may, at the time the theory is proposed, be unobservable, so that observation of its state cannot be used to predict actual occurrences of those phenomena. Similarly, no techniques may be available for manipulating the mechanisms, so that the theory provides no basis for controlling the phenomena.

For instance, the theory of evolution explains the possibility of a wide range of biological developments without providing a basis for predicting or controlling most of them.

Similarly, a theory explaining the possibility of my uttering sentences of particular forms need not provide any basis for predicting when I will utter any one sentence, or for making me utter it, or even for explaining exactly why I uttered the particular sentence I did utter at a particular time. This is because the theory may simply postulate a certain kind of sentence-generating mechanism, available in my mind as a resource to be used along with other resources. How any particular resource is used on any particular occasion, may be the result of myriad complex interactions between such factors as my purposes, preferences, hopes, fears and moral principles, what I believe to be the case at the time, what I know about the likely effects of various actions, how much I am distracted and so on. The theory which explains the possibility of generating and understanding sentences need not specify all the interactions between the postulated mechanism and other aspects of the mind. So it need not provide a basis for prediction and control.

This is true of any explanation of an ability, skill, talent, or power, in terms of a mechanism (e.g. a computer program) making it possible. The explanation need not specify the rest of the system of which that resource is a part, nor specify the conditions under which the resource is activated. And even if it does, the specification need not refer to either observable conditions or manipulable conditions. So such explanations of possibilities, though they contribute to scientific understanding, need not contribute to predictions of actual events.

I believe that the stress on predictive content derives from a misunderstanding of criteria 2 and 4, namely the requirement that the theory be definite and capable of explaining

2.5.8. Unfalsifiable scientifictheories

It is not possible to refute a scientific theory, if it merely explains possibilities, and entails or explains no impossibilities. For it is a fact about the logic of possibility that 'X is possible' does not entail 'X will occur at some time or other'. Similarly 'X never occurs' does not entail 'X is impossible'. Newtonian mechanics entails that it is possible for some very large body passing near the earth to deflect the earth from its orbit, and it explains this possibility: but the fact that this never occurs casts no doubt on the theory. Similarly, a grammatical theory may explain the possibility of the utterance of a certain rather complex English sentence, and even though nobody ever utters that sentence naturally, this casts no doubt on the theory. A psychological theory may imply that it is possible for a human being to count backwards from ninety-nine to one to the tune of 'Silent night, holy night', without being refuted merely by the fact that nobody ever does this. Only a much more complex theory, taking into account a rich set of motives and beliefs, could ever be used to predict such a performance, and perhaps be refuted by its non-occurrence.

Lack of predictive power, practical utility, or refutability need not rule out rational discussion of the scientific merits of an explanation of a range of possibilities. Neither should it rule out rational comparison with rival explanations, in accordance with the criteria listed above. Nor does it prevent such a theory from giving deep insight, of a kind which provides a firm basis for building more elaborate theories which do permit predictions and explanations of particular events, and which are empirically refutable.

I therefore see no reason for calling such theories nonsensical, as some of the logical positivists would, nor for banishing them from the realm of science into metaphysics or pseudo-science, as Popper does, (though he admits that metaphysical theories may be rationally discussable and may be a useful stimulus to the development ofwhat he calls scientific theories).

I am not here arguing over questions of meaning: I am not arguing about the definition of 'science'. My point is that among the major merits of the generally agreed most profound scientific theories is the fact that they satisfy the criteria for being good explanations of possibilities, and therefore give us good insights into the nature of the kinds of objects, events or processes that can exist or occur in the universe.

If unrefutable theories are to be dubbed 'metaphysical', then what I am saying is that even important scientific theories have a metaphysical component, and that the precision, generality, fine structure, non-circularity, rigour, plausibility, economy and heuristic power are among the objective criteria by which scientific and metaphysical theories are in fact often assessed (and should be assessed).

The development of such 'metaphysical' theories is so intimately bound up with the development of science that to insist on a demarcation is to make a trivial semantic point, of limited theoretical interest. Moreover, it has bad effects on the training of scientists. Since Artificial Intelligence produces unfalsifiable, but rationally criticisable, theories, it should undermine this harmful trend.

2.5.9. Empirical support for explanations of possibilities

Even though a theory which explains only possibilities is not refutable empirically, that does not mean that empirical evidence is wholly irrelevant to it. For instance, if a kind of possibility explained by the theory is observed for the first time after the theory was constructed, then this is empirical corroboration for the theory, even though the theory did not specify that the phenomenon ever would occur, or that it would occur in those particular conditions.

Observing an actual instance of a possibility explained by some theory provides support for that theory at least to the extent of showing that there is something for it to explain: it shows that the theory performs a scientific function. However, the support *adds* to previous knowledge only if it is a new kind of possibility. Mere repetition of observations or experiments does not increase support for a theory: it merely checks that no errors were made in previous instances.

In these contexts all the normal stress on repeatability of scientific experiments is unnecessary andhas misled some psychologists and social scientists into making impossible demands of empirical studies of man and society. Repetition may be a useful check on whether the phenomenon really is possible (since it permits more independent witnesses to observe it), and it provides opportunities for more detailed examination of exactly *what* occurred, but is not logically necessary.

Beethoven's compositions are unique. Yet it is a fact that it was possible for a human being to create them. That possibility requires explanation.

If a phenomenon occurs only once, then it is possible; and its possibility needs explaining. Any explanation of that possibility is therefore not gratuitous, and the only question that should then arise

is not whether the explanation is science or pseudo-science, or metaphysics, but whether it is the correct explanation. In practice, this becomes the question whether a *better* explanation can be found for the same possibility, that is, an explanation meeting more of the criteria (2) to (9) above; or perhaps serving additional scientific aims besides explaining possibilities.

The frantic pursuit of repeatability and statistically significant correlations is based on a belief that science is a search for laws. This can blind scientists to the need for careful description and analysis of what *can* occur, and for the explanation of its possibility.

Instead they try to find what *always* occurs a much harder task and usually fail. Even if something is actually done by very few persons, or only by one, that still shows that it is possible for a human being, and this possibility needs explanation as much as any other established fact. This justifies elaborate and detailed investigation and analysis of particular cases: a task often shirked because only laws and significant correlations are thought fit to be published. Social scientists have much to learn from historians and students of literature despite all the faults of the latter.

I have gone on at such great length about describing and explaining possibilities because the matter is not generally discussed in books on philosophy of science, or in courses for budding scientists. But I do not wish to deny the importance of trying to construct theories which can be used to explain and predict what actually occurs, or which explain impossibilities (laws) and observed regularities. Of two theories explaining the same range of possibilities, one which also explains more impossibilities and permits a wider variety of predictions and explanations of actual events to be made on the basis of observation, is to be preferred, since it serves to a greater degree the aims of science listed previously.[3]

This discussion is still very sketchy and unsatisfactory. Much finer description and classification of different sorts of explanations is required. But enough for now!

Part Six: Concluding remarks

2.6.1. Can this view of science be proved correct?

It is not possible to *prove* that this concern with possibilities is a major aim of science, for anyone can say that his concept of science is defined in terms of different aims. However, I invite the reader to reflect on examples of what he or she recognises to be major scientific achievements, and then to ask whether *one* of the criteria by which they are so recognised is not the extent to which they contributed to the stock of conceptual or representational tools available to scientists, or extended knowledge of what kinds of objects or events or processes could occur.

I suggest that anyone who tries this will discover, possibly to his surprise, that the scientific advances which he regards as most important include not only discoveries of new laws or regularities, or explanations thereof, but also discoveries or new types of phenomena, new explanations of ranges of possibilities, new concepts, new notations, and therein new means of asking questions about the world. For example, Boyle's discovery of his law relating pressure and volume of a gas, was not so profound as the prior invention of the concepts *of pressure* and *volume*. The search for laws presupposes the search for possibilities and their explanations, and this requires concepts and notations for representing possibilities.

For reasons which I do not fully understand, Popper is apparently strongly opposed to all this talk of concepts and possibilities. (See, for instance, pp. 123-4 of his (1972) where he describes it as an error to think that *concepts* and *conceptual systems* or problems about *meaning* are comparable in importance to *theories* and *theoretical systems,* or to problems of *truth.)* As far as I can tell, his

argument rests on the curious assumption that concepts or meanings are purely subjective things, and that only complete statements containing them can be assessed or criticised according to objective criteria. I hope I have said enough to refute this.

Roughly, our disagreement seems to hinge on Popper's view that the only place for rationality in science is in the selection from among hypotheses expressible in a given language, whereas I have tried to show that there are rational ways of deciding how to extend a language, and therefore how to extend the set of expressible hypotheses. I admit that there are still serious gaps in my discussion: a theory of concept-formation is still lacking.

Finally, even if it is agreed that science uses rational *means* to pursue the aims described here, the question arises: are these *aims* rational? Is it rational to pursue them? I believe there is no answer to this. If someone genuinely prefers the life of a mystic or hermit or 'primitive' tribesman to the pursuit of knowledge and understanding of the universe, then that preference must be respected. However, I believe that the aims and criteria described here are part of the mental mechanism with which every human child is born but for which it would not be possible to learn all that human children do learn. So one can reject science only after one has used it, however unconsciously, for some years.

Similarly, rational processes of concept formation and theory construction will have to be built into an intelligent robot if it is to be capable of matching the learning ability of young children. The development of science, the learning of a child, and the mechanisms necessary for an intelligent robot all involve computational processes, which build up and deploy knowledge of the form and contents of the world. This is one of several points at which bridges can be built between philosophy of science, developmental psychology, and artificial intelligence.

The attempt to build these bridges will provide good tests for the philosophical theories outlined here. It is certain that my theories will prove inadequate. But I hope they may provide a useful basis for further research.

Endnotes

[1] Some of the work on this paper was done during tenure of a visiting fellowship at the School of Artificial Intelligence, Edinburgh University. I am grateful to the Science Research Council and Prof. Bernard Meltzer for making this possible. Several colleagues have helped me by criticising drafts. P.M. Williams, L.A. Hollings and G.J. Krige in particular wrote at some length about by mistakes and omissions. This chapter is a modified and expanded version of a paper published in *Radical Philosophy* 13, Spring 1976.

[2] This is because the definition of the set entails that it contains itself if and only if it does not contain itself. (Note added: 2001. See also A. Botterell 'Conceiving what is not there', *Journal of Consciousness Studies* vol 8, no 8, pp 21--42, 2001.)

[3] Of course, it can always happen that a modified version of the inferior explanation will turn out to be better. Dead horses can come to life again in science.

Book contents page

Next: Chapter three

Last updated: 28 Jan 2007

Chapter 3

SCIENCE AND PHILOSOPHY

3.1. Introduction

Immanuel Kant's *Critique of Pure Reason* is widely acknowledged to have been a major contribution to philosophy. Yet much of the book can also be seen as an early contribution to theoretical psychology. For example, his claim that no experience is possible without some form of prior knowledge (a claim to which we shall return in the chapter on perception) is relevant to psychologists' attempts to understand the nature of perception and learning. His notion that perception and imagination require the use not of picture-like templates, but of rule-like schemata for analysing and synthesising images, has been re-invented by psychologists in this century and plays an important role in computer-based theories of perception.

So Kant's work illustrates the overlap between science and philosophy. There are many more examples. Einstein's approach to the analysis of concepts of space and time was influenced by his reading of empiricist philosophy. Frege's attempts to answer some of Kant's questions about the nature of arithmetical knowledge led him into logical and semantic theories and formalisms which have deeply influenced work in linguistics and computer science. Marx's sociological theories were partly based on Hegel's philosophy. More recently, work by philosophers of language, such as Austin and Grice, has been taken up and developed by linguists, and the psychologist Heider has acknowledged the influence of Ryle's *The Concept of Mind.*

Philosophers' analyses of some of our most general concepts, such as *cause, individual, action, purpose, event, process, good,* and *true,* are relevant to biology, to anthropology and developmental psychology, whether or not practitioners of these subjects are aware of this.

For instance, biologists studying the evolution of intelligence need to grasp what intelligence is, and how it includes the use of some or all of these concepts. A comprehensive anthropology would include cross-cultural studies of the most general and basic systems of concepts used by different peoples. And if developmental psychologists were to do their job properly they would spend a lot of time exploring such concepts in order to be able to ask deep questions about what children learn and how. (Piaget did this, to some extent. But I am not aware of university courses in developmental psychology which include training in conceptual analysis.)

Within artificial intelligence it is not possible to avoid philosophical analysis of such concepts, for the discipline of trying to design machines which actually behave intelligently and can communicate with us forces one into analysis of the preconditions of intelligent behaviour and our shared presuppositions. For otherwise the machines don't work!

These illustrations of the connections between philosophy and the scientific study of the world are not isolated exceptions. Rather, they are consequences of the fact that the aims and methods of philosophy overlap with those of science. In this chapter I shall try to analyse the extent of that

overlap.

3.2. The alms of philosophy and science overlap

In particular, the greatest philosophers have shared with the greatest scientists the first three 'interpretative' aims listed in chapter 2, namely the aim of developing good concepts or thinking tools, the aim of finding out what sorts of things are possible, and the aim of trying to explain these possibilities. Their methods of pursuing these aims have much in common too, as will be shown below.

A fourth major aim that they appear to have in common is the aim of discovering limits to what is possible, and explaining such limits. However, in relation to this aim, the methods of scientists and philosophers tend to be rather different, insofar as philosophers often try to set up non-empirical demonstrations. And they usually fail.

By exploring the relationship between the aims and methods of science and philosophy we shall explain how it is possible for philosophy to be the mother of science, thereby perhaps making a philosophical contribution to the science of intellectual history.

Let us start with some reminders of the kinds of questions which have exercised philosophers. I shall ignore the many pseudo-questions posed by incompetent philosophers who cannot tell the difference between profundity and obscurity.

3.3. Philosophical problems of the form 'How is X possible?'

Many questions of the form 'How is X possible?' have been asked by philosophers. Some of them go back to the ancient Greek philosophers, or further, while others were first formulated much more recently. It will be seen from the long list of examples which follows that more and less specific versions of the same question can be asked. Detailed analysis in philosophy or science leads to the formulation of more specific questions, concerned with more of the fine-structure of X. Increasing specificity increases the scientific interest of the question. In particular, it should be clear that although the first question listed below is a philosophical one, more specific versions of it (e.g. the fourth one) look much more like psychological questions.

1. How is knowledge possible?
2. How is empirical knowledge possible?
3. How is it possible to acquire knowledge about the material world on the basis of sensory experience?
4. How is it possible to learn, from seen two-dimensional surfaces, that an object is three-dimensional and has unseen surfaces on the far side?
5. How is it possible to know anything about the past, the future, unobserved objects or processes, or other people's minds? (Cf. 16).
6. How is it possible to know that two events are causally connected?
7. How is it possible to know laws of nature or their explanations?
8. How is it possible to know truths of logic and mathematics?
9. How is it possible to know conditional truths, of the forms If P then Q' or If P had been the case then Q would have been'?
10. How is it possible for an infant knowing no language to learn one?
11. How is it possible to learn the language of a culture other than one's own?
12. How is it possible to use strings of symbols to describe our multi-dimensional world?
13. How is it possible to interpret flat patterns as pictures of solid three-dimensional objects?

(Compare question 4.)

14. How is it possible to use general concepts, such as *dog, triangle, game, taller,* or *between,* which apply to a very varied range of instances?
15. How is it possible to learn the names of, think about, or refer to, remote persons, places or events?
16. How is it possible to think about the past or future events? (Cf. 5.)
17. How is it possible to think or talk about non-existent things, such as Mr. Pickwick, Ruritania, the accident that nearly happenedthis morning, or the choice I considered but did not make?
18. How is it possible to think about minds other than one's own, or about another person's emotions or sensations?
19. How is it possible to have idealised concepts which go beyond the limits of what we can experience, such as *perfectly thin, perfectly straight, perfectly parallel, exactly the same shade of colour, exactly the same weight,* or *exactly the same shape"*
20. How is it possible for a finite mind to think about such infinite totalities as the set of integers, the set of points on a line, or the set of all possible English sentences?
21. How is it possible to have a concept of a causal connection which is more than the concept of an instance of a well-confirmed regularity?
22. How is it possible to understand scientific theories referring to things which can never be perceived?
23. How is it possible to understand metaphors?
24. How is it possible to understand metaphysical questions?
25. How is it possible for a person, or a culture, to discover that its conceptual system is inadequate, and improve it?
26. How is it possible for there to be valid reasoning which is not purely logical, such as inductive reasoning or reasoning using diagrams?
27. How is it possible for an identity-statement, such as The Evening Star is the Morning Star', to be true, yet have a different significance from another identity statement referring to the same thing, such as 'The Morning Star is the Morning Star'?
28. How is it possible for two predicates, such as 'polygon with three sides' and 'polygon with three angles', to describe exactly the sameset of objects yet have different meanings?
29. How is it possible for there to be formal, or syntactic, tests for valid (truth-preserving) reasoning?
30. How is it possible to have knowledge which one can use yet not formulate (e.g. knowledge of how one's native language works, or knowledge of the difference between Beethoven's and Schubert's styles of composition)?
31. How is it possible for there to be minds in a material universe?
32. How is it possible for physical and chemical processes to influence or even give rise to such things as sensations and feelings? (or vice versa?)
33. How is it possible for there to be such a thing as self-deceit, or unconscious beliefs, attitudes, desires, fears, or inferences?
34. How is it possible for a set of experiences, either at the same time or at different times, to be the experiences of one mind?
35. How is it possible for a set of experiences, beliefs, thoughts, decisions, intentions, and actions all to 'belong' to one mind?
36. How is it possible for deliberation, choice, or decision to exist in a deterministic universe?
37. How is it possible for a mind to continue to exist while unconscious?
38. How is it possible to think of oneself as being in a world whose existence is independent of one's own (or any mind's) existence?
39. How is it possible to distinguish moral or aesthetic evaluations from personal likes or dislikes,

or to think rationally about moral problems?
40. How is it possible to use moral language of a kind which does not reduce to descriptive or emotive language?
41. How is it possible for a norm to exist in a community without being accepted by any individual in the community?
42. How is it possible for democracy to exist in a state with millions of citizens with different and conflicting aims?
43. How is it possible for a social system to be just?
44. How is it possible rationally to weigh up short term and long term harm and benefit?
45. How is it possible to search in a sensible direction for the solution to a problem without knowing what form the solution will take?
46. How is it possible for an event to be temporally related to another distant event?
47. How is it possible to identify and reidentify places?
48. How is it possible for objects to change their properties and relationships (and remain the same objects)?
49. How is it possible for there to be anything at all?
50. How is it possible for people to invent philosophical problems?

Many of the questions in the list have controversial presuppositions: it is often disputable whether the X in 'How is X possible?' is possible at all! Many attempts have been made to prove the impossibility of some X, for instance where X = meaningful talk about God or infinite sets, or rational discussion of moral issues, or even such obviously possible things as: change, a man over-taking a tortoise in a race, knowledge about the past, knowledge about material objects, or deliberation and choice.

Lunatic though it may at first appear, serious thinkers have put forward demonstrations that these are impossible. Equally serious thinkers have put great intellectual effort into attempts to refute such demonstrations. The process may appear a waste of time, but has in fact been very important. The discovery, analysis and, in some cases, refutation of such paradoxical proofs of impossibility has been a major, though haphazard, stimulus to philosophical progress and the growth of human self consciousness. It leads to a deeper understanding of the phenomenon whose possibility is in dispute. In some cases (e.g. Zeno's paradoxes) it has even led to advances in mathematics.

Often, a philosopher asks 'How is X possible?' only in the context of asking 'What is the flaw in so and so's alleged proof that X is impossible?' But there is also a more constructive philosophical tradition, first consciously acknowledged by Immanuel Kant, of granting that X is possible and attempting to explain how it is, in the light of careful analysis of what X is. This is the philosophical activity which merges into scientific theorising.

In what follows I'll try to analyse the similarities and differences in aims and methods: a step towards a scientific theory explaining the possibility of the growth of scientific and philosophical knowledge.

3.4. Some general similarities and differences between science and philosophy

One of the differences between science and philosophy concerns the range of possibilities attended to. Philosophers have mostly been concerned with possibilities which are known to everyone or at least to educated laymen in their community, whereas one of the characteristics of scientific research is that it uses sophisticated apparatus and techniques, and highly specialised explorations, in order to discover new possibilities which are not discoverable simply by reflection on common experience.

I do not mean that all the possibilities discussed by philosophers are obvious: they may be known to all of us without our realising that we know them (like the possibility of truly unselfish action). Some of the things we know are not evident to us until we have engaged in the philosophical activity of

digging up unacknowledged presuppositions. For instance, most people if simply asked how many different kinds of uses of language there are, are likely to come up with only three or four, such as the text-book favourites: exclamations, questions, commands and assertions (statements). But even though they do not think of more without prodding, they do in fact know of many possible uses of language not covered by this list, such as betting, congratulating, pleading, exhorting, warning, threatening, promising, consoling, reciting, calling someone, naming someone or something, welcoming, vowing, counting, challenging, apologising, teasing, declaring a meeting open or closed, and several more. (See J.L. Austin, *How to Do Things With Words*.)

Similarly, there are many psychological possibilities which we all know about, but do not find it easy to recall and characterise accurately when theorising about the mind. I shall draw attention to many examples in later chapters. So, both philosophy and science use specialised techniques to find out what sorts of things are possible, but their techniques and consequently the ranges of possibilities unearthed, are different. Philosophers dig up what we all know, whereas scientists mainly to extend what we know, about possibilities.

In both cases a preliminary characterisation of a kind of possibility is subject to correction, in the light of an explanatory theory.

One of the faults of philosophers is that they tend to ask questions which are not nearly specific enough. If one simply asks 'How is knowledge possible?' or 'How is knowledge of other minds possible?', these questions do not explicitly specify the requirements to be met by explanatory answers, since they do not describe in sufficient detail what is to be explained. They specify many requirements implicitly, because we all know a great deal about the possibilities referred to, but until they have been described explicitly, people can unwittingly select different subsets for consideration, and so philosophical debates often go on endlessly and fruitlessly.

The criteria listed in Chapter 2 for assessing explanations of possibilities, presuppose that there are detailed specifications of the range of possibilities to be explained. Otherwise there is no agreed basis for assessing and comparing rival theories. This preliminary analysis of the range of possibilities to be explained is often shirked by philosophers.

Even when philosophers do a fairly deep analysis, it is not presented in a systematic and organised form but rather in the form used for literary essays. The result is that philosophers often simply talk past each other. (This also happens in psychology for similar reasons, as may be confirmed by looking at the cursory 'definitions' of mental concepts such as *emotion, memory .perception, learning,* etc., which precede lengthy chapters on empirical results and proposed theories.)

In both philosophy and science, if progress is to be made, and seen to be made, the task of constructing an explanation of the possibility of X must be preceded by at least a preliminary characterisation of the range of possible kinds of X. This preliminary characterisation may be based on close examination of a wide variety of examples of X, taken from common experience, in the case of philosophy, or from specialised experiment and observation. The specification may include such things as the types of components, the types of organisation of those components, the types of behaviour, the types of functions, and the types of relations to other things, found in specimens of X, i.e. internal and external structures, functions and relations. In both philosophy and science, the construction of an explanatory theory will suggest ways of improving or correcting such 'observations'.

Having got a preliminary characterisation, that is, a preliminary answer to the question: What sort of things are X's? or What sort of X's are possible?, the scientist or philosopher can then begin to construct a theory describing or representing conditions sufficient to generate the possibility of instances of X (knowledge, perception, truth, scientific progress, change, falling objects, chemical

processes, or whatever it is whose possibility is to be explained). Whether one is a philosopher or a scientist, the conditions for adequacy of an explanatory theory, and the criteria for comparing the merits of rival explanations of a range of possibilities are the same, namely the sorts of criteria listed in chapter 2.

Despite the overlap, there is an important difference. Often philosophers are content to find some theoretically adequate explanation of a set of possibilities without bothering too much whether it is the *correct* explanation. So they ask 'How *might* X be possible?' rather than 'How *is* X possible?', or 'What *could* explain the possibility of X?' rather than 'What *does* explain the possibility of X?' However, every answer to the latter necessarily includes an answer to the former, and in that way science subsumes philosophy, which is very like the relationship between A.I. and psychology (see chapter I). In practice, the difference between the two approaches becomes significant only when alternative answers to the first question have been formulated, so that something can be done to find out which is a better answer to the second.

3.5. Transcendental deductions

When one has such a theory T explaining the possibility of X's the truth of T *is* a *sufficient* condition for the possibility of X. However, it may not be the *correct* explanation, for instance if T itself is false. In general it is not possible, either in science or in philosophy, to establish conclusively that some theory is true: the most one can do is determine which, if any, of several theories is, for the time being, best. And even that is not always possible when a subject is in its infancy.

However, some philosophers have not been satisfied with this, and have tried to show that no other theory besides their own could possibly give the correct explanation. An argument purporting to show that T is not merely *sufficient* to explain the possibility of X, but also *necessary,* is called a ' transcendental argument'. (As far as I know, this notion was invented first by Kant.)

No attempts to construct valid transcendental arguments have so far been successful. For instance, Kant tried to show (in *Critique of Pure Reason)* that explaining the possibility of distinguishing the objective time order of events from the order in which they are experienced must necessarily involve assuming that every event has a cause; but quantum physics shows that one can get along without assuming that every event has a cause. Strawson tried to show (in *Individuals)* that our ability to identify and re-identify material objects and persons was a necessary part of any explanation of the possibility of identifying other things such as events, processes, states of affairs, pains, decisions, and other mental phenomena.

But he made no attempt to survey all the possible theories which might one day be formulated, including the varieties of ways in which computers or robots (and therefore people) might be programmed to use language, and his arguments seem to be irrelevant to the detailed problems of designing mechanisms with the ability to refer to and talk about things. (This criticism requires further elaboration.)

Such attempts at transcendental deductions are over-ambitious, for to prove that some theory T is a *necessary* part of any explanation of the possibility of X would require some kind of survey of all possible relevant theories, including those using concepts, notations and inference procedures not yet developed. It is hard to imagine how anyone could achieve this, in science or in philosophy. Scientists rarely try: They are not as rash as philosophers.

One reason why philosophers feel they must bolster up their explanations with 'transcendental arguments' is that they dare not admit that philosophy can be concerned with empirically testable theories, so they try to show that their theories are immune from empirical criticism. However, I shall

show below that this is inconsistent with the practice of philosophers.

We now look a little more closely at similarities and differences between methods of science and philosophy.

3.6. How methods of philosophy can merge into those of science

The procedures by which philosophy can make progress in the task of describing and explaining possibilities shade naturally into scientific procedures. So by describing such philosophical procedures and the processes by which they transform a problem, we begin to explain how it is possible for philosophy to contribute to science. The overlap with AI (when AI is done well) is specially significant.

The relevant philosophical procedures concern the following:

a. Collection of information about what sorts of things are possible,
b. Construction of new characterisations or representations of those possibilities (i.e. answers to the question 'What is X?'),
c. Construction of explanations of those possibilities, and finally testing and refinement of explanatory theories. This last step can, as in all science, lead back to modifications of earlier steps.

A first step is collecting information about the range of possibilities to be explained. For instance, before attempting to explain the possibility of knowledge one must ask 'What is knowledge?'. This involves collecting examples of familiar kinds of knowledge, and classifying them in some way. (Knowledge of particular facts, knowledge of generalisations, knowledge of individuals, knowing how to do things, etc.) Closely related possibilities should also be surveyed, e.g. believing, learning, inferring, proving, forgetting, remembering, understanding, doubting, wondering whether, guessing, etc. Functions of knowledge can then be listed and classified.

All this gives a preliminary specification of some of the *fine structure* of the range of possibilities to be explained, an answer to the question 'What is X?' (or, 'What are X's?). One can go on indefinitely attempting to improve on the preliminary specification, by covering a wider range of cases, giving more detailed specifications of each, and revising the classification.

This process may at first rely only on what Wittgenstein (in *Philosophical Investigations,* Part I, section 127) called 'assembling reminders'. These are examples of possibilities which when stated are obvious to common-sense, since we have all experienced similar cases, though we may not find them easy to think of on demand, like the examples of possible uses of language noted above. Much analytical philosophy, and most of Wittgenstein's later philosophy, consists of this kind of common-sense natural history.

An obvious extension of this activity is the use of experiments, instruments, measurement, fieldwork, and other tools of science to find and describe new examples of X, or new facts about old examples. Chapter 1 explained how artificial intelligence can contribute to this fact-gathering process in philosophy by providing examples of new forms of behaviour.

So the fact-collecting of philosophers merges into the fact-collecting of scientists. However, new empirical research may be premature if common sense knowledge about possible sorts of X's has not yet been made explicit and systematised. (Hence the futility of much psychological research, e.g. on decisions, learning and emotions.) So philosophical methods of analysis should come first in cases where relevant information is part of common sense for instance in the study of mind and society. (Some linguists have appreciated this, but few psychologists or social scientists. Fritz Heider was a notable exception: see his *Psychology of Interpersonal Relations.)*

In philosophy, as in science, fact collection is rarely useful unless guided by a problem or explanatory theory. The mere collection of possibilities is of little interest except insofar as a theory can be found to explain and organise them. And theories are important only if they help us solve problems or puzzles. How theories are generated is still largely an unsolved problem. No doubt chance plays a role, but individuals like Kant, Einstein and Newton would not have made so many theoretical advances if they had not employed (albeit unconsciously) rational procedures for making the best of chances which came their way.

Artificial Intelligence in its attempts to design intelligent (i.e.rational?) learning planning and problem-solving systems necessarily overlaps with philosophical attempts to explain the nature of theories and theory formation (asoutlined in Chapter 2)

3.7. Testing theories

Once a theory T has been found which meets some or all of the criteria listed in the previous chapter (see sections 2.5.4-6) for explaining the possibility of X's, the question arises whether it is the *correct* explanation. Whether in philosophy or in science, answering this question requires testing the theory on new examples of X, or new, more detailed, descriptions of old examples, in order to see whether it is sufficiently general and explains enough fine structure. The theory can also be related to other known facts to see whether it is inconsistent with them and therefore false: i.e. its plausibility can be tested.

Emotivism is a philosophical theory purporting to explain how it is possible to use moral language meaningfully. However, fact-collecting of the sort described above showed the theory to be insufficiently general, for it was unable to account for facts about moral language which were not at first obvious to proponents of the theory, but are part of common sense. For instance, the theory interpreted moral language as performing functions like expressing the speaker's emotions, evoking similar emotions in hearers and causing hearers to act in certain ways. This fails to account for the empirically established possibility of unemotional hypothetical discussion among rational people of what, morally, ought to be done in certain situations. So the theory must either be rejected or modified to deal with this use of moral language. (I have listed a range of facts which theories like emotivism cannot account for, and proposed an alternative theory, in my two papers on 'better': see bibliography.)

This example refutes the widespread assumption that philosophical theories are not empirically testable. The assumption is probably based on a misconstrual of what philosophers actually do when they use empirical facts to test or support their theories: they use widely known common sense possibilities, rather than facts based on specialised empirical investigation. So the work can be done in an armchair no laboratory is needed, nor fieldwork. (The situation is similar when a linguist investigates his own language.) Because the information is so readily available its *empirical* nature is not recognised. (R.M. Hare made related points in his 'Philosophical Discoveries').

However, when the stock of relevant possibilities available to common sense is exhausted and has to be extended by more specialised empirical investigations, then philosophy merges into science. For instance philosophical investigations of the function of moral language and attempts to explain its possibility should, if properly conducted, overlap with linguistics and the psychology and sociology of morals. (Equally, the psychology and sociology, if done properly, would start with philosophical analysis of known possibilities.) For another example of philosophical use of empirical facts, this time from cognitive anthropology, see Bernard Harrison, *Form and Content.*

3.8. The regress of explanations

When a philosopher constructs his theory T, to explain a certain range of possibilities, it will not be long before someone asks for an explanation of the possibilities assumed in T. This may also lead towards scientific theorising and testing. For instance, Emotivism assumes (correctly) that it is possible for people to influence one another's actions and emotions by talking, and uses this to explain (wrongly) how moral language is possible. But the assumed possibilities also need explaining: and this leads directly into scientific studies of language and mind, e.g. studies of how utterances can influence attitudes.

Similarly, philosophers have often tried to explain the possibility of knowledge on the assumption that it is possible for things to be learnt from experience, and in particular that it is possible for ideas to become 'associated' with one another. But these assumed possibilities also need explaining, and this leads directly into scientific studies (in artificial intelligence and psychology) of ways in which information can be acquired and stored so as to be available for future use, and so as to enable one piece of information to 'evoke' another (which invdves tricky problems ofindexing and retrieval).

3.9. The role of formalisation

As specifications of phenomena to be explained become more detailed and cover a wider range of possibilities, so that more complex constraints have to be satisfied by the explanatory theory, it becomes necessary to invent special symbolisms and technical concepts in order to formulate theories which are sufficiently rich, powerful and precise.

In this way philosophy sometimes becomes more mathematical, as can be seen especially in the case of logic but also in philosophical studies of probability, in philosophy of science, and even in some branches of moral philosophy. Increasingly the formalisms of Artificial Intelligence will be used, as philosophical theories become more complex and precise, and too intricate to be evaluated without the aid of a computer. This parallels the ways in which scientific theories become more and more mathematical.

For instance, if, instead of the usual vague and general philosophical discussions of how perception can yield knowledge, an explanation is required *ofhow specific sorts* of perceptual experiences can yield knowledge *of specific sorts* of spatial structures, for instance an explanation of how certain views of a cube enable one to see that it is a cube with an interior and with faces on the far side, etc., then a mathematical formulation is inevitable. (N.B. 'Mathematical' does not mean 'numerical' or 'quantitative'.)

University courses in philosophy will need substantial revision if the appropriate theory-building and theory-testing skills are to be taught.

3.10. Conceptual developments in philosophy

In philosophy as in science, attempting to explain things can lead to new ways of looking at or thinking about the old facts, and this requires new sets of concepts. For example, the development of philosophical theories explaining the possibility of various uses of language can lead to criticism of old metalinguistic concepts or invention of new ones. Examples are: Kant's distinction between 'a priori' and 'analytic'; Frege's rejection of the subject/predicate distinction in favour of the function/argument distinction for describing sentence structures; the rejection by J.L. Austin and others of a four-fold classification of sentences into statements, questions, exclamations and imperatives; the discovery (explained, for instance, by J. Kovesi in his *Moral Notions)* that 'evaluative' is not a suitable label for the kinds of uses of language which have attracted attention in

moral philosophy and aesthetics; modern criticisms of Kant's distinction between analytic and synthetic statements; and Kuhn's attempt to replace the concept 'scientific theory' with 'paradigm'.

My own attempt (in chapter 7) to replace crude distinctions between verbal and nonverbal symbolisms and reasoning processes with more precise distinctions is another example. My use of the concept of 'explaining how Xis possible' is another. Further examples will be found in the chapter on numbers (chapter 8).

New concepts can change our view of what it is that we are trying to explain, so that a new specification is given of the old possibilities. Similar processes in the history of science have been described by Kuhn (1962, pp. 129-134), such as the change in the boundary between the concepts 'chemical compound' and 'physical mixture' resulting from the atomic theory of chemical composition.

In philosophy and in science, conceptual changes generate new specifications of what needs to be explained, and so can lead to new theories. The process of growth of human knowledge seems to be full of 'feed back' loops.

3.11. The limits of possibilities

I have said a lot about overlaps between aims of philosophy and the first three aims of science, namely the discovery, description, and explanation of possibilities. But science attempts also to find limits of possibilities: laws of nature. Is there a counterpart in philosophy?

Certainly some philosophers have tried to show not merely how things are but also how they must be or cannot be. Empiricists try to show that all significant knowledge *must* be based on sensory experience. Rationalists try to show that certain important kinds of knowledge *cannot* be empirical. Dualists try to show that there *must* be more than a material world if consciousness as we know it is possible. Logicists try to argue that mathematical concepts *must* be definable in terms of logic, if they are to have their normal uses. Moral or political philosophers often try to argue that their own moral or political principles must be accepted if morality or society is to be possible at all. Such theses are often based on attempts at 'transcendental arguments', which I have already criticised as over-ambitious, in the discussion of Kant, above.

Kant claimed to have unearthed various laws and principles which were part of the fundamental constitution of the human mind, so that all human thought and experience necessarily had to conform to them. However, such claims are very rash, in view of the fact that both biological and cultural evolution are known to be possible. We have already seen that thoughts that were impossible for ancient scientists are possible for modern scientists. The same contrast can be made between children and adults. This suggests that insofar as human minds have a 'form' limiting the nature of the world they experience, this form can vary from culture to culture and from time to time in one culture or even in one person, or robot.

The same is probably true of forms of language, society, morals, religion and science. If there are limits to this variation, they will have to be found by scientific investigations, not introspection or philosophical argument. The limits can hardly be studied before the mechanisms of individual and social development are understood, however. We must not try to fly before we can walk, even if we are philosophers.

However, there are many more mundane kinds of limits of possibility which philosophers characteristically attend to in their attempts to analyse familiar concepts. For instance, it is impossible for someone to be a spinster and married; it is impossible to admire someone for his honesty and breadth of knowledge yet never believe a word he says; it is impossible to be interested in botany yet

never wish to look at or learn anything about plants; it is impossible to be intensely angry with someone yet not believe that person has done anything you dislike or disapprove of; it is impossible to drive a car with care and recklessly at the same time (though it is possible carefully to drive over a cliff, to commit suicide). These are not laws' limiting what is possible in the world. Rather, they express defining conditions, or logical consequences of defining conditions, for the use of our concepts. Kant called such propositions 'analytic'.

Making such 'definitional' necessities and impossibilities explicit is part of the task of analysing how our concepts work. This in turn is a useful means of drawing attention to the presuppositions we all make about what sorts of things are possible in the world, and about useful ways of sub-dividing these possibilities. Looking at such subtle differences as the difference between 'with care' and 'carefully' (which are different since they have different boundaries) we learn to articulate our implicit common-sense knowledge about possible configurations of human beliefs, motives, decisions and actions. This is a contribution of philosophy to psychology and AI. (See chapter 4 for more on this.)

The role of necessities and impossibilities in philosophy is a large topic, and I have by no means exhausted it. All I wanted to show here is that the scientific aim of discovering limits of what is possible in the world is not an aim philosophers can or should share unless they are prepared to go beyond philosophical argument.

However, it is important for philosophers to expose present limits of our conceptual and representational apparatus often as a first step towards overcoming those limits. I am trying to expose, and remove, limits of our normal ways of thinking about philosophy and science.

3.12. Philosophy and technology

A theory which explains old possibilities may have surprising new implications. Technology includes the use of ingenuity to invent previously unthought of possibilities which can be explained by available theories. But this is also a major part of pure science, as when the kinetic theory of heat explained the possibility of a lowest temperature and the theory of relativity was used to demonstrate and explain the previously unsuspected possibility of conversion of mass into energy, and of the bending of light by gravitation. The realisation of such new kinds of possibilities in suitable experimental situations can provide dramatic new support for the theories which explain them. So can new ways of realising old possibilities. Philosophical theorising can also lead to the invention of possibilities previously unthought of and possible new means of realising previously thought of situations. So philosophy, like science, has its technological application.

For instance, philosophers have tried to use theories of language to show the possibility of logical languages which in one respect or another (e.g. precision, clarity, economy of rules) improve on natural language, or social theories to demonstrate the possibility of improving on existing social structures, or epistemological theories to demonstrate the possibility of improving on prevailing standards of rigour in science or mathematics. Similarly there is a technological theme to this book, insofar as it uses a theory of the relation between philosophy and science in an attempt to show the possibility of new types of collaboration between philosophers and scientists who study man, or engineers who try to design intelligent machines.

3.13. Laws in philosophy and the human sciences

I have tried to show that philosophy and science have overlapping interests, and partially similar methodologies, and that philosophy can generate science. The affinities between science and philosophy seem to be strongest in the case of the sciences which study man. For it is unlikely that

these sciences will, in the foreseeable future, go beyond theories which describe and explain possibilities (the things people and social systems can do).

It seems very unlikely that they will discover new laws with predictive content and explanations of those laws, apart from such trivial laws as are based on common sense, such as the law' that no normal person in our culture calmly invites a total stranger to chop his leg off! Some alleged laws are very likely to be culture-bound regularities, modifiable by training, propaganda, or economic pressures. Other apparent laws 'discovered' by empirical research are in fact just disguised tautologies, true by definition, for instance: 'Other things being equal, people tend to choose alternatives which they believe will bring about what they desire most'; or 'Persons are more likely to believe a statement if it is made by someone they respect, other things being equal'.

But the lack of substantial laws does not leave the human sciences without content, for there are many kinds of social and psychological phenomena whose *possibility* is well known and needs to be explained, even though the prediction and explanation *of particular instances* is out of the question, since it depends enormously on highly complex individual histories, decision-strategies, beliefs, interests, hopes, fears, ways of looking at things, and so on.

To turn to the search for *probabilistic* or *statistical* laws, when the hope of *universal* laws has been abandoned, as so often happens, is to reject the opportunity to study and interpret the rich structure of particular cases as a way of finding out what possibilities they exemplify.

Insofar as there are laws and regularities to be discerned among all the idiosyncracies of human behaviour, they can hardly be understood and explained before the possibilities they limit have been described and explained. Outside novels, there are so far few, if any, rich and systematic descriptions or explanations of human possibilities, so the human sciences will need to join forces with philosophy in the study of possibilities for some time yet.

3.14. The contribution of Artificial Intelligence

But not only with philosophy, for in the new discipline of artificial intelligence theories are emerging, in the form of specifications for computer programs, which, for the first time, begin to approach the complexity and generative power needed for the description and explanation of intelligent behaviour while also accounting for immense individual differences (as pointed out by Clowes, in 'Man the creative machine').

When such theories are embedded in computers and shown by the behaviour of the computer actually to work, then this establishes that they do not rest on presuppositions of the type they are trying to explain. (However, at present, A.I. models explain only a very tiny fragment of what needs to be explained.)

It may turn out that the combination of skills and knowledge required to construct non-circular and rigorous explanations of any significant range of human possibilities cannot exist in any one scientist nor in any team of scientists, philosophers, and engineers, small enough to co-operate. Human possibilities may be too complex to be understood and explained by humans. But the time is not yet ripe for drawing this pessimistic conclusion, and even if it is true, that is no reason for not trying.

3.15. Conclusion

The best way to make substantial new progress with old philosophical problems about mind and body, about perception, knowledge, language, logic, mathematics, science and aesthetics, is to reformulate them in the context of an attempt to explain the possibility of a mind. The best way to do this is to attempt to *design* a working mind, i.e. a mechanism which can perceive, think, remember,

learn, solve problems, interpret symbols or representations, use language, act on the basis of multiple motives, and so on.

Computers cannot yet do these things in a way which compares with humans, and perhaps they never will. But computer programs provide the only currently available language for formulating rigorous and testable theories about such processes. And only with the aid of computers can we explore the power of really complex and intricate theories. (Part two of this book elaborates on the kind of complexity involved.) So I conclude that in order to make real advances in problem areas mentioned above, philosophers, like psychologists and linguists, will need to learn about developments in the design of computing systems, programming languages and artificial intelligence models, even if they do not write programs themselves.

The ('meta-level') concepts *used for describing* computing systems, programming languages, hardware and software architectures, etc. are as important as, or perhaps even more important than the concepts *used in* programming languages.

The attempt to design a mind is a very long term research enterprise. I expect that it will provide the best illustration of the overlap between science and philosophy.

Book contents page

Next: Chapter four

Last updated: 28 Jan 2007

CHAPTER 4

WHAT IS CONCEPTUAL ANALYSIS?

4.1. Introduction

Elsewhere in this book, I have frequently referred to an activity of philosophers known as conceptual analysis. This has been practised in various forms and for various purposes by a wide range of philosophers and scientists. It has been particularly associated with mid-twentieth-century philosophy in Oxford and Cambridge, for instance the work of L. Wittgenstein, J. Wisdom, J.L. Austin and G. Ryle. As I see it, the main difference between these and earlier philosophers is that the latter were somewhat less self-conscious about the activity. However, on the whole recent analysts agree with previous philosophers that the main function of conceptual analysis is to help clarify or resolve philosophical problems, and occasionally also to provide a basis for criticising some uses of language. For example, in *A Plea for Excuses* Austin claimed that the analysis of the concepts *reason, excuse, justification,* and related concepts would not only help to clarify philosophical problems about freedom but also show some errors in the utterances of judges and in writings on jurisprudence.

I have tried to suggest that, besides these uses, conceptual analysis has another important purpose, namely to find out things about people and the world. However, this requires a far more disciplined and systematic approach to the analysis of concepts than is to be found in the work of most philosophers. (This is partly because their goals are different.)

We have a very rich and subtle collection of concepts for talking about mental states and processes and social interactions, including: *abdicate, abhor, acquiesce, adultery, adore, admire, angry, astonish, attend* and *avid,* to mention a few.

These have evolved over thousands of years, and they are learnt and tested by individuals in the course of putting them to practical use, in interacting with other people, understanding gossip, making sense of behaviour, and even in organising their own thoughts and actions.

All concepts are theory-laden, and the same is true of these concepts. In using them we are unwittingly making use of elaborate theories about language, mind and society. The concepts could not be used so successfully in intricate inter-personal processes if they were not based on substantially true theories. So by analysing the concepts, we may hope to learn a great deal about the human mind and about our own society. This point does not seem to be widely understood: this is why so many people (including many philosophy students) dismiss conceptual analysis as being 'merely concerned with meanings of words'.

Most of the theoretical presuppositions of our ordinary concepts are not concerned with laws or regularities, but with possibilities. For example, the use of a concept like *careful* is based on our knowledge that people can act in certain ways, not on any laws about how they always or usually act. The chapter on the mechanism of mind outlines some results of my own attempts to analyse familiar concepts concerned with actions and related mental processes. These analyses revealed a host of

human possibilities, and the mechanism sketched in that chapter is intended to provide the beginnings of an explanation of those possibilities, showing how conceptual analysis can contribute to psychology and artificial intelligence.

Similarly, by analysing concepts related to space and physical motion, e.g. *bigger, longer, inside, push, pull, carry, fetch, throw, impede, collide,* and so on, we may expose some unarticulated theories about our physical environment which govern much of our thought and behaviour. This task is not so urgent because physics and geometry have already made a great deal of progress, often going beyond our common-sense theories. To some extent this has been a result of conceptual analysis: the most striking example being Einstein's analysis of concepts of space and time. However, further conceptual analysis is required for improving our understanding not of the physical world itself, but of how people of various ages and cultures think about the world (consciously and unconsciously). Intelligent machines may need to think of the world as ordinary people do, rather than as quantum physicists do. [Note added: 2001. The recent growth of interest in the study of ontologies in AI and software engineering illustrates this point.]

It has been easier to make substantial progress in the physical sciences partly because the physical world is much simpler than the world of mental and social processes. Moreover, our interactions with the physical world are not as rich as our interactions with people so there is more scope for commonsense to have evolved mistaken theories about matter.

In the rest of this chapter, I shall try to list some of the methods which are useful in analysing concepts. Most of this will be familiar to analytic philosophers, especially those who have studied the work of Austin and Wittgenstein. However,

I have found that the techniques are very hard to teach, and hope that by formulating these procedures, I may help both to clarify how the method works and to provide beginners with a basis for developing the skills involved.

I can only list some techniques for collecting 'reminders' about how our concepts work. The task of organising and explaining the phenomena by means of some kind of generative theory is very difficult. It is similar to the construction of scientific theories. I do not claim to be able to teach people how to be good scientists. (That will have to wait until we have computer programs which behave like good creative scientists, when we shall be in a better position to think about what it is to teach someone to be a scientist!) What follows is merely a sketch, with a few hints. The topic deserves a whole book, and should be susceptible of a better organised presentation than I can manage.

4.2. Strategies in conceptual analysis

When trying to analyse a concept (e.g. *knowledge, truth, emotion, imagination, physical object),* some or all of the following moves may be helpful.

a. Collect descriptions of varied instances of the concept, and also descriptions of non-instances which are similar in some ways to instances. For example, consider the following examples of imagining something: imagining that the Conservatives will win the next election, imagining that you are very rich, imagining that you are falling off a cliff, imagining that there's a donkey in front of you, imagining that time travel will occur one day, imagining that 39875 is the largest possible number. Do these have anything in common? How do they relate to utterances like 'I can't imagine what she sees in him', 'He's a very imaginative dancer', etc. How do they differ? How do they differ from remembering something, learning something, believing something, reading about something, expecting something, planning, and dreaming? Try to formulate rules or definitions which will sort candidates into instances and non-

instances, and test your rules or definitions on those previously collected. Try to test them more thoroughly by searching for new difficult cases (which friends and colleagues may be more likely to provide since they will not be committed to your definitions.)

b. Try criticising and extending the definitions given in dictionaries. Dictionary writers are not normally trained in conceptual analysis, and may make mistakes. Moreover, the aim of a dictionary definition is not to explain how a concept works (e.g. what knowledge is presupposed by its use and how it is related to a family of concepts). Rather, the aim is merely to enable someone who already grasps the concept to attach a label to it. So dictionary definitions are usually much too brief and simple to be very useful for analysis of concepts. This comes out most clearly in the attempt to program computers to understand natural language: for such a purpose each word needs to be associated with much more elaborate rules for its use than will normally be found in a dictionary entry. In spite of this, dictionary entries may be good starting points when you are short of ideas.

c. Using a dictionary, and Roget's *Thesaurus,* try to collect lists of related words and phrases: analyses of different items in the same list will probably illuminate each other.

   For example, if analysing the concept *imagine,* look also at *image, imagination, suppose, consider, think, think about, think of, visualise, remember, invent, refer to, have in mind, . . .* Similarly, in analysing the concept *know,* we would need to look at *notice, discover, learn, believe, accept, understand, remember, forget, infer, evidence, reason, test, proof,* and many more. Having found some related but different concepts, try to find examples which illustrate one concept but not the other, and vice versa.

   Try to work out why each example fits one concept but not the other(s). For example, search for examples of knowing X without believing X, or examples of believing X without knowing X. (See Austin's use of examples to analyse the difference between 'by mistake' and 'by accident' in *'A Plea for Excuses'.* My chapter on the mechanism of mind was based on an attempt to extend his work.)

d. Try to collect lists grouped in different ways. For instance, one list given above included mental states and processes related to imagining. Another list would involve *uses* to which imagining may be put, for example, drawing something, solving a problem, trying to recall exactly what happened, entertaining people, anticipating difficulties while making plans, etc. One can then ask how it is possible for the process to be used in these various ways.

   This calls for a collection of examples of each kind of use to be thought about carefully, with a view to postulating some underlying mechanism. Another list might include a range of different kinds of things we can imagine (a visual scene, hearing a tune, doing something, a war starting, a mathematical theorem being false, etc.). (One of the things people find hard to learn is the technique of generating examples of things they already know about, including words and phrases. Wittgenstein was a master at this, though he was not very good at analysing the similarities and differences between the examples.)

e. There is a collection of very general categories we use in much of our thought and language, such as: event, act, state of affairs, process, disposition, ability, regularity, cause, explanation, function, object, property, relation, etc. (For example, if you find odd the assertion that apples hang on trees very slowly, this is because you (perhaps unconsciously) recognise that hanging is a *state* whereas 'slowly' can describe only *processes,* like growing.)

   Try fitting these categories to the lists of related concepts, to help bring out differences between them. For example, learning something is a process, knowing or believing something a state one is in (perhaps resulting from such a process). Believing something is a state

involving a property of oneself, whereas knowing something involves an extra relation to the world (e.g. getting something right). Since knowing something is a state not an event (contrast learning, or discovering, or noticing), those philosophers and psychologists who refer to 'the act of knowing' are either revealing their inability to analyse their own concepts, or else using technical jargon which is bound to cause confusion because of superficial resemblances to concepts from our ordinary language. (I do not wish to deny that ordinary language is itself sometimes muddled.)

Some mental states, for example, believing that there is a tiger in the next room, can explain behaviour, such as running away, but do not involve an ability or any disposition to behave. However, the combination of the belief and another state, such as fear of tigers, may generate a disposition to lock doors, run away, or call for help, depending on circumstances. Some states, for example, knowing how to count, involve an ability which may or may not ever be manifested in behaviour, whereas others, for example, being an enthusiast (e.g. about golf, gardening, or Greek sculpture), involve a tendency or even a regularity in behaviour. 'He smokes' reports a habit which is manifested (much to the annoyance of many non-smokers), whereas 'he would like to smoke' reports an inclination which may be successfully suppressed forever, so that there need not be any behavioural manifestations.

Desiring and wanting are states, whereas deliberating is a process, and deciding an event which terminates such a process and initiates a state of being decided.

Very often noun phrases look as if they denote objects, whereas analysis shows that they do not. Having an image is being in a certain mental state. The state may explain various abilities or actions. Some people think of an image as an object which is somehow involved in the state of having an image much as a nose is involved in the state of having a nose. However, it may be that this is not how the concept works, and that to talk of the image is merely a short-hand and indirect way of talking about a very complex mental state: when we say that a house has a shape we are not saying that besides the house there is some other object, its shape; rather we are alluding to an aspect of the state of the house, namely how all its parts are related to one another.

Similarly if someone has a visual image: this is a matter of being in a state in which one is able to do a variety of things which one can normally do only when there is something one can see. It does not follow that the image is some kind of object like a picture though no doubt, as with all mental states and processes, there is some kind of symbolism used (probably unconsciously) to represent the thing imagined. (For more on this see Pylyshyn, *'What the mind's eye tells the mind's brain").*

f. For each concept being investigated ask whether it refers to a *specific* kind of thing (event, state, disposition, etc.), or whether it covers a whole lot of different kinds of examples, in which case it *is polymorphous* (Ryle, *The Concept of Mind).* For example, the polymorphous concept *motive* covers desires, purposes, attitudes, attempts to achieve something, attempts to prevent or avoid something, and perhaps character traits ('the motive was greed'). If the concept covers many different sorts of cases, this is rarely simply because the word is simply ambiguous. So you can then ask why all these cases are grouped under a common description: do they fulfil a common function? do they have a common explanation? do they have a common relationship to some other things?

   For example, motives have in common the fact that (when combined with beliefs) they can *explain* decisions, intentions, and behaviour. But this shifts the burden to the concept *explain,* or *explanation',* why are there so many different sorts of things we call explanations, and do

they have anything significant in common? (An important and still open research question.) *Careful* is another example of a polymorphous concept: different sorts of things are involved in careful driving, careful teaching, careful selection of words in an essay, careful breaking of sad news, careful cleaning of a precious vase, etc. Here it is relatively easy to see what is in common to all these cases, namely reference to goals, possible undesirable occurrences, a collection of risks or dangers, paying attention to the risks, and doing whatever is required to minimise them.

g. If the concept appears to be polymorphous, ask whether there are some 'central' and some 'peripheral' or 'derivative' cases, and try to account for the difference. For example, describing a person as 'moody' or 'unco-operative' seems to be central compared with describing a car that way. Ask what distinguishes central from peripheral or metaphorical cases: is it a difference in the number of preconditions satisfied? If so, why does the concept have those preconditions? What is their point?

h. Ask what the role of the concept is in our culture. Is it merely a convenient descriptive symbol? If so, why should we want to describe those things? Does it have some non-descriptive function? For example, does it express approval? Is its use characteristically abusive, or a means of showing off? Is it part of a system of concepts whose use depends on the existence of some kind of social institution? What is the point of the institution? For example, is it used to apportion blame or responsibility in order to decide questions of redress? What would it be like to live in a culture without that institution? Is there some aspect of the concept which would remain useable without that institution?

   Examples of concepts which seem to depend on more or less complex social institutions are: courage, dignity, disapproval, honour, shame, embarrassment, owning, owing, impertinence and gallantry. Wittgenstein (in his *Philosophical Investigations)* and his followers have argued that very many mental concepts, including 'following a rule', are essentially social. I think that they exaggerate because of their ignorance of possible computational models of mental processes.

i. Ask what sorts of things can be explained by instances of the concept. Does it explain events, processes, states, abilities, non-occurrences, the loss of an ability, success, failure, a single occurrence, a number of occurrences, etc.?

   For example, knowledge explains (or is able to explain) success; fatigue and confusion explain failure; desire explains attempts.

   Does the explanation function as a cause, an enabling condition, a purpose, a justification, an excuse, a mechanism, a law, or what?

j. Ask the following range of questions about instances of the concept under investigation.

   1. What sorts of things can bring them about?
   2. What sorts of things can prevent them?
   3. What sorts of things can facilitate their occurrence?
   4. What can cause variations in the instances?
   5. What sorts of effects can they have?

   Sometimes it is possible to distinguish 'standard' from 'non-standard' causes, effects, etc. For example, there is something irrational about beliefs which are caused by desires ('wishful thinking') but not about actions caused by desires. (Why?)

   Sometimes it is useful to distinguish events and processes a person can bring about from those

which merely happen. You can decide to stop walking or trying to find something out, but you cannot decide to stop knowing or believing something. You can decide to try to get something, but you cannot decide to want it. Why not? (Answering this question would extend the theory of chapter 6.)

k. If you have managed to collect a number of examples of related concepts, see if you can find a set of relatively 'primitive' concepts and relations, which can be used to generate a lot of the examples, by being combined in different ways. (That is try to find a 'grammar' for the phenomena.) This is a useful first step towards building a good theory of how the concepts work, as opposed to merely describing lots of facts about their relations.

   Linguists are increasingly trying to do this though it is not clear how far they appreciate the intimate connection between the study of our language and the study of our world.

   For example, the verbs of motion mentioned earlier all seem to involve a subset of the following ideas:

   1. Something has a position which changes.
   2. Something is an agent (it may or may not also change position, and may or may not change the position of other objects).
   3. There is a route for the motion of each object, with a starting and a finishing location.
   4. Something may be an instrument, used by an agent, possibly to move an object.
   5. Moving things have absolute and relative speeds.
   6. If A causes B to move, A may be on the side away from which B is moving or on the side of B to which it is moving.
   7. The movement of B may merely be initiated by A (pushing something over the edge of a table) or may be entirely due to A (throwing something, pushing it along).
   8. The agent may have a purpose in moving the object.
   9. There may be a previous history of movements or locations referred to (e.g. if A retrieves B).
   10. There may be more than one stage in the motion (e.g.A fetches B).
   11. A may do to B something which tends to produce motion, but the motion may be resisted, e.g. pushing an object which is too heavy, pulling an object with a string which stretches or breaks.
   12. The agent may also be the supporter of the object moved, e.g. in carrying it, or may be supported by it, e.g. in riding it.

   Different combinations of these (and other) ideas can be used to generate whole families of related concepts, often including concepts for which we do not (yet?) have labels. For example, I do not think English contains a word which refers to a process in which an agent A carries an agent B to some location, and then A picks up some object and is carried, by B, back to the starting point.

   Perhaps this is an important part of some social activity in some other culture. Some sort of obstacle race?

   The 'primitive' ideas used as the basis for generating such a family of related concepts may themselves be susceptible of further analysis. Moreover, some concepts require mutually

recursive definitions: for example, *believe* and *desire* cannot be defined independently of each other.

The sort of analysis suggested here for concepts of motion is now familiar to linguists and people working in artificial intelligence (for example Schank and Abelson, who also explore analogies between such physical processes and mental processes like communicating information. See Bibliography.)

Similarly, in analysing a concept like *know,* or *knowledge,* it will be necessary to distinguish a variety of elements and relations which can enter into scenarios involving knowledge. A person (or other knower) will be involved, as will things in the world about which something is known. There will be a state of mind of the person, in which some aspects of the things and their relationships will be represented, that is, a belief is involved, though not necessarily consciously. There will be something which gives rise to the belief, either at the time the person knows or at some earlier time, for example, a process of perceiving something, doing an experiment or test, or perhaps acquiring the information indirectly from other knowers, or inferring it from some other knowledge.

There will be a relation between the source of the belief and the belief which certifies or justifies the belief (e.g. the evidence is good evidence). There may be sentences, spoken, uttered, or merely thought, which state whatever it is that is known, and in that case the sentences can be decomposed (usually) into fragments with different relations both to items in the world and aspects of the knower's mind. There may or may not be *uses* to which the knowledge is put, including answering questions, interpreting one's experiences, making plans, acting in the world, understanding other people's sentences, formulating new questions, etc. (Again, study of a system of concepts from ordinary language can contribute to psychology, and to the attempt to design artificial minds.)

In two papers on *ought, better* and related concepts (1969 and 1970), I have tried to show how a variety of uses can be generated in a fairly systematic fashion. Similarly, much important work in the development of mathematics, for instance Euclid's, and later Hilbert's, work on the foundations of geometry can be seen as a form of conceptual analysis, though usually of a very reductive sort (that is many concepts and theorems are reduced to a very small number).

l. When analysing a concept it may be helpful to try to list ways in which one can teach a young child or a foreigner learning one's language, how the concept works. What sorts of examples would make good illustrations, and why? What sorts of things would be worth mentioning as *not* being examples, and why are they likely to be confused? What sorts of things need not be mentioned because they can be taken for granted? Why? What would Martians have to be like in order to be capable of learning the concept?

m. Try to list ways in which you can test the truth or falsity of statements involving the concepts in question, including cases which might be difficult. For example, how do we decide whether a person has a certain attitude, such as anti-semitism? Is asking the person an adequate test?

When is it adequate and when not? What patterns of behaviour are adequate tests? Are they *decisive,* or are they merely *indicative?* Why? Are there some situations in which no decisive test is possible, so that doubts cannot be removed? For example, a racialist who has excellent motives for concealing his attitude, and who is an excellent actor. (As we shall see later on, there is no reason to suppose that there should be behavioural tests for all internal computational states and processes, either in a computer or in a person or animal.)

n. Sometimes it is useful to ask whether being in a certain state presupposes having some knowledge, or exercising some intellectual ability. For example emotions like surprise,

dismay, embarrassment, shyness and humiliation presuppose a lot of knowledge. You can long for your mother only if you know you have one, know she is not present, and can imagine a possible state of affairs in which the two of you are together. Can a goldfish long for its mother? If not, why not?

The widespread belief within our culture that intellectual and emotional phenomena are quite disparate can be refuted by detailed conceptual analysis.

o. Often some question about the analysis of a concept can be investigated by telling elaborate stories about imaginary situations. So science-fiction writers are good sources of material for this activity. For example, imagine a time when machines are available which will make a complete copy of a human body (including the state of the brain), except that cancer cells are replaced with healthy cells.

Suppose that in such a society it is commonplace for incurable cancer sufferers to agree to have their bodies copied by this machine, while under total anaesthetic, followed by cremation of the cancer-ridden body. The new one is allowed to take its place so people come home from hospital saying I'm glad to be back, and I feel much better now that I've got my new body'. In such a society is our concept 'murder' applicable to their treatment? Is the concept 'same person' applicable to the person who goes into the hospital and the person who comes out? (For more on this see my 'New bodies for sick persons'.)

Another example: people disagree over whether it is essential to the concept 'emotion' that emotions involve felt bodily changes. One way of convincing yourself that such physiological processes are not essential is to imagine a society of Martians who are very much like us with very similar sorts of social institutions and similar ways of seeing, thinking, and acting, but who do not have the bodily reactions which we (or some of us) feel in certain emotional states. So they have hopes, disappointments, pleasant and unpleasant surprises, they feel pity, loneliness, dismay when their plans go wrong, they are anxious when there is a high probability of things going wrong, they are proud of their achievements, envious of others who are more successful, greedy for wealth, and so on. By describing the behaviour and social interactions of such beings in great detail, and imagining what it would be like to communicate with them, you should be able to convince yourself that you would find it perfectly natural to use our emotion concepts in talking about their mental states.

You would say 'He's terribly embarrassed about the attention he's getting', even though he feels no hot flush in the cheeks or any other physiological change characteristic of embarrassment in humans.

Of course, this sort of investigation does not produce knock-down arguments, because people can differ in how their concepts work. For example, mathematicians use a concept of *ellipse* which includes circles, whereas for non-mathematicians a necessary condition for something being an ellipse is that it has major and minor axes of differing lengths. Similarly, there may be some people for whom the accompanying physiological changes are necessary conditions for the applicability of concepts like *envy, embarrassment, loneliness,* etc. However, what one can demonstrate to such people is that by insisting on these necessary conditions they are making it impossible for themselves to describe situations which might one day arise, without inventing a whole lot of new terminology which may prove very hard to teach. Whereas I would claim that my use of the non-physiological concept of emotion in no way interferes with my communication with other people, and allows me the power to read science fiction without any feeling of linguistic distortion.

p. Try to test your theories by expressing them in some kind of computer program or at least in a

sketch for a design of a working program. For example, try to design a program which can communicate with people using the concepts. If you have analysed the concepts wrongly then this will show up in some failures of communication between the computer and people (just as the misunderstandings of children and other learners show up). Or test your analysis by designing a program whose behaviour is intended to *instantiate* the concept, then see whether the actual behaviour is aptly described using the concepts in question. You will usually find that you have failed to capture some of the richness of the concept. For example, for a while some people hoped that programs written in the language PLANNER would capture the essence of the concept of a goal, or purpose. But the behaviour of the programs clearly quashed this hope. (E.g. see Winograd, 1972.)

Of course, sometimes a little thought makes this elaborate kind of test unnecessary. Nevertheless, the methods of A.I. provide a useful extension to previous techniques of conceptual analysis, by exposing unnoticed gaps in a theory and by permitting thorough and rapid testing of very complex analyses.

This account of conceptual analysis is by no means complete. For more detailed examples, refer to the writings of philosophers mentioned and also A.R. White's *Attention,* and his contribution to *Owl of Minerva,* (ed. Bontempo and Odell), and Margaret Boden's *Purposive Explanation in Psychology.* Philosophers usually do not pay enough attention to problems of describing mental processes. Neither do they normally attempt the kind of system-building involved in designing a 'grammar' for a collection of concepts in the manner hinted at above. For instance, is there some sort of grammar for concepts related to attention? In other words, is there a relatively small subset of concepts in terms of which all the others can be defined? I believe the answer is 'Yes' but to establish this will require designing a fairly detailed model of a person, capable of generating a large number of processes involving perception, deliberation, reasoning, planning, problem-solving, and execution of plans and intentions. Some small steps in this direction are taken in Chapter 6, which proposes some minimal architectural requirements for a human-like system.

Despite my disparaging remarks about philosophers, there have been some profoundly important systematic analyses, mostly produced by philosophers of logic and mathematics, such as Frege, Russell, Tarski and Prior. For example, Frege's analyses of concepts like *all, some, nobody,* and related quantifiers, led to a revolution in logic and has profoundly influenced the development of computer programming languages used in artificial intelligence (via the work of Alonzo Church). Austin's *How to do Things with Words* is another example of a philosopher's attempt at detailed and systematic analysis, which has made a great impact on linguistics and more recently on AI.

If only Wittgenstein, in his later writings and teaching, had not made such a virtue of his inability to construct systematic theories integrating the results of his analyses, a whole generation of philosophers might have been far more disciplined and productive.

Of course, there are dangers in insisting on everything being formalised and systematic. Much shallow theorising is a result of trying to fit very complex and messy structures into a neat and simple formal system. A well known example of the distorting effect of formalisation is the claim that the logical connectives of prepositional calculus adequately represent the words 'and', 'not', 'or', 'if, etc. of ordinary language. However, even if this claim is false, it remains true that the formalisation provided a basis for deeper exploration than was previously possible. For example, by describing exactly how the use of the ordinary words deviates from the truth-functional symbols, we obtain useful descriptions of how they work. (See Gazdar and Pullum 1977.) The same can be said of some other systematic but inaccurate analyses.

The two extremes to be avoided are demanding formalisation of everything at all costs, and rejecting

formalisation because some of our concepts are too complex and unsystematic in their behaviour for us to be able to represent them in elegant formal systems. One of the great advantages of using programming languages for formulating analyses of concepts (as Winograd did see his 1973), is that programming languages are well suited to include many tests for special cases and exceptions to general rules. It is much harder to use formal grammars, or axiomatic systems, for this purpose.

4.3. The importance of conceptual analysis

The activity of attempting to analyse families of related concepts can be enjoyable and interesting in its own right. Discussion of similarities and differences between *fetch, retrieve, carry,* and related concepts is the sort of thing even a child can find good fun though getting the analysis right is not child's play. But besides giving intellectual pleasure, the activity may have a useful function. For example, it is well known that many perennial philosophical problems arise out of confused reflections on things we all know, and that at least some of these problems can be solved or dissolved with the aid of conceptual analysis. I think it can also be shown that a great many debates on ethical and political issues, such as debates about the justifiability of abortion, about equality of educational opportunity, and about what sorts of decision-making procedures are democratic, are often more confused than necessary either because the participants are using concepts in a muddled fashion or because they are to some extent at cross purposes because of subtle differences in the ways their concepts work. In either case progress can be made if people learn how to analyse their own and other people's concepts.

Conceptual analysis can play a role in science and mathematics too. I have already mentioned Einstein's work involving analyses of concepts like *simultaneous,* and other spatial and temporal relations. Another example is the struggle by mathematicians of previous centuries to clarify the concepts *infinite* and *infinitesimal,* leading to the discovery of the concept of a limit, and to formal set theory.

Every science will have at its frontiers concepts which are to some extent in need of analysis and possibly improvement. Not all the problems of science are to be solved simply by collecting new facts, or by using existing terminology to build new theories. In the mature sciences, the concepts most in need of analysis will usually be highly technical, remote from the concepts of ordinary language.

However, in the social sciences and psychology, and increasingly in artificial intelligence, concepts from ordinary language play acentral role in the construction of new theories and in the description of phenomena to be explained. Thus it is important for practitioners of these disciplines to be sensitive to the need for analysis, and to be skilful at doing it.

The dangers of failing to analyse concepts properly can be illustrated by a few rather extreme examples. Someone who had not seen how the concept *bachelor* worked might think it interesting to do a survey to find our what proportion of the bachelors in some social group were unmarried. He would probably get no support from research councils. However, less obvious mistakes of the same sort could pass unnoticed, like attempts to test the hypothesis that *other things being equal* people tend to believe things which are asserted by those they respect, or the hypothesis that *other things being equal* people tend to try to achieve goals they think they can achieve, or the hypothesis that being embarrassed involves believing that other people are paying attention to you. Of course, such research goals would usually be disguised in obfuscating jargon, but that does not reduce the need for conceptual analysis. I once read a research proposal which looked very impressive until the English equivalent to the jargon emerged. The aim was to find out whether people tend, on the whole, to co-operate more successfully if they get on well together. (For some similar criticisms of Social Science,

see Andreski, *Social Science as Sorcery.)*

An example of an important piece of biological theorising whose concepts cry out for detailed analysis can be found in Dawkin's *The Selfish Gene.*

Besides the role of conceptual analysis in preventing muddled thinking and silly research, there is another important role in relation to science, namely making explicit some of what we already know, clearly a useful preliminary to attempts to add to what we know. I believe this is especially useful in fields like developmental psychology and anthropology, concerned with the study of ways of thinking and learning. Previously I listed some concepts concerned with spatial movement and indicated how one might begin to analyse some of the more complex ideas in terms of combinations of relatively primitive ones. Very young children somehow acquire both the relatively 'primitive' concepts and also a variety of complex combinations of these. It is not thought to be beyond them to grasp the difference between 'fetch' and 'send for' expressions which occur in familiar nursery rhymes. By studying these concepts we can define some of the tasks of psychology. An adequate theory of learning must account for a child's ability to master these ideas. Even very young children are capable of grasping quite abstract rules, including rules which they cannot formulate in words. For example, a three-year-old reacted to his older brother's use of 'nope' for 'no', by starting to say not only 'nope' but also 'yesp', 'okayp' and 'thankyoup'. Try formulating the rule he had invented! (Do developmental psychologists, or brain scientists, have any convincing explanation of the ability to learn these things?)

By improving our understanding of what it is that our children have to learn we may perhaps come to understand better not only how they learn, but also what sorts of things can go wrong with the learning process, and perhaps even what can be done about it. How many teachers in schools, colleges and universities have sufficient skill in conceptual analysis to be able to discern subtle differences between the concepts they are trying to teach and the concepts so far grasped by their pupils?

Other social sciences can also benefit from conceptual analysis. By doing this sort of analysis for concepts used in several different cultures, anthropologists and sociologists could enhance their studies of what is common and what varies among different modes of thinking and reasoning.

I have already alluded several times to the role of conceptual analysis in the work reported in this book. Several chapters are based in part on attempts at analysing familiar concepts. But most of the work is still sketchy and makes use of concepts which themselves require further study.

The chapter on the aims of science, for example, makes liberal use of a very complex concept which still requires further analysis, namely the concept of what is *possible.* Several other concepts used in that chapter are equally in need of further investigation.

The chapter on analogical representations attempts to analyse a familiar distinction between different sorts of symbolisms, or representations, showing that the verbal/pictorial distinction is usually misdescribed and that there are actually several different distinctions where at first there seems to be only one.

The chapter on learning about numbers begins to analyse some of our simplest number concepts, drawing attention to complexities in what a child has to learn which are not normally noticed.

The chapter on computer vision, and the ensuing discussion includes some small steps towards clarifying a collection of familiar concepts like *conscious, interest, experience.*

Nearly all of this work is incomplete, and will remain incomplete for many years. But, as I have suggested in this chapter and will try to substantiate later, the methodology of artificial intelligence will be a major spur to progress.

[[Note Added November 2001

Since this chapter was first published, the problem of 'knowledge elicitation' in designing expert systems has received much attention. It is not widely appreciated that the techniques of conceptual analysis as described here (and practised by many philosophers) are often crucial to such knowledge elicitation. There is also considerable overlap between these ideas and the Naive Physics project proposed by Pat Hayes: See P.J. Hayes, The second naive physics manifesto, in *Formal Theories of the Commonsense World* Eds. J.R. Hobbs & R.C. Moore, Norwood, NJ, Ablex, 1985, pp. 1-36

Note Added February 2007

Additional discussion of the nature of conceptual analysis, its relationship with what Gilbert Ryle called 'logical geography', and a possibly deeper notion of 'logical topography' can be found here http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html

]]

Endnotes (1) Margaret Boden, Frank O'Gorman, Gerald Gazdar and Alan White commented usefully on an earlier draft.

Book contents page
Next: Chapter five
Last updated: 28 Jan 2007

CHAPTER 5

ARE COMPUTERS REALLY RELEVANT?

Experience has shown that many readers will have been made very uncomfortable, if not positively antagonistic, by my remarks about the role of computing and computer programs in philosophy and the scientific study of human possibilities. There are several reasons for this, including (a) ignorance of the nature of computers and computer programs, (b) misunderstandings about the way computers are used in this sort of enterprise, (c) invalid inferences from the premises that computer simulations of human minds are possible, and (d) confused objections to specific theories expressed as computer simulations.

5.1. What is a computer?

It is not helpful to think of a computer simply as something which does numerical calculations, for this is only one use of a far more general facility. A computer is a mechanism which interacts with symbols. It can accept symbols, store them, modify them, examine them, compare them, construct them, interpret them, obey them (if they express instructions), or transmit them. It must therefore include a 'store' or 'memory' containing a large number of locations at which symbols can be stored. These locations must be 'addressable': that is, it should be possible for an instruction somehow to mention a location so that its contents can be examined or something new put there. The mechanism may assume that all the basic symbols stored used some fixed format, such as sequences of zeros and ones, but that is no restriction, as sufficiently complex combinations of such symbols can be used to represent anything, just as complex sequences of the simple characters on a typewriter can express poems, plays, propaganda or physical theories, or complex arrays of dots can be seen as photographs of faces.

Since the symbols stored in the computer may include instructions for it to obey, and since it can be instructed to change some or all of the symbols within it, it follows that as a computer executes instructions within itself, the instructions may change and thus the processes occurring may evolve in complex ways. In the end, the original program may have completely disappeared. Exactly how this happens may depend not only on the original program but also on the history of interactions with the environment. So no programmer, or anybody else, is responsible for the eventual state of such a mechanism or for its behaviour.

In any modern digital computer the basic symbolic processes which occur will all be very simple, such as putting a zero or a I in some location, or comparing two symbol-strings, or copying the contents of one location into another, or performing logical or arithmetical operations. But it is not helpful to think of a computer as 'simply' performing such simple operations, any more than it is helpful to think of a Shakespeare play as 'simply' composed of letters, punctuation marks, and spaces.

Computers can perform millions of their basic operations each second. Many different kinds of books can be written using the same small set of printed characters, and similarly an enormous variety of processes can be represented by complex combinations of the simple processes in a computer.

In particular, the processes need not be fully controlled by all the symbols in the store at any time. For

among the instructions executed may be some to the effect that new symbolic information should be accepted from various devices attached to the computer, such as a television camera or a microphone, or a teletype at which a person sits communicating with the computer. Some of the new symbols coming into the computer in this way may lead to changes in the stored instructions, just as much as execution of stored instructions can. (This, incidentally, is why all the philosophical debates about Godel's incompleteness theorem and related theorems proving that there are limits to what any particular computing system can do, are irrelevant to the problem of what sorts of intelligent mechanisms can be designed: for all these theorems are relevant only to 'closed' systems, i.e. systems without means of communicating with teachers,etc.)

Computing science is still in a very early phase. Only a tiny fragment of the possible range of computer programs has so far been investigated, and many of these are still only partly understood. Complex programs sometimes work for reasons which their designers only half understand, and often they fail in ways which their designers cannot understand. It follows that nobody is in a position to make pronouncements about the limits of what can be done by computer programs, especially programs which interact with some complex environment,as people do.

Attempting such pronouncements is about as silly as attempting to use an analysis of the printing process to delimit the kinds of theories that will be expounded in text-books of physics in a hundred years time. Nevertheless, people with theological or other motives for believing that computers cannot match human beings will continue to be overconfident about such matters (e.g. H. Dreyfus, *What Computers Can't Do).*

The last general remark I wish to make about computers is that the definition given above does not assume anything about what the mechanism is made of. It could be transistors, it could be more old-fashioned electronic components, it could be made of physical components not yet designed, it could somehow be made out of some non-physical spiritual stuff, if there is any such thing. The medium or material used is immaterial! All that matters is that enough structures are available to represent the required range of symbols, and that appropriate structural changes can occur in the computer. As Margaret Boden once remarked, angels jumping on and off pin-heads would do.

This is not the place to enlarge further on what computers are. Interested readers should consult *Electronic Computers,* by Hollingdale and Toothill, or Weizenbaum's *Computer Power and Human Reason.* See also chapter 8 of this book.

5.2. A misunderstanding about the use of computers

I have heard people talk as if computers were some new kind of organism, distantly related to humans or other animals, so that one might perhaps learn something about animals or their brains by studying computers!

However, computers are not natural objects to be studied. They are artefacts to be improved and used. If people had been content to *study* computers instead *of programming* them, very little would have been learnt, for a computer does nothing unless it is programmed. But what it does depends on how it is programmed. So approaching a computer with a view to finding out what it can do is as silly as it would be for a physicist to study pencil and paper with a view to finding out what they can do. One approaches a computer in order to try to make it do something. The physicist writes things down, calculates, tries out formulae and diagrams, etc. He constructs, explores and modifies a theory. That is how to use a computer in order to study intelligence: by designing a program which will make it behave intelligently one constructs a theory, expressed in that program, about the possibility of intelligence. The failure of the theory is your own failure, not the computer's.

So objections to the discipline of artificial intelligence based on the assumption that its practitioners *study* computers are completely misguided.

5.3. Connections with materialist or physicalist theories of mind

Many readers (some sympathetic and some unsympathetic) will jump from the premiss that computer programs can simulate aspects of mind, or can themselves be intelligent and conscious, to the conclusion that some kind of materialist or physicalist theory of mind is correct. Alternatively they will assume that because I stress the importance of computing studies, I support some kind of reductive materialist theory. There are two answers to this.

The short answer is that just because an electronic computer is a physical system, it does not follow that everything it successfully simulates is a physical system: there could be computer programs simulating the structures and functions of mechanisms composed of some spiritual substance!

So even if the human mind is not merely a function of the physical brain, but has some non-material or non-physical basis (whatever that may mean), then the behaviour or function of that stuff is what computer programs can simulate. In fact a program does not specify what kind of computer it runs on. The computer may use transistors, valves or spiritual mechanisms, so long as a rich enough variety of structural changes is available, as I have already pointed out.

A longer, and more important, answer is that the *ontological* status of mind has little relevance to the problems of this book. Both Dualism, which postulates some kind of spiritual entity distinct from physical bodies, and Materialism, according to which minds are just aspects of complex physical systems, lack explanatory power. That is, both of them fail on the criteria proposed in chapter 2 for adequate explanations in philosophy or science. They fail either to describe or to explain any of the fine structure of such aspects of mind as perception, memory, reasoning, understanding, deciding, desiring, enjoying, creativity, etc., or the relations between them.

In order to explain how all these things are possible, we need a theory describing or representing the structures and functions of a mechanism which can be shown to have the right sorts of abilities, that is a mechanism able to generate within itself structures and processes with the kinds of mutual relationships which we know hold between mental phenomena. For instance, we know that a certain experience, such as seeing a tool being used, can produce a change in what a person knows, and thereby can change what he is able to do and the decisions he can take in order to deal intelligently with problems. To explain how this sort of thing is possible, e.g. to explain how one can learn to operate a tool by watching its use, it will not do simply to say what kind of *stuff* the underlying human mental mechanism is made of.

Being told that a computer is made of physical components, for instance, tells you nothing about the kind of internal organisation that made it possible for the PDP-IO computer used by Winograd (1973) to hold conversations in ordinary English. Similarly, being told that the mind is spiritual or non-physical explains nothing.

For similar reasons, neurophysiology cannot help in the early stages of the search for explanations of the possibility of mental phenomena and we shall remain in the early stages for some time. Studies of neurophysiology, or the electronic basis of a computer, may explain such things as how fast the system performs, and why it sometimes goes more slowly, or why it sometimes breaks down altogether; but cannot at present explain how it is possible for the system to perform a particular type of task at all. Such an explanation requires study of the brain's *programs,* not its low level (physical) *architecture* and neurophysiology currently lacks conceptual and other tools needed for studying programs. (Study of a computer's architecture tells one practically nothing about the programs

currently running on it. The programs may change drastically while the physical architecture remains the same, and different computer architectures may support the same programs. Computers are not like clocks.)

[Note added 2001: I would now put this by saying that the virtual machine architecture is more important than the physical machine architecture. (For more on this see recent papers in http://www.cs.bham.ac.uk/research/cogaff/). The study of physical architectures would be relevant if could be used to demonstrate that certain sorts of virtual machines *could* and others could *not* run on brains. But right now we still do not know enough about ways of mapping virtual machines onto physical machines for useful constraints to be derived.]

The only kinds of explanatory mechanisms that have some hope of being relevant to explaining mental possibilities like perception, learning and decision making, are mechanisms for manipulating complex symbols, for example, computer *programs.*

People whose sole experience of computing is with programs for doing highly repetitive algorithmic numerical calculations, or programs for simulating feedback systems, may find it hard to understand how programs can be relevant to our problems. An essential antidote to this prejudice is a study of the literature of artificial intelligence to learn how, besides doing numerical calculations in an order determined by the programmer, computer programs can also construct, analyse, interpret, manipulate, and use complex symbolic structures, like lists, pictures, sentences or even sub-programs, in a flexible way determined by analysis of developments during the computation rather than following an order worked out in advance by the programmer.

All this can be summarised by saying that the known important mechanisms are not *computers* (those ugly boxes with mysterious noises and flashing lights), but *programs* or *virtual machines*. Computers are an old type of mechanism: they are physical machines. Programs are a new type. A simulation program could drive not only a physical computer, but, if ever one were made, a computer composed entirely of spiritual stuff (The program, not the medium, is the message.)

5.4. On doing things the same way

The persistent objector may now argue that the explanatory power of computer programs is doubtful, since even if a program does give a machine the ability to do something we can do, like understand and talk English, or describe pictures, that leaves open the question whether it does so *in the same way* as we do; so it remains unclear whether the program gives a correct explanation *of our* ability.

The objector may add that it is clear that existing computers do *not* do things the way we do, since, at the physical level they use transistors and bits of wire, etc., whereas our brains do not, and even at the level of programs they have to employ interpreters or compilers which translate the high level intelligent and flexible symbol-manipulating programs into sequences of very simple and very mechanical instructions which have to be followed blindly, whereas there is no evidence that humans do this.

This objection (which seems to pervade the book by H.L. Dreyfus, *What Computers Can't Do),* is based on the concept 'doing things in the same way', which requires some analysis.

The notion of doing something *in the same way* is systematically ambiguous. Two persons may calculate the answer to an arithmetical question in the same way insofar as they both use logarithms but in different ways insofar as they use logarithms with different bases. It is all a matter of how much and what sort of detail of a process is described in answer to the question In what way did he do it?' That some very detailed description would be different in the case of a computer does not imply that there is no important level at which it does something the same way as we do. We don't say a

Chinaman plays chess in a different way from an Englishman, *simply* because he learns and applies the rules using a different language, so that his thinking goes through different symbolic processes. He may nevertheless use the same strategies.

The same problem arises about whether two computer programs producing equivalent results do so in the same way. Two programs using essentially the same algorithm may look very different, because they are written in different languages or in different programming styles. Any program is a mixture of 'main ideas' and implementation details. The same may be true of human abilities.

The problem of knowing the way in which a computer does something is no different in principle from the problem of knowing the way in which a person does. it. In both cases there are questions that can be asked, and tests that can be given, which provide useful clues. (Compare Wertheimer's tests for whether children understand and apply a technique for finding areas of a parallelogram in the same way as he does, in *Productive Thinking,* chapter I. He sees whether they can solve a very varied range of problems.)

Insofar as anything clear and precise can be said about 'the way' in which a human being does something (e.g. plays chess, interprets a poem, or solves a problem) the appropriate procedure can in principle be built into a suitable simulation, so that we *ensure* that the machine does it in the same way. For instance, programs can be written to do multiplications using ordinary decimal arithmetic, or binary arithmetic, or alternatively using natural language.

Finally it should be noted that it is very unlikely that there is only *one* way in which something or other is done by *all* human beings, whether it be perceiving faces, remembering names, playing chess, solving problems, or understanding a particular bit of English: we all have our own quirks and foibles, so it-is unreasonable to deny this right to a complex computer simulation.

I do not wish to argue that *every* aspect of the human mind can be simulated on digital electronic computers, any more than an astronomer's explanation of an eclipse explains or predicts every aspect of the motion of the earth, moon and sun. For instance, certain types of human experience seem to be possible only for beings with human bodies, or bodies with very similar structures. Thus, feeling thirst, nausea, muscular exhaustion, sexual desire, the urge to dance while listening to music, or the complex combination or bodily sensations when one is about to lose one's balance whilst walking on ice, may be forever inaccessible to computer programs within immobile rectangular boxes, or even to humanoid mobile robots who are made mainly of plastic and metal. (For more on these general issues, see the contributions by H.L. Dreyfus, N.S. Sutherland, and myself to *Philosophy of Psychology,* ed. S.C. Brown.)

These abstract debates about what can and cannot be done with computer programs are not too important. Usually there is more prejudice and rhetoric than analysis or argument on both sides. What is important is to get on with the job of specifying what sorts of things are possible for human minds, and trying to construct, test, and improve explanations of those possibilities. Anyone who objects to a particular explanation expressed in the form of a program, should try to construct another better explanation of the same range of possibilities, that is, better according to the criteria by which explanations are assessed (see chapter 2). The preferred explanation should account for at least the same range of possibilities with at least as much fine structure.

The rest of this book will be concerned mainly with the description of some important possibilities known to common sense, together with some rather sketchy accounts of what good explanations might look like. I shall frequently point out ways in which the attempt to design computer simulations can subserve the endeavour to understand the human mind.

[[Note added 2001:

After this book was published there was a revival of interest among many AI researchers in "connectionist" architectures. Some went so far as to claim that previous approaches to AI had failed, and that connectionism was the only hope for AI. Since then there have been other swings of fashion. It should be clear to people whose primary objective is to understand the problems rather than to win media debates or do well in competitions for funding that there is much that we do not understand about what sorts of architectures are possible and what their scope and limitations are. It seems very likely that very different sorts of mechanisms need to be combined in order to achieve the full range of human capabilities, including controlling digestion, maintaining balance while walking, recognising faces, gossiping at the garden gate, composing poems and symphonies, solving differential equations, and developing computer programs such as operating systems and compilers. I don't know of an any example of an AI system, whether implemented using neural nets, logical mechanisms, dynamical systems, evolutionary mechanisms, or anything else, that is capable of most of the things humans can do including those items listed above. This does not mean it is impossible. It only means that AI researchers need some humility when they propose mechanisms. **]]**

[[Note added 20 Jan 2002:

A number of arguments against computational theories of mind have been advanced since this book was written. Many of them use arguments that were already rebutted in this chapter, or put forward views that were expressed in this chapter. For example, the argument that brains work in different ways from computers therefore computational theories of mind must be incorrect is rebutted above by pointing out that systems may be different at one level of description and the same at a more abstract level of description. Abstraction is often very useful, as demonstrated by the history of science in general and physics in particular. The argument that intelligence or mentality requires embodiment is rebutted by pointing out that some aspects of mind may depend on details of the body whilst others do not. Of course, that leaves unanswered the important research question: which forms of embodiment can support which forms of mentality?

Many critics of AI and some defenders of AI have based their argument on the assumption that AI in some sense presupposes that all computation is Turing Machine computation. I have tried to argue in recent years that the notion of "computation" is not sufficiently well defined to support such criticisms. In particular I have argued that the notion of "computation" employed by most users of computers, designers of computers, programmers, and AI researchers, has nothing to do with Turing machines but is an extension of two notions which go back to long before Turing, namely

a. The notion of a machine that can control something, possibly itself
b. The notion of a machine that operates on abstract entities, such as numbers, or census information.

Both ideas were well advanced before the beginning of the twentieth century, for instance in automated looms, mechanical calculators and Hollerith machines for sorting and collating information. In the middle of that century advances in science and technology

made it possible to combine those ideas in new ways, providing far greater speed, power, flexibility (e.g. self programming), and cheapness. These points are elaborated in a paper on the irrelevance of Turing machines to be published during 2002, and other papers available here: http://www.cs.bham.ac.uk/research/cogaff/

Despite all the progress of the last half century, it is clear that we still have much to learn about the nature of information and varieties of machines, including virtual machines, that can process information -- a theme developed in these talks: http://www.cs.bham.ac.uk/~axs/misc/tdks/ **]]**

Book contents page
Next: Chapter 6

PART TWO: MECHANISMS

CHAPTER 6

SKETCH OF AN INTELLIGENT MECHANISM[Note 1]

6.1. Introduction

Much of this book is concerned with describing aspects of the human mind. In the present chapter, I shall try to provide an overall theoretical framework by describing very briefly and crudely a type of mechanism which could be represented on a computing system and which would simulate some of the important general features of the human mind mentioned in other chapters. Such a computer model would provide at least a tentative and partial explanation of certain forms of human possibilities, including the possibility of accidental discoveries, and creative redeployment of old resources.

In particular, I want to undermine a common misconception about computers, namely that however complex the programs that run in them they are always essentially unintelligent, uncreative mechanisms, blindly following simple rules one at a time. Such a description may well be true of the underlying electronic components, just as it may well be true to say that a human brain is always essentially an unintelligent uncreative bundle of nerve-cells (or an assemblage of atoms) blindly reacting to one another in accordance with chemical and physical laws of nature. But just as the latter description may omit some important features of what a brain can do, so also the former description omits important 'high-level' features of complex computer programs. What is true of a computer need not be true of a program, just as what is true of a brain need not be true of a mind. In both cases the whole is far more than the sum of its parts.

I am not trying to explain phenomena which are unusual, hard to observe, and known only to experimental psychologists. The facts about people that I take for granted and attempt to account for are facts which we all know, though we may not all reflect on them, they are part of our shared common-sense.

6.2. The need for flexibility and creativity

In particular, I shall try to sketch the overall architecture of a computing system which could cope with a variety of domains of knowledge in a flexible and creative way, so that, like people, it can use available information, skills and procedures in order to solve new problems, or take decisions in new situations, in ways which were not explicitly foreseen or planned for by the programmer. The architecture to be described is not physical but computational. It concerns global organisation, rather than detailed mechanisms, and the sub-mechanisms are virtual machines rather than physical machines. (Though they are all implemented in physical machines, e.g. brains.)

Many notable examples of creativity are discussed in A. Koestler's *The Act of Creation.* However, we can also observe frequent examples of what seems to be essentially the same kind of flexibility and creativity in the daily life of ordinary persons, in our efforts to cope with spilt milk, ungrammatical sentences, unfamiliar typewriters, blind alleys, broken suspenders, lost keys, illegible handwriting, mixed metaphors, puzzle pictures and veiled insults. The child who uses his old counting ability as a

basis for answering new questions (like 'what number comes before five?') is as creative as any artist or scientist. How can we explain this flexibility and creativity?

What is required is a design for a computing system which is able to cope with types of possibility not covered by the programmer's analysis. More precisely, it is necessary to combine into a single system, competence in a variety of domains, in such a way that expertise in two or more domains can be combined creatively and flexibly in dealing with novel situations or problems. Instead of the programmer doing the analysis of all types of possibilities in advance, the program should be able, in at least some cases, to do the analysis when it is appropriate to do so, and to record the results for future use.

6.3. The role of conceptual analysis

Some insight into the mechanisms underlying human flexibility can be found in a philosophical analysis of such familiar concepts as *notice, alert, interested, puzzled, surprised, understand, cautious, attend, careless, reckless, discern, try, recognize that,* and many more. This analysis shows that to explain, by means of a computer simulation, how it is possible for available resources to be deployed in an intelligent and creative way, we need at least to construct a system which can act on the basis of multiple purposes or motives. Moreover, in the course of executing some action it must be able:

a. To notice something it was not explicitly looking for

b. To interrupt, abandon, modify or suspend some or all of the current action,

c. To search through its stock of resources for items which satisfy a current requirement or need, possibly in an unforeseen way (see (a)),

d. To relate parts and effects of one action to purposes preferences or other motives *besides those which generated the action*, or more generally, to relate facts to problems and purposes not currently being attended to.

These abilities require the system to contain mechanisms which facilitate communication of information between different sub-processes in an unplanned way: the programmer need not have anticipated each possibility inherent in the system. I shall now give a sketchy description of how this might be achieved. Several steps in the construction of such a system have already been taken by people designing artificial intelligence programs (e.g. Sussman, 1975).

Note: 2004

Sussman's book (based on his PhD thesis) seems to be available here:
http://portal.acm.org/citation.cfm?id=540310&dl=ACM&coll=portal

---

6.4. Components of an intelligent system

I shall describe (at a very high level of abstraction) some of the *structures* the system will have to contain; some of the *procedures* (or programs) that will be needed to inspect, construct, manipulate, and use those structures, and some of the *processes* that will be generated when those procedures are executed. The structures and processes may be either inside the mechanism or in the environment.

However, it will be seen to be useful to blur the distinction between the mind of the mechanism and the environment. (This blurring in one form or another has a long philosophical history. See, for

example, Popper's *'Epistemology without a knowing subject',* reprinted in his *Objective Knowledge.* As he points out, Plato, Hegel and Frege had similar ideas.)

We shall discuss interactions between the following structures:

1. an environment,
2. a store of factual beliefs and knowledge,
3. a store of resources (for instance a dictionary and previously learnt procedures for making things, solving problems, etc.),
4. a catalogue of resources,
5. a motivational store,
6. a process-purpose index (action-motive index),
7. various temporary structures associated with ongoing information processing.

Many more or less temporary internal and external processes (actions) will be generated by these structures. There will also be the following more permanent processes ensuring that the actions which occur are relevant to the current motives and that intelligent use is made of previous knowledge and new information:

1. central administrative processes, not to be confused with an homunculus (see also p. 244, Chapter 10, below);
2. a set of monitoring processes, including both permanent general-purpose monitors and others which are more specialised and are set up temporarily in accordance with current needs;
3. a retrospective analysis process, reviewing current beliefs, procedures and plans on the basis of records of previous occurrences.

The system must have several kinds of processes running simultaneously, so that implementing it on a computer will require multi-processing time-sharing facilities already available on many computers, and used in the POPEYE vision program described later in chapter 9. This *global* parallelism is an important requirement for our mechanisms, though concurrent processes can be implemented in a very fast serial machine.

6.5. Computational mechanisms need not he hierarchic

The main parts of the mechanism will be described separately in terms of their functions. However, computing models, unlike previous kinds of mechanisms, should not be thought of as composed of several interlocking parts which could exist separately, like parts of an engine or human body. Normal concepts of part and whole do not apply to computing structures and programs.

For instance, two data-structures stored in the memory of a computer, containing pointers to their elements, may contain pointers to each other, so that each is an element of the other. This can be illustrated by so-called 'list-structures'.

Thus, a list A may contain, among other things, the list B, while list B contains the list A. A is then an element of B and B an element of A (which is not possible for physical collections). A list may even be an element, or part, of itself. Examples of circular structures will be found below in chapter 8. (For further details consult a manual on some list-processing programming language, e.g. Burstall, *et al.*

1973, or Foster, 1967, or a manual on Lisp, Prolog, Scheme, or Pop-11).

Similarly, computer programs may be given names, in such a way that at a certain point in the set of instructions defining one program. A, there is an instruction of the form If condition X is satisfied then run program B', while program B contains a similar call of program A. Program A may even contain an instruction to run itself. (These are examples of 'recursion'.) Such circular programs are able to work provided certain conditions are satisfied which can be roughly summed up by saying that during execution the series of nested, or embedded, calls to programs (sub-routines) must eventually produce a case where a particular program can run without meeting the conditions which make it call another: it can then do its job and feed the result back to the program which called it, which can then get on with its job, and so on. This is commonplace in programming languages which permit recursion, such as ALGOL, ALGOL68, LISP, or POP-2.)

In such cases, we can say that program A is part of program B, but B is also part of A. More complex chains or networks of such circular relationships between programs are possible. Similarly, human abilities, as we shall see, combine to form complex systems in which normal hierarchic part-whole relationships are violated: each of two abilities may be a part of the other. For instance, the ability to read someone's hand-writing may be a part of the ability to understand his written sentences, and vice versa.

Because these ideas have been made precise and implemented in the design of computing systems, we can now, without being guilty of woolly and unpackable metaphors, say things like: the environment is part of the mechanism (or its mind), and the mechanism is *simultaneously* part of (i.e. 'in') the environment!

We turn now to a sketch of structures, programs and processes in a mechanism to simulate purposiveness, flexibility and creativity. I cannot give more than a bird's eye view of the system at present. My description is deficient, in that it does not provide a basis for a team of experienced programmers to construct such a system. At best, it provides a framework for further research.

6.6. The structures

A structure is a complex whole with parts standing in various kinds of relationships. The chapter on numbers (Chapter 8) gives several examples. As parts are removed, replaced, or added, or relationships between parts changed, processes occur within the structure. In order that a structure be usable by the system, there must be mechanisms (already to be found in many computing systems) which are able to identify the parts, read off or compute their relationships, match one structure against another, interpret one structure as representing another, and perhaps perform deductions in order to extract implicit information about the contents of the structure. The system may also be able to modify or 'update' the structure, by modifying the components or their properties and relationships, or adding new components. All the structures listed below will be capable of changing, but some will be much more dynamic than others. Such manipulable structures are often referred to as 'data-structures'. They function as complex symbols, or representations.

The different structures about to be mentioned are listed separately in terms of their different functions in the system. But they need not exist separately. As already remarked, one structure may be part of another which is part of it. Some of the structures are in the mind (or computer), some not.

6.6.(a) The environment

This is a domain in which configurations can exist and processes occur, some but not all of them produced by the mechanism itself, and some, but not necessarily all of them, perceivable by it. For

instance, the environment will normally be a space-time domain inhabited by the mechanism.

But for human beings it may also include, at the same time, a more abstract culturally determined domain, such as a kinship system, or a system of socio-economic relationships, within which the individual has a location. Some of the 'innards' of the mechanism or person may also be thought of as part of the environment, since the system can examine and act on them! (See chapter 10 for more on this.) Similarly, parts of the environment, like internal structures, may be used as an information store (blazing a trail, writing a diary, 'reading' the weather-signs, putting up signposts), so that the environment is part of the store of knowledge, that is, part of the mind.

6.6.(b) A store of factual information (beliefs and knowledge)

This is a set of descriptions or representations of aspects of the form and contents of the environment (including descriptions of some of the system's own 'innards'). It may include specifications of the current situation, previous history, (especially records of the system's own actions and their effects), and predictions or expectations about the future.

Several different kinds of language or symbolism may be used, for instance sentences, networks representing sets of relationships, maps, diagrams, and templates. Some of the information may be procedural, for example, in the form of routes and recipes. The information will necessarily be incomplete, and may contain errors and inaccuracies. There may even be undetected inconsistencies, and mistakes about the system's *own* states and processes. We do not necessarily know our own minds.

What gets into the store will depend not only on what stimuli reach the sense organs, but also on what languages and symbolisms are available for expressing information, and on what kinds of perceptual analysis and recognition procedures (i.e. the monitors mentioned below) are available and active. (What is already in the store will also make a difference. Where things are stored will depend on indexing procedures used.)

In order that its contents be readily accessible, this store of beliefs will have to have an index or catalogue associated with it, possibly including general specifications of the *kinds* of information so far available or unavailable. For instance, it should be possible to tell that certain types of information are not present without exhaustive searches. (How long does it take you to decide whether you know what Hitler ate at his last meal?) The index may be implicit in the organisation of the store itself, like the bibliographies in books in a library, and unlike a library catalogue which is kept separate from the books. If books contained bibliographies which referred directly to locations in the library (e.g. using some internationally agreed system for shelf-numbers) the analogy would be even stronger.

6.6.(c) A motivational store

In a mind there will be at any time many current purposes and sub-purposes, preference criteria, constraints on permissible actions, plans for current and future actions, specifications of situations to be avoided, etc. These likes, dislikes, preferences, principles, policies, desires, hopes, fears, tastes, revulsions, goals, ambitions, ideals, plans and so on, have to be accessible to the system as a basis for decision-making or execution, so they will need to be formulated, in an appropriate symbolism, in a motivational store.

If they are not explicit, but are implicit in decision-making procedures, then it will be much harder for the system to become aware of the reasons for what it does, and to revise its decision-making strategies. (Compare the discussion of consciousness in chapter l0, below. A more detailed analysis would distinguish first-order motivators, such as goals or desires, from second-order motive-

generators or motive comparators, e.g. attitudes, policies and preferences.)

Some of the contents of the store will have been generated on the basis of others, for instance as means, plans, or strategies for achieving some end. (This store must not be thought of as a 'goal-stack' with only the last added goal accessible at anyone time as in some over-simple computer models.)

The representational devices may be varied: for instance some *motivational* information might be stored in an apparently *factual* form within the previously mentioned store of beliefs, for example, in sentences like 'Jumping off objects is dangerous', or 'Nasty people live in town T'. This can work only so long as adequate procedures are available in at least some contexts for finding and using this sort of information when it is relevant to deciding what to do.

The processes produced by the mechanism, that is its actions, whether internal or external, will be generated, modified, controlled, interrupted, or terminated, by reference to the contents of the motivational store, in ways to be explained briefly below. Such purposive actions may include planning processes, the construction of new motives, problem-solving processes, external movements, manipulations of objects in the environment, processes of modifying plans or actions generated by other processes, and also perceptual or monitoring processes.

One of the constraints on the design of a human-like intelligent system is the need to act with speed in many situations. This has some profound design implications. In order that rapid decisions may be taken in a complex world there will have to be a very large set of 'rules of thumb', including rules for deciding which rule to use, and rules for resolving conflicts. This is almost certainly incompatible with assumptions made by economists and some moral philosophers about how (rational) people take decisions. For instance, there need not be any overall tendency for the rules to optimize some abstraction called 'utility'.

At any time, some of the purposes or other motivational factors may not yet have generated any process of planning or action: for instance, a purpose may have been very recently generated as a new sub-purpose of some other purpose, or it may have a low priority, or there may not yet have been any opportunity to do anything about it, or it may be a conditional purpose (do X if Y occurs) whose condition has not been realised, or some other purpose or principle (for example, a moral principle) may override it. *Thus many existing motivational factors may generate no decisions.*

Similarly, plans and decisions that have been formulated on the basis of motives may still not have generated any action for analogous reasons.

6.6.(d) A store of resources for action

This includes not only usable objects in the environment, such as tools, materials, sources of information, teachers, and so on, but also the store of factual beliefs, and, most importantly, a set of readily available programs, or procedure-specifications, some of which may use one another recursively (as explained above in section 6.5, p. 116).

Among the resources should be linguistic or symbolic abilities. Those are needed for formulating problems, purposes, procedures and factual information. Chapter 2 indicates some of the reasons why notational resources can be very important. Chapter 7, below, explains why different kinds of symbolisms may be required for different sorts of tasks or sub-tasks. Other resources would include procedures for constructing plans or routes (for example, with the aid of maps), procedures for getting information and solving problems, such as problems about why an action went wrong, and procedures for constructing, testing and modifying other procedures. (See Sussman 1975 for a simple working example.)

That is to say, the resources store will include collections of 'intelligent' programs of the sorts

currently being produced by workers in artificial intelligence. The concept of a resources store, like the concept of an environment, expands to swallow almost everything! This is why a catalogue is necessary.

6.6.(e) A resources catalogue

It is not enough for resources, like objects and abilities, to be available. In order that they be intelligently usable, the system must have information about them, such as what kinds of purposes or functions they are typically useful for, the conditions under which they are applicable, likely side-effects, possible dangers, and any other information which may be relevant to selecting a particular resource and embedding it in a plan, perhaps with modifications to suit current needs and conditions.

It must be possible for new information about typical causes and effects, or requirements, of old resources to be added to the catalogue. The system may have to use pointers in the catalogue in two directions, namely, starting with some purpose or need, it should be able to use the catalogue to get at available resources which might meet that need. So pointers are needed from purpose-specifications to resources. However pointers are also needed the other way, since in selecting one resource it may often be important to know what sorts of uses, and effects, it can have besides the one which led to its selection: if some other typical use of the resource matches another current motive or need, then the resource may be 'a stone that kills two birds'. Alternatively, if a resource selected as a possible means to one end has a typical effect which will frustrate some other current purpose (or principle, or preference, etc.), then an alternative resource should be sought, or the other purpose abandoned. Those are some of the design implications that follow from the need to cope with multiple motives.

Sometimes information about typical uses and side effects of a procedure (or other resource) can be got by inspecting its *structure*. But often such things are learnt only by experience of using the resource and in the latter case we need *explicit* additional entries.

For a very large store of resources, as in the human mind, the catalogue will have to be highly structured, for instance in the form of a tree, with lower levels giving more details than higher levels. The organisation of the catalogue may be partly implicit in the searching and matching procedures. As indexing can never be perfect, the system will have typically human failings, no matter how fast and large a computer is available. (This is contrary to some optimistic pronouncements about the way bigger and faster computing systems will enable super intelligences to be made.)

This catalogue of resources, like the index to factual beliefs, need not be physically separate from the store of resources: it may be partly implicit in the organisation of the store.

6.6.(f) A process-purpose index (or action-motive index)

The central administrative mechanism (described below) will set various processes (internal and external actions) going, on the basis of analysis of contents of the current motivational base together with analysis of the resources catalogue and the store of information about the environment. Some of the processes will consist of sets of parallel or sequential sub-processes, which in turn may have a complex inner structure. Some of the processes may be lying dormant, waiting for starting signals, or resumption signals from monitors (see below). This is commonplace in computer operating systems.

In order to be able intelligently to modify ongoing processes, terminate them, interrupt or suspend them, change their order, and so on, in the light of new information, the system will have to have information about which processes and sub-processes are generated by any given motive, and which motives lay behind the initiation of any one process.

The function of a process-purpose index is to store this information about the reasons for various

actions. It may need to be modified whenever a new process is initiated or an old one terminated, or when any of the reasons for doing something change, for example, if one of the birds a stone was intended to kill turns out to be already dead. The system will thus have access to the reasons why it is doing things. Faults in the procedures for keeping the process-purpose index up to date may account for some pathological states.

So if a process generated by purpose PI accidentally achieves a purpose P2, and this is detected by the monitors, then the index shows which other processes were generated by P2, and can therefore be terminated, unless the index still contains a pointer from one of them to some other as yet unfulfilled purpose P3. Other uses of the process-purpose index will be mentioned below.

Perhaps one of the most important reasons why it is necessary to be able to be in the midst of several different processes at once, is that this provides opportunities to *learn* from accidental interactions between processes. The process-purpose index, which relates current activities to the reasons for doing them makes it easier to achieve such learning. For example, one might learn that a certain purpose can be achieved in a new way, because of an unexpected interaction between the old strategy for achieving it, and some other activity.

The *process-purpose index* should not be confused with the relatively more static, less changeable, *resources catalogue*. Their functions are different. For instance, a particular procedure may be selected, using the resources catalogue, in order to achieve purpose PI, and then executed. While the process of execution is going on, the same procedure may be selected again in order to achieve another purpose P2. We thus have two processes (actions) running in parallel in order to achieve different purposes, yet the same procedure (or program), a relatively permanent resource, controls them.

A clear example of this is a person playing two games of chess simultaneously, and using the same strategy in the two games for at least part of the time. If one of the opponents makes a move requiring that strategy to be abandoned, the process of executing it has to be terminated in one game but *not* in the other.

The resources catalogue contains the relatively permanent information (modifiable in the light of experience) that this strategy is normally useful in such and such circumstances for achieving certain types of advantages. The process-purpose index, however, relates not the strategy itself, but, for example, two current *executions* or *uses* or *activations* of the strategy, which may have reached different stages of advancement, to two current purposes. Similarly the ability to multiply may be used twice over in evaluating the following expression:

$$\frac{(17-12) \times (6+5)}{(3 \times 2)}$$

The process-purpose index would also have an important place in planning activities, when instead of real executions of strategies the index would contain pointers to representations of possible executions of strategies.

6.6.(g) Temporary structures for current processes

At any time, various ongoing processes will have associated with them structures containing information about partial results, current values of variables, next instruction or procedure step, current subgoals, where to send results, and so on.

Once the importance and ubiquity of such structures in a complex goal-directed information processing system has been understood, the distinction sometimes made between two kinds of

memory -- short-term and long-term -- evaporates, for instance, in connection with a plan carried out over a period of several years.

Note added April 2004

That point was very badly expressed, or just wrong. What I think I was trying to say in 1978 is that there are different memories with different time-scales and different functions, and assuming there are only two kinds, short-term and long-term memory, as some people appeared to claim in those days, was a mistake.

The full details of these temporary structures need not be globally accessible in the same way as some of the previous a structures. That is to say, they are private to the particular processes which use them. It may be, however, that certain local computations, are automatically reported up to a global level whenever they occur, such as estimates of time or computing space, or other resources needed for a process to be completed. This might be done by monitors, described below.

Some of the temporary workspace may be outside the system, for instance a shopping list, an artist's rough sketches, an engineer's calculations. Even a half-completed object is an extension of short term memory for the constructor.

Note added April 2004

Examples would be a partially completed painting, a partial mathematical proof, a half-built shelter. A similar point was made long ago by Herbert Simon in connection with insects that produce the next step in a complex task by reacting to the current state of the environment in which they are building something. This notion is often referred to as 'stigmergy' and the phenomenon was known to entomologists in the 1950s. Good ideas are often re-discovered.

These more or less temporary structures are of no use to an intelligent system unless mechanisms are available which can bring about the sorts of processes already hinted at and elaborated below. Typical mechanisms would be procedures for accessing, using, and modifying the resources, catalogues, plans, etc. The whole system needs some kind of overall control. This is the business of a central administrative process, for which computer operating systems provide a very first (very rough) approximation.

6.6.(h) A central administrator

[[Paragraph added in 1986:

To some extent, parts of the system may (and will) work autonomously in parallel, e.g. posture control, control of breathing, and control of saccadic eye movements. However, since two or more needs may require incompatible actions, and since coordinating two actions rather than performing them separately may improve overall performance, it may be useful for some 'central' system to resolve conflicts and co-ordinate decisions. A 'central administrative process' may have this role. **]]**

The central administrative process will at various times survey the motivational base and purpose-process index and select from the unfulfilled purposes a subset for generating further planning and action. This selection may be driven partly by previously selected purposes or principles, and may use current information, such as estimates of likelihood of success or failure, knowledge about opportunities and resources available now and in the future, the current state of other actions, and so on.

Sometimes no selection can be made until a change has been made in *the set of purposes* for instance by inventing a compromise between two conflicting purposes. In at least some cases, the selection must be automatic to avoid an infinite regress of decision making.

Similarly, after certain motives or purposes have been selected for action, then in at least some cases they must invoke suitable action-generating procedures automatically, since if everything required prior deliberation or planning, nothing could ever get started. This automatic activation can happen when a current purpose closely matches a catalogued specification of a typical use of an available procedure. Monitors would be employed to reduce the risks inherent in some automatic activation.

When no matching procedure is found for a certain purpose P, in the resources catalogue, it may be possible instead to find a match for the new purpose of *making a plan for achieving P*. For instance, if the purpose 'Go to Liverpool' fails to match any current plan, then 'Make a plan for going to Liverpool' may match a typical use (that is, making plans for going places) of a procedure for constructing *routes* (for example, find an atlas containing both your current position and the destination, then . . . etc.). Again, even if one does not yet know a procedure for making *objects* of a certain type, one may have a *procedure for constructing a suitable procedure* for making those objects, by analysing specifications of a required object and available tools and materials.

In short, when first-order matching fails, second-order matching may succeed. Perhaps in some cases even higher-order matching (make a plan for making a plan for achieving P) may succeed.

Similarly, if several procedures are found to match the purpose, then a new purpose may have to be set up, namely the purpose of choosing between the available alternatives. If a choice cannot be made using the information in the resources catalogue, it may be necessary to try out some of the alternatives. (See chapter 8 for more on the difference between examining and executing procedures.) This kind of comparison of alternatives may occur at various stages in the construction of one plan, contrary to the games-theoretic analysis of human decision-making which assumes that we always choose between *complete* alternatives, without saying anything about how we construct those alternatives.

When the administrator has failed to find or produce a plan for a certain purpose, a second-order task may have to be added to the motivational base as a new unachieved purpose (i.e. finding a plan), to be attended to later if anything relevant crops up, in ways described below. (This can produce accidental learning.) Alternatively, if the original purpose was very urgent, or there is nothing else to do at the time, then trial and error with back-tracking may be used.

Note added April 2004

Some of the above ideas later turned up in SOAR, a problem solving and learning system developed in the early 1980s. SOAR detects when an *impasse* occurs and switches to a new task, resolving the impasse. However, as far as I know, SOAR did not include the option in the previous paragraph, namely *deferring* the process of dealing with an impasse until some unspecified future date. I believe SOAR also did not include the point in the next paragraph about checking unexpected benefits and side-effects of proposed new solutions, which became an important feature of many planning systems apparently inspired by Sussman's HACKER (Sussman 1975) referred to above. For more on the ideas in SOAR see

Newell, A. (1980b). Reasoning, problem solving and decision processes: The problem space as a fundamental category. In R. Nickerson (Ed.), *Attention and Performance VIII.* Hillsdale, NJ: Erlbaum.
http://sitemaker.umich.edu/soar

Should a suitable procedure for achieving P be found or constructed, in any of the above ways, then analysis of its typical uses and effects (recorded in the resources catalogue), or analysis of its structure, may show that it, or a modified version, will enable more than one current purpose or motive to be fulfilled. If several suitable alternatives are available, this analysis may provide a reason for choosing between them. Or it may show that the procedure would interfere with some other current purpose. A process (action, procedure-execution) is then generated by executing the selected procedure with suitable arguments or parameters bound to its variables (for example, 'destination' might be bound to 'Liverpool' in the previous example).

The central administrator (and perhaps also some of the other currently running programs) must be able to interrupt, terminate, modify, or restart current processes (though some may be less controllable than others, for instance if they are so well-tried that possible interrupt points have been kept to a minimum). These control decisions will be taken on the basis of new information from *monitors* (described below), using the purpose-process index as described above. So the index must be changed every time a process is begun, modified, halted, or found to be capable of serving an unexpected purpose as a side effect, as well as when ongoing processes set up new sub-goals and generate corresponding sub-processes.

Some processes which include complex sets of sub-processes, may have to have their own private purpose-process indexes in their private work spaces (see p. 124), as well as being more briefly represented in the main index. They may also have their own central administrators!

Chapter 10 attempts to relate the idea of central decision-making processes to the distinction between what we are and what we are not conscious of.

6.6.(i) Perception and monitoring programs

Mechanisms must be available for inspecting the environment in which the system acts, such as familiar types of sense-organs for inspecting the external environment, and less familiar mechanisms for accessing structures within the system (the internal environment). However, all that a *physical* sense-organ can do is produce some kind of spatial or temporal array or manifold of physical values (as a television camera or microphone does). This does not yet amount to perception: it simply amounts to the production of a new structure within the system. Whether anything is thereby perceived, and what is perceived, depends on what procedures (or programs) are available for analysing the new structure, finding relationships between its parts, perhaps manipulating or modifying it (for example, correcting misprints or other errors), interpreting it, and making use of all this either immediately or later on in the performance of actions or solving of problems.

Such perceptual procedures may involve computations of arbitrary complexity, using a great deal of background knowledge, like the perceptual procedures involved in a medical diagnosis or the tuning of a car engine. Even ordinary perception of simple shapes and familiar physical objects can be shown to presuppose considerable factual and procedural knowledge. This is why perception cannot be separated from cognition. See chapter 9 for more details.

So the system needs a collection of perceptual procedures, for analysing and interpreting various kinds of structures in various kinds of contexts. The limits of these procedures together with the limits

of the sense-organs and the current store of information about the environment will define what the system is capable of perceiving. Systems with the same physical sense organs may therefore have quite different perceptual abilities as we know from variations in human perception. Thus there cannot be any such thing as perceiving things 'directly' or 'as they are in themselves'. As Max Clowes once put it: "We inhabit our data-structures". The same must be true of intelligent machines. So the objective/subjective distinction evaporates. (Compare Boden, 1977.)

The range of types of objects, properties and relationships that human perceptual procedures are capable of coping with is enormous. So in a sensible system they will not all be applied to every possible chunk of sensory input or meaningful structure. For instance, when you last read a page of typescript you probably did not use your ability to notice that the letters on the page were in vertical columns as well as horizontal rows; and while listening to someone talking one language you know, you do not apply the analysis procedures which would enable you to recognise in his syllable-stream the sounds of words of another language you know. Did you notice the let' in letters' or 'horizon' in 'horizontal' above? If every available analytical and interpretive procedure were applied, their outputs would form an enormous information store, and the system would then have the problem of perceiving its contents in order to make use of the information.

It seems not only sensible, but also to correspond to human experience, to have only a small selection of available perceptual programs running at any time in relation to any one piece of 'perceivable' structure, such as the structures mentioned in the previous sections or those produced by sense-organs. There are serious problems in explaining how appropriate programs are selected.

The active analysis programs may be called *'monitors'*[note 2] and it seems to be necessary to have two main kinds of monitoring general purpose and special purpose. The former involves frequent and large-scale application of relatively simple analyses and tests which have a good chance of being relevant to a wide range of purposes and circumstances. (Is anyone calling out my name? Is something on my retina moving?) The special purpose monitors may be more complex, and will be set up only when there is a specific reason to expect that they will find something or that if they find something it will be very useful in relation to current motives.

In either case the monitor need not itself complete the analysis and interpretation of new information. Instead, what it finds may act as a *cue* (or reminder, or stimulus) which will invoke (e.g. via a catalogue or index of resources) more complex object-specific or problem-specific procedures.

For instance, if the environment is a spatial domain, then a visual retina might be designed with very many relatively simple general purpose monitoring procedures 'wired into the hardware', for efficiency, instead of being expressed as programs. So the retina might be divided into many small regions, each being constantly monitored to see whether any change has occurred in some physically detectable property (brightness, colour, graininess of texture). If a change is noted, the monitor sends an interrupt signal to inform processes which may need the information. Other general purpose monitors might be constantly monitoring these monitors to see whether something which has consistently been reporting changes stops doing so. There may be general purpose monitors not only at the interface with the physical environment, but also at several other interfaces. Perhaps every time one of the globally accessible structures (such as the motivational base or process-purpose index) is accessed or modified by any current process, a general purpose monitor will note this and send an appropriate signal or take appropriate action (such as recording the fact for future reference). In recently developed programming languages this is achieved by 'pattern-directed procedure activations'. It is also a common feature of computer operating systems, for example, to prevent unauthorised access to information.

A very useful general purpose monitor would be one on the lookout for 'I've been here before'

situations: this might enable loops, infinite regresses, and unnecessarily circuitous procedures to be detected. However, the concept of 'the same state as before' admits such varied instantiations that it cannot be tested for in general by any one procedure. General tests might therefore have to be restricted to a few possibilities, like a return to the same geographical location, much more specialised monitors being required if other kinds of repetition are to be detected another source of fallibility in complex systems.

This will not work if records of previous states are not retained. Alas, people do not remember everything, not even their own actions. Repetitions often go undetected, like recounting their exploits or telling you a joke for the nth time. However, we shall see below that remembering apparently useless things may be an essential pre-requisite for certain kinds of intelligent behaviour and learning.

A 'found something' signal from a general purpose monitor may function simply as an *invitation* to some other program or monitor to look more closely, applying special purpose perceptual procedures to see if the occurrence is important to current motives or processes. Depending on what else is going on, the invitation may be ignored, or the new information may simply be stored, without further analysis, in case it will be useful later on. (Note that this presupposes some indexing procedure.)

Special purpose monitors may be much more complex, may have a much more transient existence, and may be set up at all levels of complexity in the system. For instance, in dealing with someone we know to be 'difficult' we need to be on the look-out for danger-signals in their behaviour. And while searching for a proof of some mathematical formula, one may have good reason to suppose that if certain sorts of intermediate results turn up in one's calculations they will enable an easy proof to be found, whereas if others turn up they will show that the formula was not provable after all. In that case one could set up monitors to be constantly on the lockout for the 'accidental' production of such results. (For examples, see Wertheimer *Productive Thinking.)* The tests for the occurrence of such special cases need not be at all trivial, and it may be necessary to make inferences from obscure cues, learnt in the course of considerable previous experience.

Watching out for multiplication or division by zero when simplifying equations illustrates this: zero may be heavily disguised in an expression like:

$$a^2 + (a + b)(b - a) - b^2$$

So the monitoring required will have to be pretty sophisticated. The same applies to detecting signs of irritation, dismay, incomprehension, etc. in one's spouse or pupils.

Normally the 'something found' signal from a special purpose monitor would be less likely to be ignored than signals from general purpose monitors, partly because the latter will always be crying 'wolf' and partly because the setting up of a specialised monitor will reflect the importance of its results, for current purposes.

Discoveries of the analytical and interpretative programs constituting monitors may be added (perhaps after some filtering by intermediate monitors) to the belief system (see section 6.6.(b)), forming a record of events and discoveries. At this stage a particularly important general purpose monitor should be available to try matching each addition to the belief system against currently unfulfilled purposes, or at least a 'high priority' subset of current motives, to see whether the new information satisfies or obstructs any of them. For example, the newly discovered fact or technique may be a solution to a problem you were thinking about yesterday. If it is a *general* purpose monitor it will have to use *crude* matching techniques, so some relevant relationships will be missed unless specialised monitors are set up. (Again, we see how fallibility is a necessary consequence of complexity.)

Not every piece of new information can be stored permanently. The problems of indexing, shortage of space, searching for what is relevant etc., would make this unworkable. But it may be possible to store information for a short time in case it turns out to be relevant to some process or purpose other than that which generated it. This will be most useful in the case of 'raw' data acquired for one purpose but potentially useful for others. If only the *interpretation* of such data is stored, then useful information may be lost. So besides the interpretation made for one purpose it may be useful also to store, at least temporarily, the original uninterpreted information in case it turns out to be relevant to other purposes. It must therefore be stored in a globally accessible structure.

In order to be really flexible and creative, the system will have to be able to activate specialised monitors, from time to time, which ask the following questions about new items of information as they turn up:

i. Does this imply that a particular current purpose has been achieved or frustrated?

ii. Does it imply that particular current purposes are unexpectedly near to or far from being achieved?

iii. Does it imply that a current purpose can be achieved more efficiently or quickly or at less risk or cost, or in a more enjoyable way, etc., by modifying an ongoing process or terminating it and starting with a new strategy: that is, is there a better way of doing what is currently being done? what is currently being done?

iv. Does it imply that any current purposes are mutually incompatible?

v. Is this worth examining more closely to see if questions like (i) to (iv) get a positive answer after specialised investigation.

Although such questions may occasionally be answered by a simple match between a current purpose and new information, at other times the full problem-solving power of the system may be needed in order to detect the relevance of a new fact, another example of the recursive, or non-hierarchic, nature of computational systems. For instance, a stored resource may not be found by a straightforward search in the resources catalogue. However, some further analysis of what is needed may solve the problem of where to search. Alternatively, it may later be found to be related to a current problem only when, by chance, it is turned up as a result of a search generated by some *other* need, and a monitor, or the central administrator, causes its relevance to the earlier purpose to be investigated. The person who is looking for both a screwdriver and eating utensils may be more likely to recognise the knife on the table as a potential screwdriver than the person who is simply looking for a screwdriver. But he must also be able to relate the structure of the knife to the function of a screwdriver.

6.6.(j) Retrospective analysis programs

For efficient and creative learning, the system will need to analyse fairly lengthy and detailed records of events. Such records will, as already pointed out, need to contain more detailed information than is *obviously* relevant to current needs, information retained in case it turns out to be useful.

For instance, examination of a series of failures over a long period of time may suggest a generalisation about what caused them, leading to a modification of some old procedures. (Of course, some people never learn from their failures, especially their failures in dealing with other people. Why not?)

Similarly, if successes are sometimes achieved unexpectedly, the system should go back and try to

find out whether enough information was previously available for the bonus to have been predicted and planned for, in which case some existing planning procedures will again need to be modified.

Records of events must also be searched for refutations of previously accepted generalisations, and for new patterns suggesting deeper explanations of previously known phenomena. In some cases, retrospective analysis of difficulties in getting at relevant stored resources may show the need for reorganisation of the catalogue of resources or the index to information. Thus, all sorts of comparisons need to be constantly going on, relating new information, old information, current motives, and possible future motives.

Once again, retrospective analysis cannot be done simply by a general purpose program, if it is to be at all deep. There must be a preliminary general analysis of unsolved problems to suggest that certain particular types of questions need to be investigated, and appropriate special purpose investigation procedures invoked or constructed.

Normally many questions like 'Was my failure due to bad luck or was there something wrong with the procedure by which I worked out a strategy?' will remain unanswered. Unanswered questions can be added to the store of unfulfilled purposes, thereby enlarging the motivational base and possibly influencing the course of events later on, if for instance, one of these problems turns out accidentally to match some information generated by another purpose.

Moreover, these unsolved problems may themselves generate new processes of experimentation or exploration, for instance in order to test some tentative hypothesis about the scope of a regularity or the explanation of a surprise. Without a major driving force provided by the need to answer questions and solve problems, it is hard to see how human infants could possibly learn as much as they do in the first years of life. It is paradoxical that the words 'play' and 'toy' are often used to denote this most important of all human activities and its instruments. It is also worth noting that unless the system in some way consciously or unconsciously distinguishes errors in its own procedures from failures due to the environment, it cannot modify its procedures and learn. Thus even new-born infants and any organism that learns, must have a rudimentary concept of 'self', contrary to popular opinion.

6.7. Is such a system feasible?

It will not be easy to construct a working computer model containing these structures, procedures and processes. Many problems have to be solved which are hardly even mentioned in the ludicrously brief specification I have given. A suitable type of environment must be chosen for the initial attempts, with a rich but interconnected variety of possible structures and processes. There are many difficulties in enabling so many processes to interact. Existing computing systems used in artificial intelligence are too small. Symbolisms will have to be developed for expressing various sorts of purposes, possibilities and plans, and for formulating entries in catalogues, the purpose-process index and the belief system.

No simple and uniform notation can be expected to work for all cases: sometimes a desired object may have to be represented in terms of a function that it can fulfil, sometimes in terms of a verbal description of its structure, sometimes in terms of a procedure for constructing it, and sometimes in terms of a template or model, similar in structure to it. Usually a combination of representations will be needed.

A language which is suitable for formulating a procedure (or program) so that it can be executed efficiently need not be equally good for constructing the procedure in the first place nor for describing how that procedure works so that its uses and limitations can be understood. The system may have to use one language while a procedure is constructed and debugged, after which it is translated (that is,

compiled) into some less accessible, less intelligible, less easily modified but more efficiently executed form.

Programming such a system would be an enormous task, yet it seems that existing expertise makes it all possible in principle. For instance there are complex operating systems which permit several different processes to run on a single computer, as if in parallel (because small chunks of each are run in turn), interacting with each other as they go, and this would enable several monitoring programs and administrative programs to run at the same time as programs for planning, executing actions and retrospective analysis. The POPEYE perception project, described in a later chapter, illustrates the possibility of such parallelism, in a simple form.

6.8. The role of parallelism

It may be necessary to have a computer with several different processors, each containing its own time-sharing system, each processor being devoted to some major function described above, but all having access to the same set of stored structures, but it is not obvious that this is necessary. It is often supposed that the human brain has some tremendous advantage over electronic computers because it operates in parallel, but there is no reason to suppose that having very large numbers of processors working in parallel would be an advantage if they all had the opportunity of producing complex changes in a central store: the mess might be impossible to control.

Parallel processors might be of use only for relatively simple, general purpose monitoring of the kinds already described, such as the monitoring of a retinal array for simple events, and perhaps the monitoring of stored symbols for crude and obvious matches with widely broadcast current requirements.

Since all this can in any case be simulated on a single, serial processor, the distinction between serial and parallel physical processors has not much theoretical significance for our purposes. This is not to deny that parallel processing (which can in principle occur on a serial processor) is crucial for the kinds of interactions between processes described above.

> **[[Note added 2001.**
>
> This point became clearer in the 1990s and beyond when AI researchers saw the importance of *architectures* for complete systems, instead of concentrating only on representations and algorithms. See my 1992 review of Penrose: A. Sloman, 'The emperor's real mind', Review of Roger Penrose's *The Emperor's new Mind: Concerning Computers Minds and the Laws of Physics*, in **Artificial Intelligence**, 56, pp. 355--396, (available at the CogAff web site **]]**

6.9. Representing human possibilities

All this is part of an answer to the question: How should we represent the human mind in such a way as to do justice to the enormous variety of possibilities inherent in it? What a person experiences, thinks and does over a period of time is a single process with many sub-processes, but weknow that if he is a normal person then his mind contains, during that time, the potential for many other processes which might have occurred. (Compare Ryle, 1949, on dispositions). A computing system of the sort just described could represent an explanation for such human possibilities. Of course, the proposed system could not account for all major human possibilities without further complex mechanisms. For instance, nothing has been said yet about how the possibility of moods and emotions is to be explained. This would involve various kinds of disturbances of the central processes. (For instance, feeling startled is sometimes a result of rapid automatic re-organisation of a collection of plans and

actions in the light of a sudden awareness of new danger.)

How wide a range of possibilities existed in such a system at any time would depend on such things as how wide a range of resources were stored in it, how complete the catalogues of resources and beliefs were, and what kinds of matching mechanisms were available. These, in turn, would depend, as in the case of a human being, on how much and what kind of previous learning had occurred. A mind contains a culture. So anthropology must be part of psychology, and vice versa.

Nobody could hope to design a complete adult humanoid robot. At best, it may be possible to produce a baby mind with the ability to absorb a culture through years of interaction with others.

6.10. A picture of the system

It is possible to give a crude representation of some of the range of possibilities inherent in such a computing system, or a person, by means of a flow-chart. This consists of a set of boxes (nodes in a graph) representing possible states or sub-processes of the system, joined by arrows representing the transitions that can occur between them.

If four arrows lead from box A to other boxes, this implies that there are four possible states, or sub-processes, which can occur after the one represented by A. Which one does occur will, normally, depend on what sorts of things occur in phase A, which in turn may depend on features of the context, including a long history of previous processes. (The normal approach in social science and much psychology is to shirk the task of understanding such dependencies, by representing the transitions as probabilistic, and studying the frequencies with which they occur in a large sample examined superficially, instead of studying particular transitions in depth.)

Each box represents a *state* of the whole system, so a flowchart of this sort should not be confused with a chart in which the boxes represent *mechanisms* of some kind and the arrows indicate *flow* of something like energy or information. Mechanisms are not states or phases, and flow between parts is not the same thing as transition between states of the system. A flow chart is not an architecture diagram (though it may imply some architectural features.)

The chart summarising many examples of familiar kinds of human behaviour, follows. (In the original book it was on pages 138-9).

This kind of flow-chart (see next page) can be misleading in various ways.

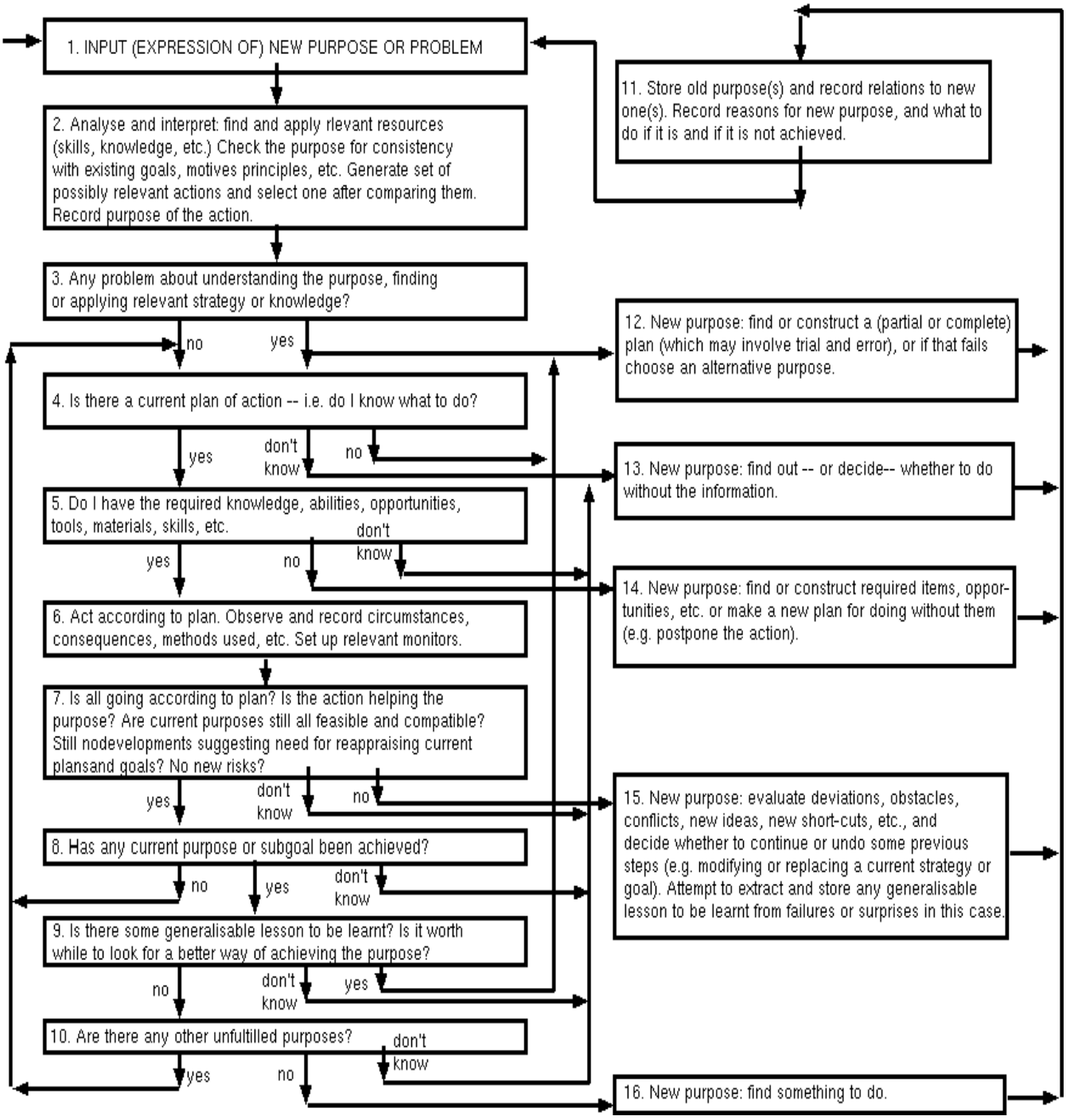
a. Each box represents a sub-process which may have very great internal complexity, including many possible alternative sub-processes.

b. Because the chart has loops, the same box may be entered twice, or many times but each entry will represent a different sub-process, and this is not represented in the chart.

c. The action of monitors is not represented. They run concurrently with the processes depicted and can detect unexpected events and cause interruptions and jumps, for instance into box 8 or box 15.

d. The chart does not indicate ways in which a path may have to be re-traced. For instance, while executing some plan one may reach box 5, then find that a required object is missing, which leads, via boxes 14 and I I, back to the top of the chart. The sub-process of box I I, namely recording the reason why the object is wanted and what is to be done with it when found (compare what was said above about the process-purpose index), ensures that when the object is later found or constructed there is a jump back to where the system was in the original process (box 5), so that it can continue, unless an

embedded entry in box 9, has led to a revision of plans (and the process-purpose index) in the meantime. Further, if the object cannot be found, it may be necessary to go back to an even earlier phase, and perhaps choose another plan for which the missing object is not necessary. Thus, the possible jumps back to an earlier phase are not represented explicitly on the flow chart.[note 3]



6.11. Executive and deliberative sub-processes

It is perhaps worth noting the difference between the clockwise loops (on the left) and the anticlockwise loops (on the right). The former may be called *executive* loops, since they represent the kinds of processes which can occur when there is a plan of action and everything is going more or less according to plan. The anticlockwise loops may be called *deliberative* loops, since they represent the kinds of things that can happen when new planning is required, so that a question has to be answered, or an unexpected new obstacle or resource has turned up: the kinds of things which may require further intelligent deliberation and decision-making, using the agent's full resources.

The plans, or procedures, which generate uninterrupted executive processes may themselves have been stored or constructed only after previous processes of deliberation, involving many circuits round the anticlockwise loops. Even when things do not go wrong, there is always the *possibility* of dealing with difficulties and surprises, representedby the arrows going from left to right.

A system in which everything *always* worked exactly as described above would be much more efficient and rational than a human being. Nevertheless we know that human beings are often *capable* of doing the kinds of things the system can do, such as noticing unexpected obstacles and changing plans. The system does not therefore explain what people actually do; rather it generates, and thereby explains, a framework of possibilities which, for various reasons, may often not be actualised even though they would be appropriate, as in failure to recall a well-known fact or name. For reasons already mentioned, even a computing system of this kind must be fallible when it is very large.

6.12. Psychopathology

Notice that this outline of the structure of an intelligent mechanism gives enormous scope for analysis of various kinds of pathological conditions in which things go wrong. Indexes and catalogues may be destroyed or corrupted. Plans, procedures, and factual records may be destroyed or corrupted.

A spell in a peculiar environment may cause procedures and beliefs to be constructed which interfere with efficient functioning in other environments, and may be hard to erase or modify. The mechanisms which manage the purpose-process index may have faults. Monitors may fail to work normally, or else their 'something-found' messages may not reach appropriate destinations. A certain class of records may be intact, but the procedures for interpreting the symbols used may be faulty. Procedures for relating new information to the index of current processes and their purposes may be faulty. Good plans may be constructed, but mechanisms for executing them may be faulty. Alternatively, execution of available plans may proceed faultlessly, but the processes of constructing new plans may fail for one reason or another.

Various sorts of learning catered for in the above scheme may fail to occur. These are very general kinds of pathology. Other more specific kinds would require a quite different analysis.

Clearly, the task of interpreting and diagnosing pathological behaviour in such a complex system must be extremely difficult. It cannot be done without a good theory of the normal structure and functions of the system. This is why I have little faith in current methods of psychotherapy.

6.13. Conclusion: what is a mind?

This ends my sketch of the main features of a mechanism able to account for some of the main features of human thought and action, that is, able to answer a large number of questions of the form 'How is X possible?' In order to prove that such a mechanism is possible, it is necessary to design one in much more detail, filling in the form with much more content. The attempt to do this will probably

show up important kinds of hidden circularity, incompleteness or inconsistency in my description, leading to a revision of the specifications.

In order to demonstrate that this sort of mechanism provides an adequate explanation of the possibilities available to a human being, it is necessary either to analyse the specifications of the mechanisms and of the possibilities to be explained, and then prove mathematically that the mechanism does generate the required range of possibilities and nothing which it should not generate, *or else* to construct the mechanism and run it experimentally in a wide variety of circumstances to ensure that it produces an adequate variety of behaviour, with the required fine structure.

The former is likely to be well beyond the possibilities of mathematical analysis available in the foreseeable future, even though the mathematical analysis of programs and proof of their correctness is a developing discipline. In particular, it assumes that we can produce complete specifications of the possibilities to be explained, whereas one of the lessons of artificial intelligence is that attempting to design a working system often leads you to revise and extend your specifications. The experimental method may require the development of computers which have much faster processors and larger memories than at present.

Whichever approach is taken, it is necessary to have a good initial specification of the range of human abilities to be explained, and this is best achieved by combining philosophical techniques of conceptual analysis with the methods of social science and psychology.

Since each of the abilities makes use of many others, like a family of mutually recursive computer programs, there is no logical order in which they should be described: no ability is basic to the others. Further, none of them can be described completely without describing many others. This makes the task of constructing such descriptions, difficult, confusing and very frustrating.

The abilities which the above system is required to explain include:

a. The ability to perceive and have perceptual experiences
b. The ability to learn: skills, particular facts, general facts
c. The ability to think about things including things near and remote, things previously met and new possibilities
d. The ability to deliberate, decide, plan and act
e. The ability to relate a purpose to available resources
f. The ability to notice unsought-for facts
g. The ability to reason, that is, the ability to use available knowledge to extend one's knowledge the ability to construct or manipulate symbols and representations, both verbal and non verbal, for such purposes as storing or communicating information, reasoning, deliberating, guiding actions, and so on

To specify these abilities in detail is to give at least part of an answer to the question: what is a mind? or what is a human mind? The partial answer is of the form: *a mind is something which can do such and such sorts of things*. To explain these abilities, that is, to explain how a single integrated system can do all these things, is to explain how it is possible for minds to exist. This does not merely make a contribution to the scientific study of man. It also brings many old philosophical discussions about the nature of mind and its relation to the human body several steps forward. (But it need not include anything that Aristotle would have disagreed with.) In the process it is certain that many detailed problems in different branches of philosophy will be solved, rejected as confused, or brought nearer solution. The remaining chapters of this book address a few of the more detailed problems.

Endnotes

(1) An early version of this chapter appeared as *Memo 59* of the Department of Computational Logic, Edinburgh University, in 1972. A slightly revised version appeared in the *A.I.S.B. Newsletter,* February 1973. A much earlier version, called 'Chapter C' was circulated privately.

> **[[Note Added 1 May 2004**
>
> On re-reading this chapter I have become aware how much of my work over the last few decades has simply been elaboration and in some cases correction of the ideas in this chapter. Even the SimAgent toolkit, developed over the last 10 years to support work on architectures for agents with human-like capabilities has many features whose inclusion can be traced back to some of the requirements described in this chapter. The toolkit is summarised in:
>
> http://www.cs.bham.ac.uk/research/poplog/packages/simagent.html
> http://www.cs.bham.ac.uk/research/cogaff/talks/#simagent
>
> Papers and slide presentations on architectures are here:
>
> http://www.cs.bham.ac.uk/research/cogaff/
> http://www.cs.bham.ac.uk/research/cogaff/talks/ **]]**

(2) In the A.I. literature they are sometimes called *demons*.

(3) Flow-charts constitute a programming language. My remarks indicate that the language is too limited in expressive power. I never use them in my own programming, and do not teach students to use them, since careful layout in a language like LISP or POP2 (augmented with good iteration constructs) can achieve the same clarity without the same limitations.

> **[[Note added in 2001**:
>
> Two themes that are implicit in this chapter turned out to be important in later work, namely the role of real-time constraints in a fast-moving world, and the potential for a mechanism of the sort described here to get into an emotional state (See: A.Sloman and M.Croucher 'Why Robots Will have Emotions' in IJCAI 1981, available, along with other relevant papers at the cogaff web site.)
>
> The two themes are closely connected. The real-time constraints generate a need for various kinds of interrupt mechanisms, alluded to in this chapter. The potential for interrupts, which can disturb current activity is intimately connected with emotional states (in at least one of the many senses of 'emotional': some authors use the word so loosely as to cover all affective states including motives and attitudes.) **]]**

Book contents page
Next: Chapter seven
Last updated: 28 Jan 2007 (Minor re-formatting)

CHAPTER 7

INTUITION AND ANALOGICAL REASONING

The previous chapter listed varieties of information that must be represented in an intelligent system. Nothing was said about how different types of symbolism could be used for different purposes. This chapter explores some of the issues, relating them to philosophical debates about inference and reasoning.

Note:

This is a revised version of

> A. Sloman, (1971) 'Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence', in *Proceedings 2nd IJCAI (1971)* Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971, and in J.M. Nicholas, ed. *Images, Perception, and Knowledge* Dordrecht-Holland: Reidel. 1977
>
> Also available online http://www.cs.bham.ac.uk/research/cogaff/04.html#analogical
>
> See notes at end for related papers written later.

7.1. The problem

Within philosophy, there has long been a conflict between those who, like Immanuel Kant, claim that there are some modes of reasoning, or inferring, which use 'intuition', ' insight', 'apprehension of relations between structures', etc., and those who argue that the only valid methods of inference are logical, for instance the use of syllogisms and rules of predicate calculus. This dispute is relevant to problems in psychology, concerning non-verbal forms of thinking and remembering (for example, the problem whether there is such a thing as 'iconic' memory).

It is also relevant to problems about the nature of mathematics and science. For instance, many mathematicians adopt a logicist' position and argue that the only acceptable mathematical proofs are those using the formalisms and inference rules of symbolic logicians. They claim that where diagrams, or intuitively grasped models are used, these are merely of 'psychological' interest, since, although they shed light on how people arrive at valid proofs, the *real* proofs do not contain such things. According to this viewpoint, the diagrams in Euclid's *Elements* were strictly irrelevant, and would have been unnecessary had the proofs been properly formulated. (For some counter-arguments, see Mueller, 1969.)

This issue is clearly relevant to teachers of mathematics and science. Teachers who accept the logicist' position will be inclined to discourage the use of diagrams, pictures, analogies, etc., and to encourage the use of logical notations, and proofs which are valid according to the rules of prepositional and predicate logic.

Kant's theories were opposed to this logicist position, insofar as he argued that important kinds of

mathematical knowledge could be both *a priori* and *synthetic,* that is, non-empirical and non-analytic. I think he had an important insight, though it has not been possible until recently to say very clearly what it was. The issues can be clarified by discussing different kinds of symbolisms, or representations, and their roles in various kinds of reasoning. Some irrelevant metaphysical digressions can be avoided by noting that such reasoning can occur in computers, as well as in human minds.

One interpretation of what Kant was trying to say is that we sometimes, for instance in mathematical thinking, use non-verbal 'analogical' representations, and make inferences by manipulating them, instead of always using logic. His claim is that these non-logical (but not illogical) modes of thinking may be valid sources of knowledge.

This topic is closely related to current problems in artificial intelligence, for it turns out that different forms of representation may differ greatly in their computational properties.

In particular, methods of representation and inference which meet the approval of logicians will not necessarily be the best ones to use in a computer program which is to behave intelligently. Not all workers in A.I. would accept this. For example, McCarthy and Hayes (1969) argued that an intelligent computer program will need to be able to prove by methods of logic that a certain strategy will achieve its goal. They claimed that this would be an essential part of the process of decision making. I doubt whether they still hold the same views (see Hayes, 1974), but the position they once advocated is worth refuting even if they have changed their mind, since it is very close to the views of many philosophers, especially philosophers of science.

7.2. Fregean (applicative) vs analogical representations

The main point I wish to make in this chapter is that there are many different types of language, or representational system, and many different ways of making inferences by manipulating representations. The forms of inference codified by logicians are relevant only to languages of the type analysed by Gottlob Frege (see Bibliography), in which the basic method of constructing complex symbols is by applying function-signs to argument-signs. Much mathematical and logical notation, and many (though not all) of the constructions of natural languages are Fregean. For instance, a first rough Fregean analysis of 'Mary shot Tom's brother' would be something like:

```
Shot (Mary, the-brother-of (Tom))
```

where the predicate 'shot' is treated as a two-place function and 'the brother of as a one-place function. Pictures, maps, diagrams, models, and many of the representations used in computer programs are not Fregean. Some of them are 'analogical'.

This contrast between Fregean (or 'applicative') and analogical representations will be more precisely defined later. It is often referred to by people who do not know how to characterise it properly. For instance, it is sometimes assumed that analogical representations are continuous and the others discrete, or that analogical representations are essentially non-verbal (that is, that verbal languages do not use them), or that analogical representations are isomorphic with what they represent. These mistakes (which will be exposed later) also go along with a tendency to assume that digital computers cannot construct or use analogical representations. (See the writings of Pylyshyn.)

Terminology is also often confused. What I have called 'Fregean' or 'applicative' representations are sometimes called 'symbolic', linguistic', 'formal', 'prepositional', or 'verbal'.

The word 'symbolic' is unsatisfactory, since the ordinary use of 'symbols', 'symbolism' and 'symbolic'

is much more general (for example maps can be said to be symbolic, even though they are analogical). I shall use 'representation' and 'symbol' and their derivatives more or less interchangeably as very general terms, and will refer to any system of representation or symbolism as a language, as in 'the language of maps'. I shall use 'Fregean' and 'applicative' interchangeably.

One of the main aims of this chapter is to show that inferences made by manipulating non-Fregean representations may be perfectly valid. I believe this is at least part of what Kant and Intuitionist mathematicians (for example Brouwer) were trying to say.

Before developing the point in detail, I would like to stress that I am not taking sides in the dispute among psychologists who argue over whether people use 'iconic' forms of memory, and reason with images. I believe that contributions from both sides are often riddled with confusions, related to the mistakes referred to above. It is especially important to notice that the points I make about analogical representations are quite neutral on the question whether such representations occur in the mind or not. Even if they occur only on paper (for example in maps and diagrams) the point is that they can still be used in valid reasoning.

Useful discussion of these issues is impossible without careful definitions of some of the main concepts, such as 'valid', 'inference', logic', 'verbal', 'analogical', 'Fregean' (or ' applicative'). However, before attempting to be more precise, I shall present a few examples of reasoning with non-Fregean symbolisms.

7.3. Examples of analogical representations and reasoning

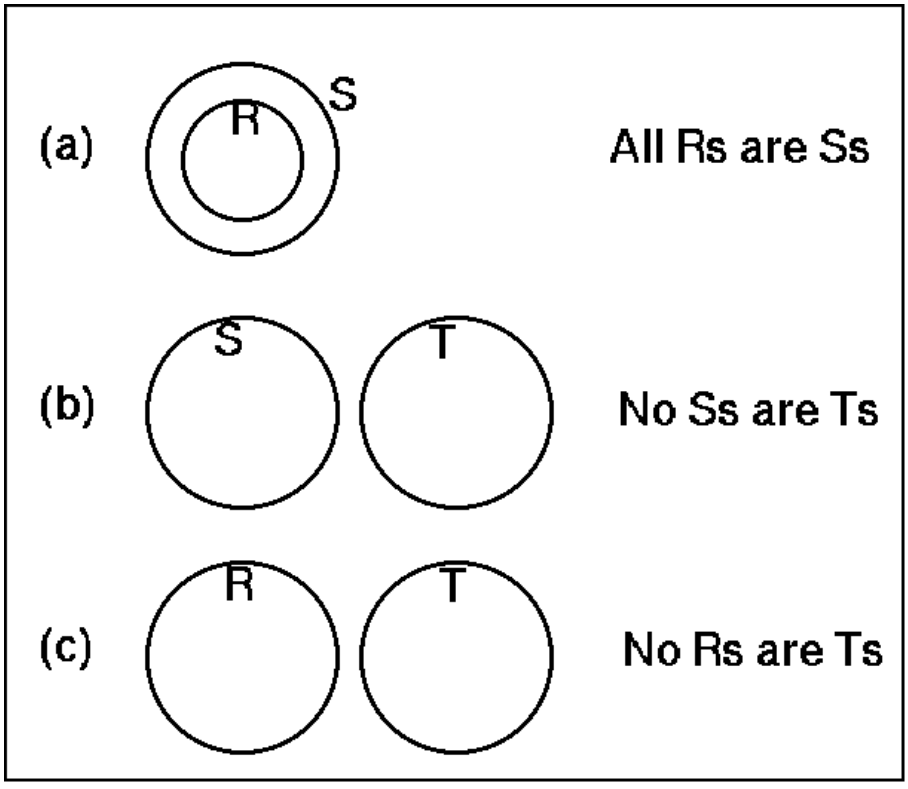


Figure 1

We can reason about set-theoretical relationships using Euler's circles. Suppose we use a circle

marked R to represent people in a certain room, a circle marked S to represent students, and a circle marked T to represent taxpayers. Then in figure 1, the three diagrams (a), (b) and (c) all represent possible states of affairs. Geometrical relations in the picture analogically represent relations between sets of people. Whether any one of them represents the way things are in the world is a contingent matter, a matter of fact. It depends, in the case of (a) and (c), on who is in the room at the time in question. This is analogous to the way in which the truth-value of a sentence depends on how things are in the world.

Whether a picture correctly depicts the world is, in each case, a contingent question which can only be answered by examining the world; but we can still discover, without examining the world, that certain *combinations* of correctness and incorrectness are necessarily ruled out. For example, no matter how things are in the world, we can use our understanding of the methods of representation employed in such diagrams to discover that it is impossible for (a) and (b) correctly to represent how things are, while (c) does not, given the stated interpretations of the diagrams. This has to do with the impossibility of creating a diagram containing (a) and (b) simultaneously, without the relation (c). *How* we discover this is not obvious, but *that* we can is.

We are also able to use our understanding of the syntax and semantics of English to tell that the following argument is valid:

> All the people in the room are students.
> No students are taxpayers.
> **Therefore:** No people in the room are taxpayers.

In both the verbal and the diagrammatic representation there are problems about possible ambiguities of reference or meaning. In both cases it is hard for people to explain why the inferences are valid. Nevertheless, we can tell that they are, and the study of such reasoning has occupied great logicians since Aristotle, leading to many logical symbolisms designed to capture the essential form of a variety of inferences.

It is worth remarking that when Euler's circles are used for this kind of reasoning, the three diagrams of figure I are normally superimposed in one diagram. This makes it harder to perceive that a method of reasoning from 'premisses' to a 'conclusion' is involved. By contrast, in verbal arguments the premisses and conclusion normally have to be formulated separately. In some of the examples which follow, I shall collapse the different representations involved into one diagram or picture, in the usual way.

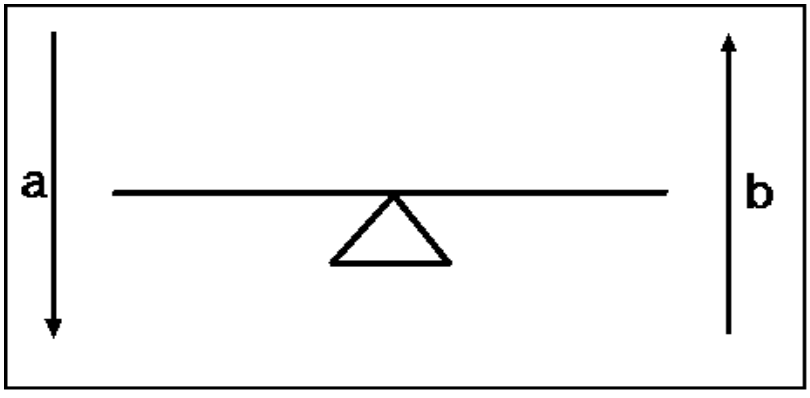


Figure 2

Here are some more examples. In figure 2, the horizontal straight line is to be interpreted as representing a rigid straight rod, pivoted at the middle on a fixed support. In figure 3 each circle

represents a rigid wheel free to rotate about a fixed axle passing through its centre, and contact between circles represents contact *without slipping* between the wheels.

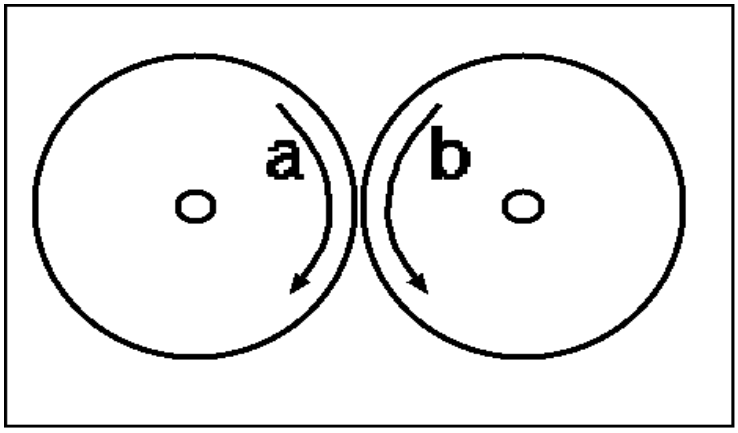


Figure 3

In both figure 2 and figure 3 the arrows represent direction of motion, (of what? how can you tell?), so the figures represent changing configurations. However, the arrows labelled (a) are to be interpreted as assumptions, or premisses, and the arrows labelled (b) are to be interpreted as conclusions, inferred from the rest of the picture. In both cases, we can consider a bit of the world depicted by the diagram and ask whether the arrow (a) correctly represents what is happening, and whether arrow (b) correctly represents what is happening. In each case, it is a contingent matter, so empirical investigation is, required to find out whether the representation is correct. (Just as empirical investigation may be used to check the truth of premisses and conclusion in a logical argument.)

However, we can tell non-empirically that it is impossible for arrow (b) to be an *incorrect* representation while arrow (a) and the rest of the diagram represents the situation correctly given the specified interpretations of the arrows, and other features of the pictures. So we can say that the inferences from (a), and the rest of the picture, to (b) is valid, in both figure 2 and figure 3. Both examples could have been replaced by two separate pictures, one containing only arrow (a) and one containing arrow (b), as in figure 1.

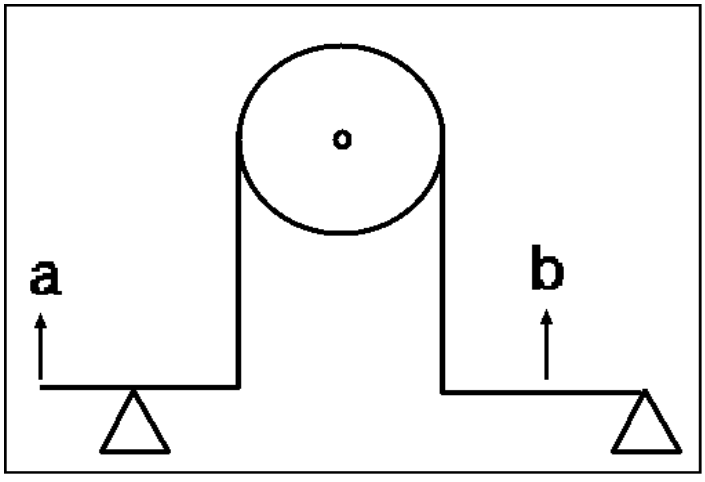


Figure 4

Far more complex examples of inferences about mechanical systems, using diagrams could be given.

Figure 4 is relatively simple. In figures 4 and 5, horizontal lines again represent rigid levers pivoted at the points indicated by small triangles. The circles represent pulleys free to rotate about their centres, but not free to move up or down or sideways.

The vertical lines, apart from arrows, represent inelastic flexible strings, and where two such lines meet a pulley on either side, this represents a string going round the pulley. Where a vertical line meets a horizontal line, this represents a string tied to a lever. As before, the arrows represent motion of the objects depicted by neighbouring picture elements. Once again, we can see that what is represented by the arrow marked (b) can be validly inferred from what is represented by the arrow marked (a) and the rest of the picture.

Where the inference is more complicated, some people may find it harder to discern the validity. In the case of logical or verbal inferences, this difficulty is dealt with by presenting a proof, in which the argument is broken down into a series of smaller, easier arguments. Something similar can be done with an argument using a diagram.

For example, figure 5 (below) is just like figure 4, except for additional arrows. The arrows marked (c), (d), (e), (f) and (g) can be taken as representing intermediate conclusions, where each can be validly inferred from the preceding one, and (c) can be inferred from (a), and (b) from (g). Using the transitivity of valid implication, we see that (b) is validly inferrable from (a). Notice that it is not always immediately obvious what can and what cannot be validly inferred. For instance, if the length of an arrow represents speed of motion, do the inferences remain valid?

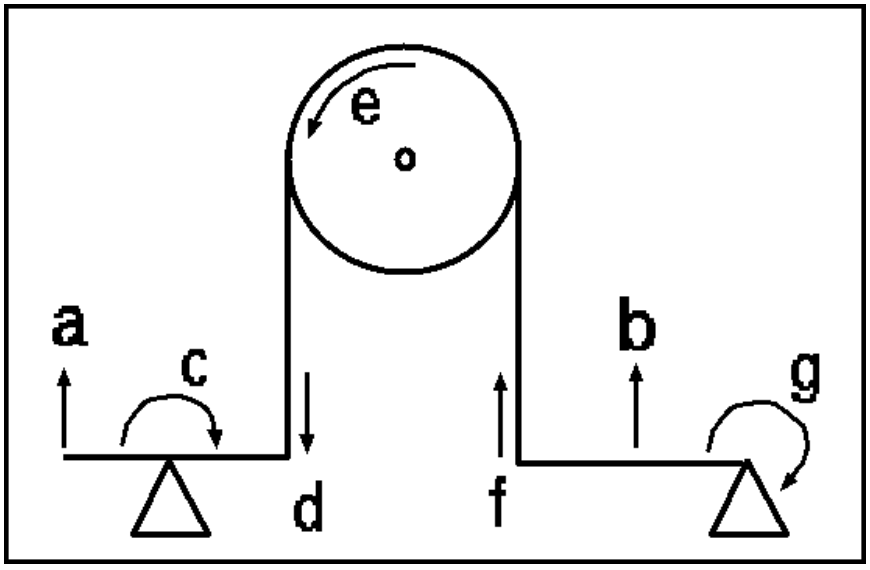


Figure 5

It is possible to give a computer program the ability to reason about mechanics problems with the aid of such diagrams. To do so would require us to formulate quite precise specifications of the significant properties and relations in the diagrams, and the rules for interpreting them, so that the computer could use these rules to check the validity of the inferences. Funt (1976) has done this in a program which makes inferences about falling, sliding and rotating objects.

I have experimented with similar programs. Making a program solve problems *intelligently* would involve giving it procedures for searching for significant paths through such diagrams, analogous to the path represented by the arrows (c) to (g), indicating a chain of causal connections relating (a) and

(b). Finding relevant paths in complex configurations would require a lot of expertise, of the sort people build up only after a lot of experience. Giving a computer the ability to acquire such expertise from experience would be a major research project given the current state of artificial intelligence. (At the time of writing a group at Edinburgh University, directed by Alan Bundy, is attempting to give a computer the ability to reason about simple mechanical problems described in English.)

I believe that our concept of a causal connection is intimately bound up with our ability to use analogical representations of physical structures and processes. This point is completely missed by those who accept David Hume's analysis of the concept of 'cause', which is, roughly, that 'A causes B' means 'A and B are instances of types of events such that it has always been found that events of the first type are followed by events of the second type'. His analysis explicitly rejects the idea that it makes sense to talk of some kind of 'inner connection' between a cause and its effect. I suspect that we talk of causes where we believe there is a representation of the process which enables the effect to be inferred from the cause using the relations in the representation. The representation need not be anything like a verbal generalisation. However, analysis of the concept 'cause' is not my current task, so I shall not pursue this here.

So far my examples of valid reasoning with analogical representations have all used diagrams. It does not matter whether the diagrams are drawn on paper, or on a blackboard, or merely imagined. Neither does it matter whether they are drawn with great precision: detailed pictorial accuracy is not necessary for the validity of examples like figure 4. It is also worth noting that instead of looking at diagrams (real or imagined), we can sometimes do this kind of reasoning while looking at the physical mechanism itself: the mechanism can function as a representation of itself, to be manipulated by attaching real or imaginary arrows, or other labels, to its parts.

So by looking at a configuration of levers, ropes and pulleys, and finding a suitable chain of potential influences in it, we can draw conclusions about the direction of motion of one part if another part is moved.

It is so easy for us to do this sort of thing, for example when we 'see' how a window catch or other simple mechanism works, that we fail to appreciate the great difficulty in explaining exactly how we do it. It requires, among other things, the ability to analyse parts of a complex configuration in such a way as to reveal the 'potential for change' in the configuration. We probably rely on the (unconscious) manipulation of analogical representations, using only procedures which implicitly represent our knowledge of the form of the world. This point is closely bound up with the issues discussed in the chapter on the aims of science, where science was characterised as a study of possibilities and their explanations.

7.4. Reasoning about possibilities

This ability to use the scenes we perceive as a representation to be in some sense manipulated in making inferences about possible actions and their effects, is central to our ability to get around in the world. For instance, the ability to select a path across a crowded room is analogous to the ability to use a map to select a route from one place to another. Using the map might be unnecessary if we could get a suitable view of the terrain from a helicopter. We frequently use things as representations of themselves!

Figure 6 gives a very simple illustration of the use of a map to make a valid inference. It is instructive in that it also shows a relationship between two representations of different sorts

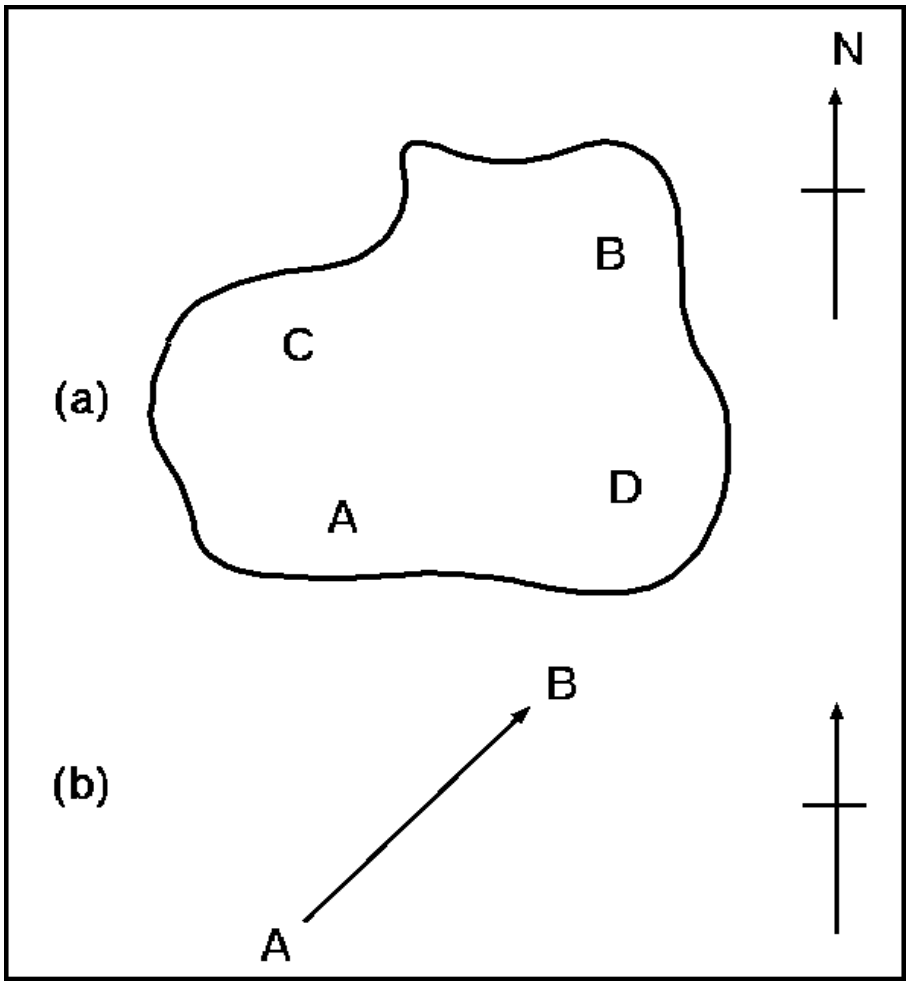


Figure 6

.

In (a) we have a map showing a few towns, marked by dots, with the usual indication of compass points. In (b) we have, not a map, but a representation of the direction (and perhaps distance) between two towns. The arrow represents a vector. Once again we can say that (b) may be validly inferred from (a), though now we have to qualify this by saying that the inference is valid only within certain limits of accuracy.

Many different uses of maps are possible. For instance, from a map showing which crops are grown in different parts of a country, and a map showing the altitude of different parts of the country, we can 'infer' a map showing which regions are both corn-producing and more than 100 feet above sea level.

When planning the layout of a room it may be useful to draw diagrams or to make flat movable cardboard cut-outs representing the objects in the room, and to use them to make inferences about the consequences of placing certain objects in certain locations. This has much in common with the use of maps.

This sort of example shows how a representation may be used to reason about what sorts of things are possible. For example, a particular arrangement of the bits of cardboard can be used to show that a certain arrangement of the objects in a room is possible. This is like the use of diagrams in chemistry to show that starting from certain molecules (for example H-H and H-H and O=O), it is possible to

derive new molecules by rearranging the parts (giving H-O-H and H-O-H).

7.5. Reasoning about arithmetic and non-geometrical relations

Reasoning with analogical representations is not restricted to geometrical or mechanical problems. Every child who learns to do arithmetic finds it useful, at times, to answer a question about addition or subtraction by using analogical representations of sets of objects. For example, a child who works out the sum of three and two by counting three fingers on one hand, two fingers of the other, then counting all the fingers previously counted, is reasoning with analogical representations. The same thing can be done with dots, as in figure 7.

An important step in mastering arithmetic and its applications is grasping that number names themselves can be used in place of dots or fingers (that is, 'one two three' followed by 'one two', matches 'one two three four five).

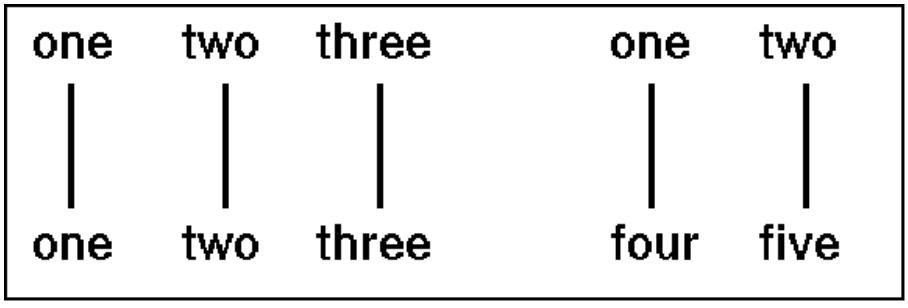


Figure 7

The diagram in Figure 7 can be used as a proof that three plus two is Five.

What is the largest possible number of persons who might have been parents of great-grandmothers of yours? What relation to you is your son's daughter's first-cousin? There are various ways you might attempt to answer this sort of question, but one of them involves drawing a fragment of a 'family tree', or possibly several family trees consistent with the problem specification. A family tree diagram is an analogical representation of a bit of the social world. Another example of an analogical representation of a rather abstract set of relationships is a chart indicating which procedures call which others in a computer program. Flow charts give analogical representations of possible processes which can occur when procedures are executed. Both sorts of diagrams can be used for making inferences about what will happen when a program is executed, or when part of a program is altered. A morse code signal is an analogical representation of a sequence of letters.

7.6. Analogical representations in computer vision

Some people working on computer vision programs have found that it is convenient to use two-dimensional arrays of numbers (representing brightness, for instance) as a representation of a visual image. (See chapter 9 for a simple example.)

Operations on the array, such as examining a set of points which lie on a straight line', or possibly marking such a set of points, make use of the fact that there is a structural relationship between the array and the retinal image. Similarly, when processing of such an image has produced evidence for a collection of lines, forming a network, as in a line drawing of a cube, then it is convenient to build up data-structures in the computer which are linked together so as to form a network of the same structure. A similar network, or possibly even the same one, can then be used to represent the three-

dimensional configuration of visible edges of surfaces in the object depicted by the line-image.

Manipulations of these networks (for example attaching labels to nodes or arcs on the network, or growing new networks to represent the 'invisible' part of the object depicted) can be viewed as processes of inference-making and problem solving, with the aid of analogical representations. It may be that something similar happens when people make sense of their visual experiences. (For more on this see the chapter on perception and Clowes, 1971, Waltz, 1975, Winston, 1975, Boden, 1977, and more recent books on computer vision.)

7.7 In the mind or on paper

It should be stressed that most of my examples are concerned with diagrams and other representations which are on paper or some other physical medium. The processes I am talking about do not have to be completely mental, though mental processes will always be involved if the representations are interpreted and used for reasoning. However, in some cases it is possible for the process to be entirely mental, when we merely imagine manipulating a diagram, instead of actually manipulating one. Reasoning of this sort may be just as valid as reasoning done with a real diagram. Unfortunately it is not at all clear what exactly does go on when people do this sort of thing, and introspective reports (for example It really is just like seeing a picture') do not really provide a basis for deciding exactly what sorts of representations are actually used. (Pylyshyn, 1973.)

Although we are still very unclear about what goes on in the minds of people, we can understand what goes on in the mind of a computer when it is building arrays or networks of symbols and manipulating them in solving some problem. By exploring such programming techniques we may hope to get a much better understanding of the sorts of theories which could account for human imaginative exercises. Our main lack at present is not data so much as ideas on how to build suitable theories.

The illustrations in the preceding sections should give at least a rough idea of what I mean by saying that sometimes valid reasoning may be done by manipulating analogical representations. Many more examples could be given. It is time now to try to formulate more precise definitions of some of the concepts used.

7.8. What is a valid inference?

Consider first an inference expressed in sentences in some natural or artificial language. There will be a set of premisses PI, P2, . .. Pn and a conclusion C, each of which is a sentence (or is expressed in a sentence). In general, whether a particular sentence says something true or something false, that is, what its truth-value is, depends partly on its form and meaning and partly on how things are in the world. So discovering the truth-value requires the application of verification procedures defined by the sentence and the semantics of the language. So each of PI, P2... Pn and C may have its truth-value determined by 'the world'. In spite of this it may be possible to discover, without examining the world, that is, without applying the usual verification procedures, that there are constraints on the possible combinations of truth-values.

In other words, by *examining* verification procedures, instead *of applying* them, we can discover that certain combinations of truth-values of statements cannot occur, no matter what the world is like. 'London is larger than Liverpool' and 'Liverpool is larger than London' cannot both be true: they are *contraries.* We can discover this by examining the semantics oflarger than'. (How is this possible?)

There are many other relationships of truth-values which can be discovered by this kind of non-empirical investigation. For instance, two statements may be incapable of both being false, in which

case they are called subcontraries by logicians.

Validity of an inference is a special case of this. Namely, the inference from PI, P2... Pn to the conclusion C is valid if and only if relationships between the statements constrain their truth-values so that it is impossible for all the premisses to be true and the conclusion false. So validity of an inference is simply a special case of the general concept of a constraint on possible sets of truth-values, namely the case where the combination

(T, T,..., T:- F)

cannot occur. So validity is a semantic notion, concerning meaning, reference, and truth or falsity, not a syntactic notion, as is sometimes supposed by logicians. They are led to this mistake by the fact that it is possible to devise syntactic tests for validity of some inferences, and indeed the search for good syntactic criteria for validity has been going on at least since the time of Aristotle,

It is an important fact about many, or perhaps all, natural languages, that syntactic criteria for some cases of validity can be found. For, by learning to use such criteria, we can avoid more elaborate investigations of the semantics of the statements involved in an inference, when we need to decide whether the inference is valid. The syntactic tests give us short-cuts, but have to be used with caution in connection with natural languages. It is not always noticed that our ability to discern the correctness of these tests depends on a prior grasp of the semantics of key words, like 'all', 'not', 'some', 'if and others, and also a grasp of the semantic role of syntactic constructions using these words. It is still an open question how ordinary people, who have not learnt logic, do grasp the meanings of these words, and how they use their understanding in assessing validity of inferences. (For further discussion see my 'Explaining logical necessity'.)

7.9. Generalising the concept of validity

Validity of inferences has been shown to be a special case of the semantic concept of a constraint on possible truth-values of a set of statements, which in turn is a special case of the general concept of a constraint on possible 'denotations' of a set of representations. This provides a basis for giving a general definition of validity.

We have seen from some of the examples of the use of analogical representations, for example, figure I and figure 2, that the question whether a particular picture, diagram or other representation correctly represents or 'denotes' a bit of the world is in general an empirical question, which involves using the appropriate interpretation rules to relate the representation and the bit of the world. (Similarly, the truth of what a sentence says is, in general, an empirical question.) We have also seen that it is sometimes possible to discover non-empirically, that is, without examining the world, that if one diagram represents a situation correctly then another must do so too. So we can easily generalise our definition of 'valid' thus:

> The inference from representations **R1, R2, . . . Rn** to the representation **Rc** is *valid,* given a specified set of interpretation rules for those representations, if it is impossible for **R1, R2 . . . Rn** all to be interpreted as representing an object or situation correctly (i.e. according to the rules) without **Rc** also representing it correctly.
>
> In this case we can say that **Rc** is jointly *entailed* by the other representations.

This definition copes straightforwardly with cases like figure I, where there are separate representations for premisses and conclusion. The other examples need to be dealt with in the obvious way by treating the single diagram as if it were a compound or two or more diagrams. For example, in figure 2 we can say that there is a 'premiss' which is the diagram with arrow (a) but not arrow (b), and a 'conclusion' which is the diagram with arrow (b) but not arrow (a).

Someone who actually *uses* a picture or diagram to reason with may modify it in the course of his reasoning, and in that case there are really several different diagrams, corresponding to the different stages in the reasoning process.

Explicitly formulating the semantic rules which justify the inference from a set of 'premiss' representations to a ' conclusion' representation, is generally quite hard. We do not normally know what rules we are using to interpret the representations we employ. Many workers in artificial intelligence have found this when attempting to write programs to analyse and interpret pictures or drawings. But the same is also true of the semantic rules of natural languages: it is hard to articulate the rules and still harder to articulate their role in justifying certain forms of inference.

In the case of artificial languages invented by logicians and mathematicians, it is possible to formulate the semantic rules, and to use them to prove the validity of some inferences expressible in the languages. In prepositional logic, symbols for conjunction '&', disjunction V', and negation '~' are often defined in terms of 'truth-tables', and by using a truth-table analysis one can demonstrate the validity of inferences using these symbols. It is easy to show, for example, that inferences of the following form are valid:

P v Q

~P

———

so: Q

(See for example Copi, *Introduction to Logic,* chapter 8.)

Similarly, in predicate logic the quantifiers ('for all x', 'for some x') may be explicitly defined by specifying certain rules of inference to which they are to conform, like the rule of 'universal instantiation' (see Copi, chapter 10). It is not nearly so easy to formulate semantic rules for words in natural languages. In fact, for some words the task would require much more than the resources of linguistics and philosophy. The semantics of colour words ('red', ' vermilion', etc.) cannot be properly specified without reference to the psychology and physiology of colour vision, for example. The principles by which we interpret pictures, diagrams and visual images may be just as hard to discover and formulate.

If the semantic or interpretative rules for a language or representational system have been articulated, it becomes possible to accompany an inference using that language with a commentary indicating why various steps are valid. A proof with such a commentary may be said to be not only valid, but also *rigorous.* So far relatively few systems are sufficiently well understood for us to be able to formulate proofs or inferences which are rigorous in this sense. Most of the forms of reasoning which we use in our thinking and communicating are not rigorous.

However, the fact that we cannot give the kind of explanatory commentary which would make our

inferences rigorous does not imply that they are not *valid.* They may be perfectly valid in the sense which I have defined. Moreover, we may know that they are valid even if we cannot articulate the reasons.

This is not to suggest that there are some inherently mysterious and inexplicable processes in our thinking. I am only saying that *so far* it has proved too difficult for us.

The use of representations to explain or demonstrate possibilities is not directly covered by the preceding discussion. However, all such cases seem to fit the following schema:

> Suppose R is a representation depicting or denoting W, where W is an object, situation or process known to be possible in the world.
>
> And suppose that Tr is a type of transformation of representations which is known (or assumed) to correspond to a really possible transformation Tw of things in the world. (See chapter 2 on the aims of science for discussion of 'really possible'.)
>
> Then, by applying Tr to R, to get a new representation, R', which is interpretable as representing an object, situation, or process W' we demonstrate that W' is possible, if the assumptions stated are true.

This seems to account for the chemical example and the use of bits of cardboard to determine a possible layout of objects in a room.

There are many problems left unsolved by all this. For instance, there are problems about the 'scope' of particular forms of inference. Are they *always* valid, or only in certain conditions? How do we discover the limits of their validity? (See Lakatos, 1976, for some relevant discussion in relation to mathematics, and Toulmin, 1953, for discussions of the use of diagrams in physics.) Does our ability to see the validity of certain inference patterns depend on our using, unconsciously, 'metalanguages' in which we formulate rules and discoveries about the languages and representations we use?

Are children developing such metalanguages at the same time as they develop overt abilities to talk, to draw and interpret pictures, etc.? Questions like these can, or course, be asked about inferences using verbal symbolisms too. (See Fodor, 1976.)

7.10. What are analogical representations?

Earlier, I introduced the idea of Fregean or applicative symbolisms, and throughout the chapter have been using the notion of an 'analogical' representation without ever having given it a precise definition. I shall try to explain what I mean by 'analogical' partly by contrasting it with 'Fregean'. I hope thereby to clarify some of the things people have had in mind in talking about 'iconic', 'non-verbal', 'intuitive', 'pictorial' symbols and modes of thinking.

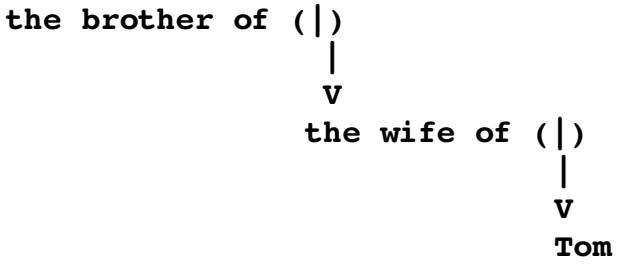
But experience has taught me that readers will project their own presuppositions onto my definitions. So I should like to stress a point which will be repeated later on, namely that there is nothing in the idea of analogical representations which requires them to be continuous (as opposed to discrete). Thus there is nothing to prevent digital computers using analogical representations. A less important source of confusion is the prejudice that analogical representations must be isomorphic with what they represent. This is by no means necessary, and I shall illustrate this with two-dimensional drawings which represent three-dimensional scenes.

The contrast between Fregean and analogical symbolisms is concerned with the ways in which

complex symbols work. In both cases complex symbols have parts which are significant, and significant relations between parts. Of course, the parts and relations are not so much determined by the physical nature of the symbol (for instance the ink marks or picture on a piece of paper) as by the way the symbol is analysed and interpreted by users. Only relative to a particular way of using the symbol or representation does it have parts and relations between parts. I shall take this for granted in what follows.

In both Fregean and analogical representations, the interpretation rules are such that what is denoted, or represented, depends not only on the meanings of the parts but also on how they are related. I shall start by saying something about how Fregean symbolisms work. Their essential feature is that all complex symbols are interpreted as representing the application of functions to arguments. Here is a simple example.

According to Frege, a phrase like 'the brother of the wife of Tom' should be analysed as having the structure:



The function 'the wife of is applied to whatever is denoted by 'Tom', producing as value some lady (if Tom is married), and the function 'the brother of is applied to her, to produce its own value (assuming Tom's wife has exactly one brother). Thus the whole expression denotes whatever happens to be the value of the last function applied.

Frege's analysis of the structures and functions of ordinary language was complex and subtle, and I have presented only a tiny fragment of it. For more details see the translations by Geach and Black, and the items by Furth and Dummett in the Bibliography. I shall not attempt to describe further details here, except to point out that he analysed predicates as functions from objects to truth-values, a notion now taken for granted in many programming languages, and he analysed quantifiers ('all', 'some', 'none', etc.) and sentential connectives ('and', 'or', 'not', etc.) also as functions.

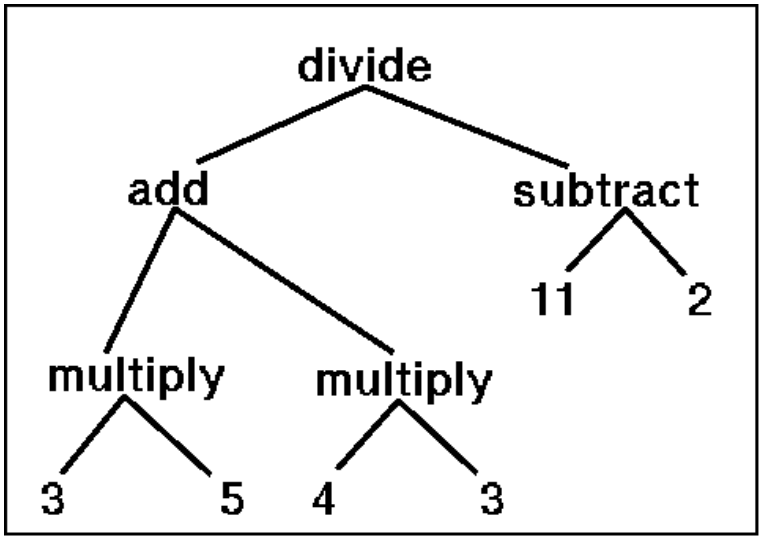
For present purposes it will suffice to notice that although the complex Fregean symbol 'the brother of the wife of Tom' has the word Tom' as a part, the thing it denotes (Tom's brother-in-law) does not have Tom as a part. The structure of a complex Fregean symbol bears no relation to the structure of what it denotes, though it can be interpreted as representing the *structure of a procedure for identifying what is denoted.* In this case, the procedure is first of all to identify whatever is denoted by 'Tom', then use the relation 'wife of' to identify someone else, then use the relation 'brother of' to identify a third object: the final value. (See also my *Tarski Frege, and the liar paradox*(1971).)

We could express this by saying that sometimes the structure of a Fregean symbol represents the structure of a 'route through the world' to the thing denoted. But this will not fit all cases. For instance, in the arithmetical expression:

$$\frac{3 \times 5 + 4 \times 3}{11 - 2}$$

it is not plausible to say that the structure of the whole thing represents a route through the world. However, given certain conventions for grouping, it does represent the structure of a rather elaborate procedure for finding the value denoted. The procedure can also be represented by a tree, as indicated below.

(Notice that in interpreting the expression this way we are using a convention about how expressions involving 'x' and '+' should be 'bracketed'.) The tree-structured procedure is executed by working up to the top of the tree from the bottom. Left-right ordering of components does not signify a temporal ordering in which the sums should be done. In some sense we can say that the sub-expressions, for example, 1 I', denote aspects of the procedure. But they do not denote *parts* of what is denoted by the whole thing. An arithmetical expression denoting the number three may contain a symbol denoting the number eleven, but that does not imply that the number eleven is in any sense part of the number three.



Representing an arithmetic expression as a tree.

By contrast, analogical representations have parts which denote parts of what they represent. Moreover, some properties of, and relations between, the parts of the representation represent properties of and relations between parts of the thing denoted.

So, unlike a Fregean symbol, an analogical representation has a structure which gives information about the structure of the thing denoted, depicted or represented.

This, then, is my definition of 'analogical'. It is important to note that not ALL the properties and relations in an analogical representation need be significant. For instance, in a diagram the colour of the lines, their thickness, the chemical properties of the paint used, and so on, need not be meaningful. In a map (for instance maps of the London underground railway system) there will often be lines whose precise lengths and orientations do not represent lengths or orientations of things in the world: only topological relations (order and connectivity) are represented. This may be because a map depicting more of the structure of the relevant bit of the world would be less convenient to use. (Why?)

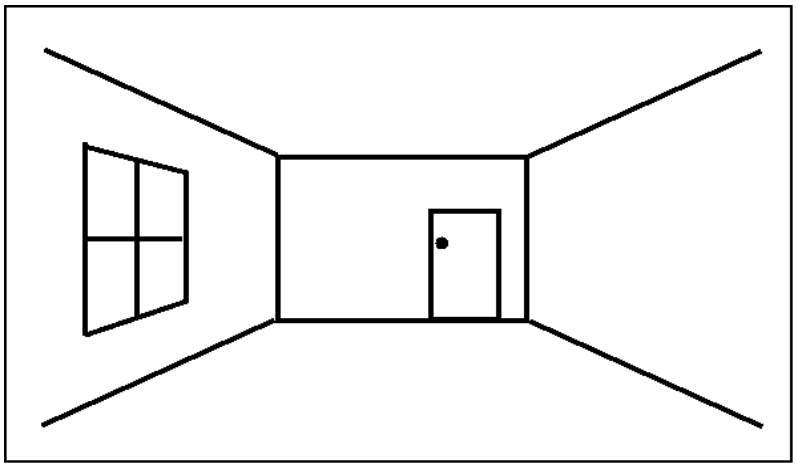


Figure 8

Further, the interpretation rules (semantic rules) need not require that properties and relations within the representation must always represent the same properties and relations of parts of what is represented. The interpretation procedures may be highly context-sensitive. For example, lines of the same length in the scene may be depicted by lines of different lengths in the picture. In figure 8, below, distances, or lengths, in the picture represent distances in the scene in a complex context-sensitive way. Further, lines of the same length in the picture may depict different lengths in the scene. Moreover, the relation 'above', in the picture, may represent the relation 'above', or 'further', or 'nearer', or 'further and higher', depending on whether bits of floor, wall, or ceiling are involved. This is connected with the fact that parts of an analogical representation may be highly ambiguous if considered in their own right. Only in the context of other parts is the ambiguity removed. Much work in computer vision is concerned with the problem of enabling global relations to ' resolve local ambiguities. (See bibliography references to Clowes and Waltz, and chapter 9.)

Figure 8 also brings out clearly the fact that although the structure of an analogical representation is related to the structure of what it represents, there is no requirement that the two be *isomorphic*. Indeed, they may have very different structures. In particular, Figure 8 is two dimensional but represents a three-dimensional scene, whose structure is therefore very different from that of the picture.

It should be obvious how to apply my definition of 'analogical' to the sorts of pictures and diagrams used earlier to illustrate inferences with analogical representations. However, it turns out that the precise details of how to interpret relations in a diagram are often surprisingly complicated. Trying to program a computer to do the interpreting is perhaps the best way of discovering the rules. Merely writing down theoretical analyses, you are likely to get the rules wrong. Embodying them in a program helps you to discover that they do not work.

7.11. Are natural languages Fregean?

Frege was able to apply his function-argument analysis to a wide variety of examples from German, and they transfer easily to the English equivalents. However, not all the complexity of natural language utterances is due to the application of functions to arguments. For example, we often use analogical representations either within sentences or in larger structures, like stories. The order of words, phrases, or sentences often depicts the order of things represented or denoted by the words, etc. Tom, Dick and Harry stood against the wall in that order.' 'He entered the room, saw the body, gasped, and ran out screaming.'

This shows that there is no sharp verbal/analogical or verbal/iconic distinction. A particular symbolism may include both Fregean and analogical resources.

In modern programming languages this is very clear, since there is a great deal of the usual function-application syntax often mixed up with conventions that the order in which instructions occur in a program represents the order in which they are to be executed (and doing them in a different order may produce quite different results). So programming languages, like natural languages, are partly Fregean and partly analogical. This is true even of a logic programming language like Prolog.

But the Fregean/analogical distinction does not exhaust the variety of important kinds of symbolising relations. For example, in a program a symbol may occur which is merely a label' its sole function is to make it easy for other parts of the program to refer to this bit, so that it does not depict either a part of something represented by the whole program nor a thing which is the argument to which a function is applied. Elsewhere in the program may be an instruction to jump to the location specified by this label. The occurrence of such 'jump' instructions can badly upset the correspondence between order of instructions in the program and the time order of events in which the instructions are executed, making programs hard to understand and modify.

The kind of self-referring metalinguistic role of labels in a computer program is clearly something different from the kinds of representation I have called Fregean and analogical.

Natural languages also use self-reference, for instance when the expressions 'the former' and 'the latter' direct attention to order of phrases in a text. They have many other devices which do not fit neatly into these two categories. For example, it is not easy to give a Fregean analysis of adverbial phrases ('He came into the room, singing, leaning heavily on a stick, and dragging the sofa behind him'). So I am not claiming that I have given anything like a complete survey of types of representation. I doubt whether such a thing is possible: for one aspect of human creativity is the invention of new sorts of symbolisms.

One conclusion which may be drawn from all this is that neurophysiologists, psychologists, and popular science journalists who take seriously the idea that one half of the human brain deals with verbal skills and the other half with pictorial and other non-verbal skills are simply showing how naive they are about verbal and non-verbal symbolisms. Presumably, when they learn that besides Fregean and analogical symbolisms there are other sorts, they will have to find a way of dividing the brain into more than two major portions. As for how we deal with combined uses of the two sorts of symbolisms, no doubt it will prove necessary to find a bit of the brain whose function is to integrate the other bits! (Programmers know that there need not be a localised bit of the computer which corresponds to sub-abilities of a complex program.)

7. 12. Comparing Fregean and analogical representations

Philosophers of science who acknowledge that scientists and mathematicians often use diagrams, models, images, and other non-verbal representations, sometimes claim that this fact is of no philosophical importance. It is a mere empirical fact, of interest to psychologists, but not relevant to philosophical studies of what is 'rational' in scientific methods.

The implication is that the use of non-logical methods of inference, and the choice of analogical representations is an irrational, or at best non-rational, piece of behaviour. Scientists are behaving rationally only when they perform logical deductions from theories and when they use observation and experiment to discover whether certain sentences express truths or falsehoods.

Against this view I shall argue that it is sometimes quite rational to choose to use an analogical rather than a Fregean method of representation. That is, there are often good reasons for the choice, given

the purposes for which representations are used, which include storing information for future use, guiding the search for good solutions to problems, enabling new versions of previously encountered situations to be recognized, and so on. I do not claim that analogical representations are *always* best.

If one were designing a robot to be a scientist, or more generally to play the role of a person, it would be advisable, for some purposes, to program the robot to store information in an analogical representation, and to perform inferences by manipulating analogical representations. (See Funt 1977 for a description of a program which solves mechanics problems with the aid of analogical representations.) So it is not merely an empirical fact that people do this too. Of course, neither people nor robots could possibly function with only analogical representations. Any intelligent system will have to use a wide variety of different types of representation and different types of reasoning strategies. But how can we decide which ones to use for which purposes? There are no simple answers.

Fregean systems have the great advantage that the structure (syntax) of the expressive medium does not constrain the variety of configurations which can be described or represented. So the same general rules of formation, denotation and inference can apply to Fregean languages dealing with a very wide range of domains. The formula **P(a,b,c)**, or its English variants, like **'a is P to b and c'**, can be used for applying a predicate to three arguments no matter what kind of predicate it is, nor what sorts of things are referred to by the argument symbols.

The following assertions use the same Fregean structure despite being concerned with quite different domains:

```
Between(London, Brighton, Cambridge)

Greater-by (three, twelve, nine)

Joins(coupling, truck 1, truck 2)
```

Contrast the difficulty (or impossibility) of devising a single two-dimensional analogical system adequate for representing chemical, musical, social, and mechanical processes. Fregean systems make it possible to think about very complex states of affairs involving many different kinds of objects and relations at once. For each type of property or relation a new symbol can be introduced as a predicate (that is, a function which, when applied to objects as arguments, yields the result TRUE or the result FALSE). The syntax for making assertions or formulating questions using all these different symbols is the same. There is no need to invent new arrangements of the symbols to cope with a new kind of domain.

The price of this topic-neutrality, or generality, is that it becomes hard to invent procedures for dealing efficiently with specific problems. Very often, searching for the solution to a problem is a matter of searching for a combination of symbols representing something with desired properties. For instance it may be a search for a plan of action which will achieve some goal, or a search for a representation of an arrangement of objects in a room, or a search for a representation of a route between two places which is shorter than alternative routes. For a frequently encountered class of problems it may be advantageous to use a more specialised representation, richer in problem-solving power than a Fregean symbolism.

What makes one representation better than another? To say that it is easier for humans, or that people are more familiar with it is not to give an explanation. An adequate explanation must analyse the structure of the symbolism and show its relationship to the purposes for which it is used, the context of use, and the problems generated by its use. This is often very hard to do, since it is hard to become conscious of the ways we are using symbols. I shall try, in the rest of this section, to give a brief

indication of the sort of analysis that is required.

A method of representation may possess problem-solving power, relative to a domain, for a number of different reasons.

a. It may have a syntax which makes it impossible to waste time exploring unfruitful combinations of symbols.

b. It may permit transformations which are significantly related to transformations in what is denoted, so that sets of possibilities can be explored exhaustively and economically.

c. It may provide a useful 'addressing' structure, so that mutually relevant items of information are located in the representation in such a way that it is easy (using appropriate procedures) to access one of them starting from the other.

d. It may provide an economic use of space, so that there's lots of room for adding new information or building temporary representations while exploring possible ways of solving a problem. Economy in use of space may also reduce the time taken to search for what is needed.

e. The representation may make it easy to alter or add to information stored, as new facts are learnt or old information is found to be mistaken or no longer necessary.

f. The system used may facilitate comparisons of two representations, to find out whether they represent the same thing, and, if not, how exactly they differ.

g. The representation may facilitate the process of ' debugging', that is tracking down the source of the difficulty when use of the representation leads to errors or disappointments.

h. The representation may allow similar methods of inference< and problem-solving to be used in more than one domain, so that solutions to problems in one domain generate solutions to problems in another domain.

These form just a subset of the problems about adequacy of representations which have had to be faced by people working in artificial intelligence. (See Hayes, 1974, Bobrow, 1975, Minsky, 1975.) The subject is still in its infancy, and criteria for adequacy of representations are only beginning to be formulated. The sorts of issues which arise can be illustrated by the following list of properties of analogical representations which often make them useful:

1. There is often less risk of generating a representation which lacks a denotation. In Fregean systems, as in ordinary language, 'failure of reference' is a commonplace. That is, syntactically well-formed expressions often turn out not to be capable of denoting anything, even though they adequately express procedures for attempting to identify a referent. Examples are 'the largest prime number', 'the polygon with three sides and four corners', 'my bachelor uncle who is an only child'.

   In analogical systems it seems that a smaller proportion of well-formed representations can be uninterpretable (inconsistent). This is because the structure of the medium, or the symbolism used, permits only a limited range of configurations. Pictures of impossible objects are harder to come by than Fregean descriptions of impossible objects. This means that searches are less likely to waste time exploring blind alleys.

2. In an analogical representation, small changes in the representation (syntactic changes) are likely to correspond to small changes in what is represented (semantic changes). We are relying on this fact when we use a map to search for a short route between two towns, and

start by drawing, or imagining, a straight line joining the two towns, then try to deform the line by relatively small amounts so as to make it fit along roads on the map.

(This is not as simple a process as it sounds.) By contrast, the differences in the forms of words describing objects which differ in shape or size may not be related in magnitude to the differences in the objects. The difference between the words 'two' and 'ten', for example, is in no sense greater than the difference between 'two' and 'three', or 'nineteen' and 'twenty'. 'Circle' and 'square' are not more different in their form than 'rectangle' and 'square'. So substitution of one word for another in a description need not make a symbolic change which is usefully related to the change in meaning. In particular, this means that the notation does not provide an aid to ordering sets of possibilities so that they can be explored systematically.

3. Closely related to the previous point is the fact that constraints in a problem situation (the route cannot go through a wall, a lever cannot bend, the centres of pulleys have a fixed position) may, in an analogical representation, be easily expressed by constraints on the kinds of syntactic transformations which may be applied to the representation. Thus large numbers of possibilities do not have to be generated and then rejected after interpreting them. So 'search spaces' may be more sensibly organised.

4. Often in an analogical representation it is possible to store a great many facts about a single item in a relatively economical way. Each part of a map is related to many other parts, and this represents a similar plethora of relationships in the terrain represented. Using a map we can 'get at' all the relationships involving a particular town through a single 'access point', for example a single dot. If the same collection of relationships were stored in sentences, then for each significant place there would be many sentences referring to it, and this would normally require a large number of repeated occurrences of the name of that place.

   Sometimes there are devices for abbreviating sentences repeating a single word, by using 'and' to conjoin phrases, for example, but one could not get rid of all repetitions of place names like this. If the sentences are stored in a list of assertions, then in order to find all the facts concerning any one place it is necessary to search for all the sentences naming it. For some places it is possible to collect together all the sentences concerning them, but since such sentences will generally mention lots of other places too, we cannot collect all the facts about a place under one heading, simultaneously for all places, without an enormous amount of repetition. This problem is avoided in a map.

   The same effect as a map can be achieved in a computer data-structure by associating with all objects a set of 'pointers' to all the stored assertions about them, that is, a list of addresses at which assertions are stored in the machine. The facts do not then need to be repeated for all the objects they mention. This sort of technique can lead to the use of structures, within the computer, which include relationships representing relationships in the world. Programmers often make their programs use analogical representations because of the efficiency achieved thereby.

5. Closely related to the previous point is the fact that it is often possible in an analogical representation to represent important changes in the world by relatively simple changes in the representation. For instance, if buttons or other markers on a map represent positions of objects, then moving the buttons represents changes in the world.

   From the new configuration the new relationships between objects (which ones are near to which others, which are north of others, etc.) are as easily 'read off as before the alteration. By contrast, if instead of representing all the initial representations by location on a map, we make a lot of assertions about their relationships, then for each change of position a large

number of changes will have to be made in the stored assertions. Of course, this problem can be minimised if we have some way of recording position without doing it in terms of relations to all the other objects, for instance by storing a pair of co-ordinates (latitude and longitude). This also requires good methods for inferring relationships from such stored positional information. Notice incidentally that the use of Cartesian co-ordinates to represent position, and more generally the use of algebraic methods in geometry, involves using sets of numbers as an analogical representation for sets of locations on a line that is, order relations and size relations between numbers represent order relations and distance relations.

7.13. Conclusion

When an early version of this chapter was published in 1971, many readers thought I was trying to prove that analogical representations are always or intrinsically better than Fregean ones. That would be absurd. I have been trying to show that questions about which should be used can be discussed rationally in the light of the purposes for which they are to be used and the problems and advantages of using them. In some circumstances, analogical representations have advantages.

The problem of deciding on the relative merits of different ways of representing the same information plays a role in the development of science, even if scientists are not consciously thinking about these issues. Similarly a child must be acquiring not only new facts and skills but new ways of representing and organising its knowledge. Very little is currently known about such processes, but the attempt to design machines which learn the sorts of things which people can learn is helping to highlight some of the problems.

The issues are complicated by the fact that one type of representation can provide a medium within which to embed or 'implement' another (see Hayes, 1974). For instance, by using a suitable method of indexing statements in a Fregean language we can get the effect of an analogical representation, as I have already indicated in discussing maps. Another example is the use of two-dimensional arrays to represent two-dimensional images in a computer. There is not really any two-dimensional object accessed by the program, rather a linear chunk of the computer's memory is organised in such a way that with the aid of suitable programs the user can treat it as if it were a two dimensional configuration addressable by a pair of co-ordinates. (Actually the physical memory of the computer is not really linear but it is interpreted as a linear sequence of locations by mechanisms in the computer.)

In chapter 8 on learning about numbers, I give examples of the use of lots of linked pairs of addresses to build up data-structures which in part function as analogical representations, insofar as the order of numbers is represented by the order of symbols representing them. This is another example of one sort of representation being embedded in another.

Computer programs can be given the ability to record and analyse some of their own actions. There will generally be a limit to what a program knows about how it works, however. For instance, programs cannot normally find out whether they are running on a computer made of transistors or some other kind. Similarly, a program may be able to record, and discuss the fact that it is accessing and modifying a two-dimensional array, or moving along a linear list of some kind, without being able to tell how the array or list is actually represented in the computer. So a program could be under the illusion that it is building and manipulating things which are very like two-dimensional pictures on paper, or very like physical rows of objects, not knowing that really it is using scattered fragments of an abstract address space managed by complex storage allocation routines and accessed by procedures designed to hide the implementation details.

When such a system is asked about its own mental processes, it could well give very misleading accounts of how they work. Phenomenologically, of course, it could not but be accurate. But it would

not give accurate *explanations* of its abilities, only *descriptions* of what it does. No doubt people are in a similar position when they try to reflect on their own thinking and reasoning processes. In particular, we see that very little explanatory power can be attached to what people say about how they solve tasks set for them by experimental psychologists interested in imagery.

One moral of all this is that often a discussion of the relative merits of two kinds of representation needs to take account of how the representations are actually constructed and what sorts of procedures for using them are tacitly assumed to be available. (For further discussion, see Hayes, 1974, Sloman, 1975.)

Very many problems have been left unsolved by this discussion. In particular, it is proving quite hard to give computers the ability to perceive and to manipulate pictures and diagrams to the extent that people do. This is an indication of how little we currently understand about how we do this.

[[Notes Added 2001: It remains very hard to implement working systems with all the features described here, though many partly successful attempts have been made.

See these two books for example (both of which contain papers that are sequels to this chapter):

J. Glasgow, H. Narayanan and Chandrasekaran (Eds), *Diagrammatic Reasoning: Computational and Cognitive Perspectives,* MIT Press, 1995,

M. Anderson, B. Meyer P. Olivier (eds), *Diagrammatic Representation and Reasoning,* Springer-Verlag, 2001.

My own papers in those books are also available online

- Musings on the roles of logical and non-logical representations in intelligence (1995)
- Diagrams in the mind? (1998/2001)

I believe that we cannot hope to understand these issues independently of understanding how human vision works. Likewise, any satisfactory model of human visual capabilities must include the basis for an explanation of how visual reasoning works. Chapter 9 of this book presents some ideas but is still a long way from an adequate theory.

Also relevant are Talks 7 and 8 here:

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/

on visual reasoning and on architectural requirements for biological visual systems, as well as more recent talks in the same directory. **]]**

Book contents page
Next: Chapter 8.

Last updated: 28 Jan 2007 (minor reformatting)

CHAPTER 8

ON LEARNING ABOUT NUMBERS: PROBLEMS AND SPECULATIONS[*]

8.1. Introduction

The aim of this chapter is both methodological and tutorial. It should help to introduce readers to some computing ideas. It also includes some theoretical speculations about learning and memory. These speculations are fairly complex, yet it is clear that they are too simple-minded to be adequate accounts of how children perform their astonishing feats of learning. Many more questions will be asked than answered. And answers offered will be tentative and provisional. Unfortunately, experienced programmers will find some of the explanations below very tedious and over-simplified. I apologise to them, and hope that non-programmers will not find the same explanations too difficult!

Here is a typical conversation with a child aged between three and a half and five years.

> Adult: Can you count up to twenty?
>
> Child: One two three four five six seven eight nine ten eleven twelve thirteen fourteen fifteen seventeen eighteen twenty.
>
> A: What comes after three?
>
> C: One two three four --- four.
>
> A: What comes after eight?
>
> C: Four
>
> A: What comes after six?
>
> C: Don't know
>
> A: What comes before two?
>
> C: One
>
> A: What comes before four?
>
> C: Five
>
> A: How many fingers on my hand?
>
> C (counting fingers): One two three four five
>
> A: What's two and three?
>
> C (counting fingers): One two three four five. Five.

Does this child grasp number concepts? Perhaps there is something wrong with the question, because number concepts are not simple things which you have either grasped or not grasped?

What are number concepts? How is it possible for them to be learnt? How is it possible for them to be

used? How is it possible to discover non-empirical facts about them? I believe we are not yet able to formulate adequate answers to these questions. What follows is offered as a preliminary exploration of some of the issues.

The method illustrated below is important. Previously (in Chapter 2), I argued that a major aim of science is to find out what is possible and explain how it is possible. We all know a great deal about what it is possible for adults and children to do with numbers. So, instead of collecting facts by doing experiments on children, we can generate requirements for explanatory theories by reflecting on the fine-structure of familiar human abilities. In other words, methods of conceptual analysis, typically practised by philosophers and linguists, can be an important source of data for psychology. (Compare chapter 4.)

I am not suggesting that conceptual analysis suffices to reveal everything we would like to know about, for example, ordinary counting abilities. The claim is only that it is foolish to embark on expensive empirical investigations before making a serious and systematic effort to articulate what you already know about the subject matter.

Here are some of the questions for which answers are lacking:

- What exactly is it that a child learns in learning about numbers?
- How is it possible for different fragments of the same number-system to be mastered by different people?
- How far does learning about numbers depend on very general learning abilities, and how far are the hurdles specific to numbers?
- How is it that a child who already seems to have the knowledge to answer a question or solve a problem, is often unable to use that knowledge?
- What enables the knowledge to be accessed at some later time?
- How can a child learn new truths about what she already knows, for instance learning that two of the numbers she has learnt add up to a third number she knows, or that two different additions generate the same result, or that some procedure (e.g. adding one) can be repeated indefinitely to yield larger and larger numbers?

I shall try to show how thinking about such apparently psychological questions can lead towards new answers to old philosophical problems about the nature of numbers, thereby providing further support for the claim that academic barriers between philosophy and science are artificial, (Some implications regarding information processing architectures for intelligent systems will emerge as a side-effect.)

[[Note added January 2002

I have just discovered the fascinating book *Wild Minds: What animals really think*, by Marc Hauser (Penguin Books 2001). Chapter 3, entitled "Number juggling", discusses and compares the understanding of numbers in very young children and in other animals. Hauser comes close to asking some of the questions asked here, and includes some speculations about possible mechanisms, but does not seem to be aware of the full variety of architectures and sub-mechanisms that might explain the observed evidence.

He repeatedly stresses the important point that it is very easy to assume that the observed behaviours of animals often suggest a unique interpretation, until we start exploring possible mechanisms that might produce those behaviours. He implicitly acknowledges that such

mechanisms can be described at different levels of abstraction, not only at the level of brain physiology.

The common trap of anthropomorphism is often a product of a lack of understanding of the variety of possible information processing architectures. Some of them are explored in these recent online presentations: http://www.cs.bham.ac.uk/~axs/misc/talks/

Presentations particularly relevant to the nature of mathematical understanding include

- Talk 7: When is seeing (possibly in your mind's eye) better than deducing, for reasoning?
- Talk 27: Requirements for visual/spatial reasoning
- Talk 14: Getting meaning off the ground: symbol grounding vs symbol attachment/tethering
- Talk 36: TWO VIEWS OF CHILD AS SCIENTIST: HUMEAN AND KANTIAN
- Talk 6: Architectures for human-like agents,

]]

8.2. Philosophical slogansabout numbers

Here are some examples of philosophers' answers to the question 'What are numbers?', and related questions:

1. Numbers are non-physical mind-independent entities, existing in their own realm which is different from the world of spatial objects. (Platonists)
2. Numbers are perceivable properties of groups of objects. For example, the number three is what is *visibly* common to the two groups

   * * *

   and

   $ $ $

   (Aristotle?)
3. Numbers are mental objects, created by human mental processes. Facts about numbers are discovered by performing mental experiments. (Kant, and the Intuitionist mathematicians)
4. Numbers are sets of sets, or predicates of predicates, definable in purely logical terms. An example of this view: the number one is the set of all sets capable of being mapped bi-uniquely onto the set containing nothing but the empty set. (Frege, Russell, and other logicists)
5. Numbers are meaningless symbols manipulated according to arbitrary rules. Mathematical discoveries are merely discoveries about the properties of this game with symbols. (Formalists)
6. Numbers are implicitly defined by a collection of axioms, such as, Peano's axioms. Any collection of things satisfying these axioms can be called a set of numbers. The nature of the elements of the set is irrelevant. Mathematical discoveries about numbers are merely discoveries of logical consequences of the axioms. (Many mathematicians)

7. There is no one correct answer to the question 'what are numbers?' People play a motley of 'games' using number words and other symbols, and a full account of the nature of numbers would simply be an analysis of these games (including the activity of mathematicians) and the roles they play in our lives. (Wittgenstein: *Remarks on the Foundations of Mathematics*)

For more details, see standard texts on philosophy of mathematics. I believe that more or less articulate versions of these philosophical theories, play an important role in many psychological and educational theories about numbers. (I have formed this opinion over many years, from wide but unsystematic reading and discussion, including attendance at lectures and seminars. So I am not in a position to document the claim. I shall continue in this chapter to make remarks about psychological theories -- if the disparaging ones are untrue I'll be delighted).

All the views listed above combine elements of truth with distortions and oversimplifications. I think that Wittgenstein's answer comes closest to encompassing the truth. In his writings he formulates many problems about mathematics, which are not answered by other theories, but his own solutions seem to me to be too shallow.

In particular, the anti-mentalism, or anti-psychologism, which pervades much of his writing prevents him from discussing mental processes in any depth. So he writes as if thinking about numbers were an essentially *social* process, consistently with his conclusion in *Philosophical Investigations* that *all* rule-following is an essentially social process, dependent on the existence of a public language.

This conflicts with a computational analysis of mental processes, according to which it is perfectly possible for a non-social mechanism to contain within itself rules which it can obey, for instance, programs transmitted genetically.

Wittgenstein's position also conflicts with any sensible account of the biological evolution of mental processes in precursors *of homo sapiens*.

I am not going to try to solve all the philosophical and psychological problems about numbers in one chapter. I shall merely try to show how we can get important new insights into the problems, and perhaps take some small steps towards formulating possible answers, if we think about the mental processes and mechanisms as if they were analogous to the processes and mechanisms involved in so-called 'list-processing' computer programs. Adequate exploration of these issues has been hampered by the current separation of philosophy and psychology, and the ignorance among most philosophers and psychologists of computing ideas.

I shall not be talking about events or processes or mechanisms in the human brain. Exactly how the brain works is as irrelevant to our problems as the detailed workings of a computer are to an explanation of a computer program written in a high-level programming language. There may be creatures on other planets, or robots, whose brains are totally unlike ours in their physiological details, yet such beings could well learn about numbers, and learn the same concepts as we do, just as two computers with quite different physical components can execute the same 'high-level' programs. (Incidentally, this undermines philosophical theories which claim that mental processes are identical with brain processes. This is as inaccurate as the claim that computational processes in a computer are identical with physical processes.)

When I talk about mechanisms involved in using numbers, I am not talking about physiological mechanisms. I am talking about aspects of the way information is organised and represented, and about the kinds of symbol-manipulating processes which may be necessary for accessing and using various sorts of representations. In particular, such processes involve the following of rules, instructions, or plans, whether consciously or unconsciously.

This illustrates how the concept of 'mechanism' is extended by developments in computing.

8.3. Some assumptions about memory

Unfortunately, my speculations about mental processes will be intelligible only if I introduce some technical ideas and assumptions, already hinted at in previous chapters, especially chapter 6. If the assumptions are wrong, then quite different theories are required. At the moment, there does not seem to be any way of avoiding these assumptions, if we are trying to explain well-known facts about what people can do.

The main assumption is that we can speak of the human mind as storing information in a vast collection of locations'. They need not be spatial locations, like shelves in a library. Positions in any kind of symbolic space with appropriate mechanisms for storing and retrieving information will do. So the word location' is being used as a technical term. For instance, radio waves are often used to transmit information, different information being transmitted at different frequencies. So information could be stored in a collection of continually reverberating radio waves, with different symbols stored at different frequencies. Each possible frequency would then be a location in the sense required here.

Similarly, possible structures of a certain class of molecules could define 'addresses' in a space. Storing information at a certain address would mean attaching that information to molecules with the structure represented by the address. This could be done more or less simultaneously in many different physical places. But the information would still be stored in one symbolic place, just as a name occurs at only one symbolic location in a telephone directory, even though there may be millions of physically distinct copies of the directory containing the name. So from now on, when I talk about locations, this is neutral as to what sorts of locations they are.

I shall assume then that a mechanism is available which can store symbols in some 'space' of locations. Further, I assume that it is possible for some of the symbols to represent locations in this space. (For instance, a directory, or catalogue, can contain entries which refer to the location' of other entries, by page or section number.) Thus the space can contain information about itself.

A symbol representing a location can be called a 'pointer', or an 'address'. So the storage mechanism can be given an address and asked to produce the symbol located there. In other words, when given a pointer, it can determine what symbol is pointed at. What is pointed at may be a complex structure containing a symbol which is itself an address of some other location, that is a pointer to another symbol. (See Figure 1.) So the space may contain chains of pointers. (In more elaborate systems, the addressing may be relative to a context or mode of operation. That is, which location is represented by a given symbol may depend on the current state of the accessing sub-mechanism. Some of the flexibility of behaviour of the system may depend on such systematic changes in the 'meaning' of symbols.)

The concept of a symbolic structure containing pointers into itself, and the investigation of processes in which such things are manipulated and used for solving problems, are among the important contributions of computing science. I shall try to show how these ideas help us to think about a child's ability to count, an ability which provides the substratum for a grasp of number concepts.

The first task is to make explicit some of our commonsense knowledge about the sorts of things we can do with number words and number concepts. Note the 'can': it is possibilities we most need to explain, not laws, that is not regularities or correlations. We know relatively few non-trivial laws of human behaviour. But we know of very many human possibilities, namely, many things at least some people can do. By thinking about possible mechanisms underlying fairly common abilities we can reveal the poverty of most philosophical and psychological theories about the nature of mathematical

concepts and knowledge. These theories do not account for the fine-structure of what we all know. All this illustrates methodological points made in chapter 2 and chapter 3.

8.4. Some facts to be explained

Reflecting on even the simplest things we know children can learn (although not all children learn all of them) shows that children can somehow cope with quite complex problems of storing, using and manipulating symbols, that is, computational problems. Some of these problems are common to many forms of learning, others peculiar to counting.

I shall start with problems involved in learning number words. These problems are common to all words. Next, there are problems concerned with the fact that number words form a *sequence* to be memorised. Some of the problems are common to many other sequences, for instance letters of the alphabet, digits in telephone numbers, the letters used to spell a word, sequences of sub-actions making up a learnt action (a dance-routine, or a method for testing faulty engines). Finally, I shall mention some problems peculiar to numbers, without offering more than tiny steps towards solutions.

There are many facts about number concepts and the ways in which they are used that I shall not attempt to analyse or explain. For instance, I shall say nothing about our ability to learn to generate an indefinitely extendable set of number names in a systematic fashion, or our ability to learn to think algebraically about numbers, for example in proving general truths about adding, subtracting, multiplying, etc., without mention of particular numbers.

In unravelling some of the hidden complexities in even the simplest abilities, I hope to give a feeling for the even greater complexities still to be explored. The intellectual tasks accomplished by ordinary children in apparently simple activities are comparable in complexity to some of the mental processes of adult scientists, engineers and artists. The children merely have less knowledge to build on.

If children have these impressive powers, why don't they use them to learn about arithmetic, reading, music, painting, and so on, without formal schooling, just as they learn to walk, talk and manipulate objects without formal schooling? Perhaps the answer is that despite all the variations in parental behaviour and home environment, nearly all children are placed in situations where learning to talk, walk, etc. are essential for them to achieve things they are highly motivated to do (like eating and interacting with other people), and moreover there are well-structured opportunities for them to learn, even though they learn things at different rates and in different orders. By contrast, very few parents and teachers are able to provide similarly highly structured and highly motivating situations to generate learning about reading, writing, mathematics, science, music, history, etc. One of the difficulties of investigating such issues without good theories of learning is that there are so many different factors which can make a difference in subtle ways. (Selfe 1977 presents relevant evidence in the drawings of an autistic child.)

8.5. Knowing number words

How is it possible for a child to learn to recognise sounds, like 'one', 'two', 'three', etc.? A simple-minded answer is that repeated exposure causes the sounds to be stored so that they can be reproduced and new occurrences recognised by matching them with stored ones. Immediately all sorts of questions need to be asked. In what form is the sound represented, that is, what exactly is stored? Is the sound analysed into recognisable fragments, such as phonemes? How are they recognised? Is some symbolic description of the sound stored, for example a description of the components of the sound? How does the child cope with variations in the sound? For instance, we may hear the same word produced by different people, or by the same person with different intonation contours or

different pronunciations.

Does the stored description cope with all variations by making use of relatively abstract specifications (whatever that means)? Or does the child store different descriptions corresponding to different ways the sound may be uttered? In the former case, how does the child learn to use descriptions with sufficient generality, and in the latter case how does she represent the fact that the different descriptions are of *the same* word?

Or is some other method used to cope with variations, such as storing a specific description (a description of a 'prototype' or 'template'), and using a flexible matching procedure so that things not quite like it will match anyway? (This kind of 'sloppy matching' is often useful in computer programs.) Or, as Kant suggested in his discussion of schemata, do we cope with variations by using rules or procedures for synthesis and analysis rather than stored templates or descriptions? For instance, a rule which says 'count the number of consecutive occurrences of "ho" in an utterance and if the result is above two then call it a laugh' can enable one to recognize laughs' of very varied lengths. (I am not suggesting, and neither was Kant, that these matching and testing processes are conscious. In any case, we know so little about the difference between what we are and are not conscious of, that we cannot draw any useful conclusions from the fact that they are mostly unconscious processes.)

For a brief introduction to further complexities of recognition, see chapter 9. The artificial intelligence literature takes the topic much further.

8.6. Problems of very large knowledge stores

We know that children learn many things. It is arguable that, using any reasonable method of counting facts, they learn millions, or at least hundreds of thousands of facts about possible appearances of things, about sounds, about possible movements, etc., in the first year or two. Given that there is a vast store of known items in a child's mind when she hears a word, what sort of process can quickly decide (not necessarily consciously) whether the word just heard matches something previously stored? Clearly a linear search through a list is out of the question, unless human minds have mechanisms which can work at far greater speeds than computers, which seems very unlikely. When listening to something quite novel, how does one decide, without searching the whole collection of stored items, that the sound just heard *does not* match anything already known?

These problems can be dealt with if the child not only stores items, but also builds an appropriate index to what it knows. For instance we use alphabetically ordered indexes to help us search books, libraries, department-stores, etc. (How? Think about how you might teach achild or a computer to use an alphabetic ordering to avoid a complete linear search.) What sorts of indexing techniques do children use, and how are they able to use them? Are we born with some sophisticated indexing strategies? Is it possible that children unconsciously use some kind of ordered set of symbols, like an alphabet, and build 'alphabetically' ordered or tree-structured catalogues of what they know, to minimise searches?

Librarians and computer scientists do not find it easy to design good methods of cataloguing and indexing. Children must be much more sophisticated, although unconsciously.

Why don't we (and children) learn things permanently as soon as we hear them? Why is repeated hearing sometimes needed for learning? One popular answer is that memory uses probabilistic mechanisms, and that repeated exposure to an item increases the probability of its being retrieved later. How this happens is rarely explained. In any case, it does not seem to be consistent with the fact that not all learning requires repetition. If someone tells you that he plans to leave you a fortune, you

will probably remember it for a long time without his having to repeat it. And faces seen once for a short time are often recognized long after, even if nothing else is recalled about the context in which they were first seen, though not all shapes are so easily remembered. So we do have some abilities to store things quickly and permanently: why are they not applied to everything we experience?

Here is a sketch of a non-probabilistic explanation of the need for repetition in some cases: the child needs to experiment with different ways of analysing, describing, and indexing new experiences. For example, it may be necessary to experiment with different ways of describing the sounds of words, so as to develop a good way to cope with variations in the sound of a word. It may even be necessary to experiment with ways of analysing a total experience before a particular sound pattern can be noticed as a significant substructure in any experience. Many adults have already developed good ways of indexing information about likely disasters and benefactions, so that they can store important items and access them later without repetition.

A closely related problem is worth mentioning. At any moment a child's experience is rich and complex. How are some features selected to be stored? How does the child decide what is worth learning? More fundamentally, how are some aspects of the current experience selected as candidates for things to be recognised if possible? How is a chunk of sound selected from the whole stream of sounds for an attempt to find a match among known items? To say that the child selects what 'interests' her is no explanation, since she can only decide that something is interesting after it has already been recognized. (Or at least some parts or aspects of it have been recognized.) These questions are taken up again in the chapteron visual perception.

8.7. Knowledge of how to say number words

Children learn not merely to recognise familiar words, but also to say them. How is the ability to say the word represented in the child's mind? Is it a set of instructions for the appropriate muscles? Or is there some representation of how the word sounds, and a general procedure which can examine a description of a sound sequence and generate appropriate instructions for muscles? This may be compared with the difference between compiled and interpreted computerprograms.

Clearly we need some general 'interpretative' procedure in order to be able to repeat a sequence of sounds which we do not recognise, for instance when imitating someone talking a foreign language, where there is no question of simply repeating something learnt previously.

Perhaps there are good reasons, if there is no shortage of space, for storing both explicit instructions for producing the sound and the specification which allows the sound to be recognised. But this raises new problems. If the knowledge of how to say the word is represented differently from knowledge of what it sounds like, how are the two items related? In particular, how is the *appropriate* knowledge found when needed?

This is just a special case of a much more general problem about how one piece of information can have other kinds of information linked to it, or associated with it. Suppose you have managed to find in your memory something matching a .word you have just heard. How does that help you access your knowledge of how to reproducethe word yourself?

8.8. Storing associations

One might think that an answer could take the form: if two items need to be associated so that when one is found then the other will be found too, then store them in adjacent locations in the memory space, just as books on related topics are often stored in adjacent locations in a library. The trouble with this is that each stored item may have to be associated with not just one, but with very many

other items, and the associations can change with time.

For example, a child has to associate the sound of a number word not only with how to say it, but also with a method of writing it down, and a method of reading written versions of it. She may even learn to say or write it in several different languages, or several different notations for numerals. Moreover, as she learns more and more about numbers, she will have to associate lots more information with each number name, including: the fact that it is a number name, that it is a word associated with certain games (e.g. chanting things in sequence), the fact that its successor is so and so, that its predecessor is so and so, the fact that it is or it is not a prime number, the fact that it is odd, or even, its 'multiplication table', its 'addition table', and so on.

Of course, not only number names generate this problem. Many known items each have to be associated with a large and growing collection of other items. For instance, in your mind your home town will be linked to very many facts which you know about the town, such as its name, its location, its direction and distance (roughly) from major towns, its population, many of its geographical details, and so on.

So we have some new problems. First, if you cannot tell when you first learn a word, say 'three', how many further items of information are to be associated with it as a result of further learning, you cannot tell how much space to reserve in the neighbourhood of the location at which a description of the sound is stored. If too little is reserved, you'll find a limit to what you can learn about the number. But people do not seem to have such limits. (For instance, think of all the things associated with the word 'word' in your mind, i.e. all the words you know.) Is there an upper limit? Some people can learn several languages!

Of course, the need for expansion could be dealt with by moving the whole collection of linked items to a larger unoccupied space if the initially reserved space overflows. Are we to assume that children have the ability to manage storage allocation like this? For instance, do they have ways of telling which of the 'free' locations have a large enough collection of free neighbouring locations? Large enough for how many additional items? Is it possible that extra chunks of space are allocated in minimal units, as in some computer storage-allocation procedures? Perhaps people solve the problem by using an abstract symbolic space of locations, like the space of decimal numbers: this has the advantage that new neighbours can always be generated for any given location. But this merely shifts the problem, for we now have to explain how information is stored about which symbols occupy which locations!

Philosophers love to analyse the concept of rationality, and to discuss rational ways of doing things. But I have yet to hear them discuss what it means to have a 'rational' way of organising and using a massive store of knowledge, subject to the constraint that in real life decisions often need to be taken fairly quickly. Attempting to design a working system forces one to address such issues.

8.9. Controlling searches

When a recognized item *is* associated with more than one other item, and some task requires one of the associated items, then how is the *right* one found? If you hear me say 'Please write down the word three', then how do you find the relevant bit of knowledge associated with the sound 'three'? That is, how do you find the specifications for writing it as opposed to saying it, or as opposed to what its successor is, or whether it is odd or not?

Obviously this search has to be controlled by the request or question. For instance, in this case, the hearer has to find something associated not only with 'three' but also with 'write down'. How is this done? There are many techniques for this sort of thing which have been explored by computer

programmers, and some of them are quite sophisticated. Do children have the ability to perform the elaborate operations used by such programmers, or do they have special techniques not yet discovered by programmers?

The problem is compounded by the fact that having learnt about some structure, we can then learn about a larger whole containing it as a part. You probably recognise not only the individual words, but also the whole phrase here: 'three blind mice'. Which method is used for obeying an instruction like

'count to three'?

Has the whole instruction been memorised and stored (e.g. because it is encountered frequently), or is there a process by which something associated with one of the words (e.g. 'three') is found because it is also associated with one of the others, or does something much more elaborate than retrieving a stored specification go on?

Is it possible that analysis of the instruction is somehow used to generate an action-specification? Quite likely we can do both, namely analyse the instruction using general principles and recognise it as a familiar whole. So how do we, and children, decide (unconsciously) which to do?

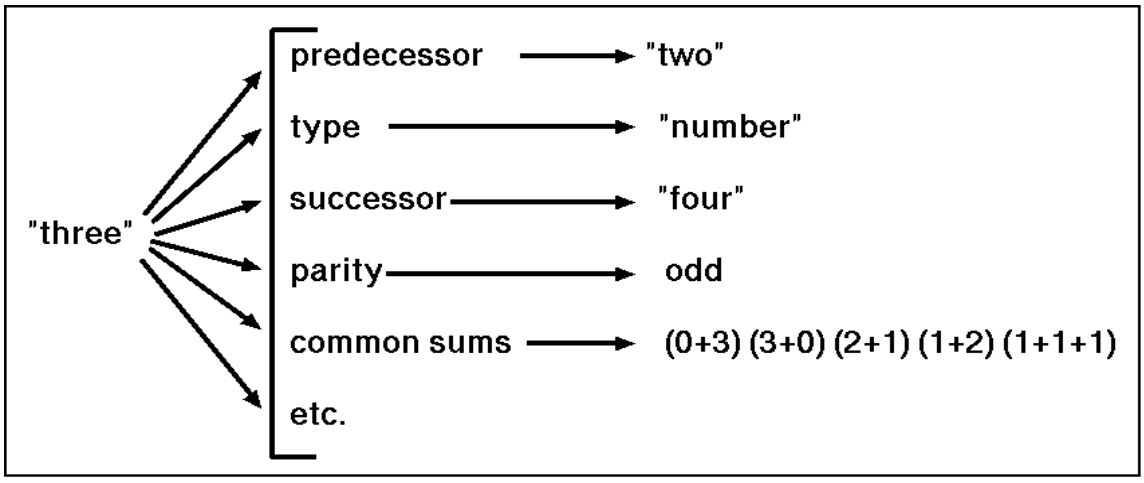
An explanatory theory, which purports to answer the questions raised here, must specify some kind of mechanism which is not merely able to hold learnt information in an efficiently accessible form, but is also capable of explaining how complex information structures are built up, how they are modified or replaced (e.g. when mistakes are discovered), and how they are used. I do not believe that educational psychologists have even the foggiest notion of what such a mechanism might be like, or what its limitations are, or what sorts of teaching strategies might interfere with its operation or facilitate learning. Some gifted teachers may have an intuitive grasp of some aspects of the mechanism, but they probably cannot articulate their implicit theories.

Computer scientists dealing with problems of managing complex collections of information in a flexible way seem to have unwittingly invented possible explanations some of which I sketch briefly below.

If we can find good theories, we may be able to do something about the large numbers of children who, for one reason or another, fail to learn so many things which might be useful or enriching to know. I believe that all normal children have the potential to learn a great deal of mathematics and other technical subjects painlessly, if only we knew how to prevent our teaching methods and attitudes to children (at home and in schools) from interfering with the learning process.

8.10. Dealing with order relations

A child can learn to answer the question 'What's after three?' How? The task is not merely to find something associated with both 'after' and 'three', since the word 'two' may also be associated with them, for three is after two. The child may also have learnt that 'five' and 'six' come after 'three' or, more specifically, that 'five' is two after 'three', 'six' is three after 'three', and so on.



Information to be stored about (associated with) "three"

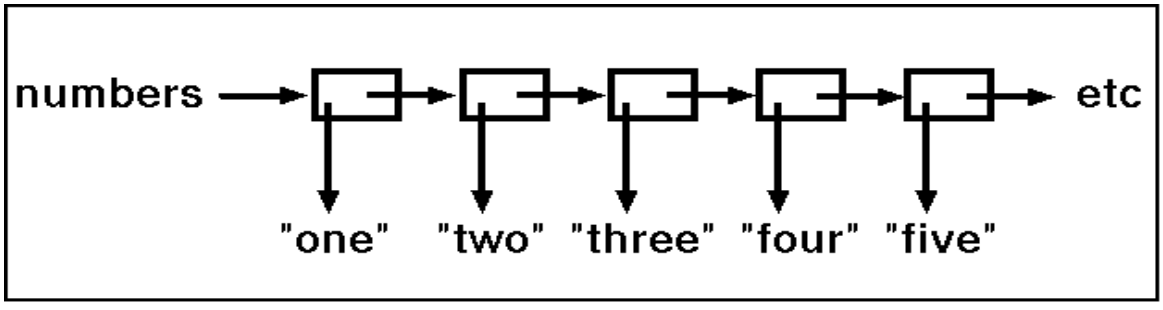
Figure 1

The problem cannot be solved by simply storing some such symbol as 'four is after three', that is, a representation of the required fact, since that would not always be the appropriate answer. For instance the question might be In the song *Ten green bottles* what comes after three?' And if the context is unambiguous it is not even necessary to mention the song explicitly in the question.

So finding the required item of information may involve analysing the question in such a way as to control the search for *relevant* links in memory.

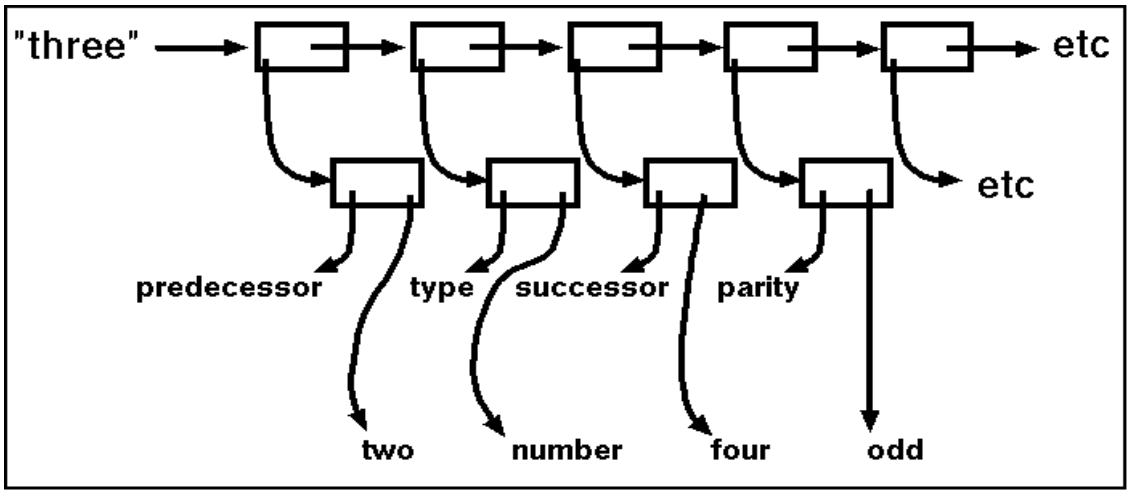
For example, it may be that each item which is associated with several others somehow has links to those others which are labelled as represented in Figure 1. It is very easy to draw diagrams like this, but not so easy to describe mechanisms which can build and use such structures. A common method used by programmers is that shown in Figure 2. A 'property-list' or 'association-list' is made up of a *chain of links* where each link contains two storage cells treated as an association by the memory mechanism, for example because they are adjacent in the memory space.

At least if the items associated with 'three' are all accessible through a linear list, then fairly obvious search procedures will enable the wanted item to be found, provided the location of the initial link of the list can be found easily.



Using a chain of two-element records to store information about order. Each link has two items of information (the main content of the link and where the next link is).

Figure 2



Using two-element links to store the information in Figure 1.

Figure 3

A chain of links may be attached to some item, for example the concept *numbers,* or the concept *three* with related items 'hung' from the chain by means of pointers giving their addresses. As Figure 3 shows, the items hung from the chain may themselves be associations, corresponding to the labelled links of Figure 1. Thus in the context of the chain attached to 'three', there is an association between 'predecessor' and 'two', whereas in a chain attached to 'four' (not shown) there would be an association between 'predecessor' and 'three'. Associations are relative to context.

Stored structures are not enough. Procedures are required for creating and finding associations in them. Such procedures are easily defined using modern programming languages. Suppose you want to search down a chain, starting from a specified link, looking for an association with a specified label (e.g. 'successor', or 'type'), because you want to find the item associated with that label in the chain. The obvious way is to see if the association pair pointed to by the given link starts with the required label, and if so to treat the second element of the pair as the desired result. Otherwise start again with the next link, whose address is in the BACK of the given link.

In a suitable programming language one could express this as Procedure-1, with the name ASSOC. (I am assuming that the subroutines FRONT and BACK when applied to a given pair produce the first thing and the second thing in the pair respectively. See Burstall *et al. Programming in POP2* for more details.)

Procedure ASSOC:

Given: initial LINK of chain, and target LABEL
Is FRONT of FRONT of LINK equal to LABEL?
If so, result is BACK of FRONT of LINK. STOP.
Otherwise, assign BACK of LINK to LINK, and restart, with LABEL as target.

Procedure-1

So ASSOC('THREE', 'TYPE') could represent the application of this procedure to a memory structure like Figure 3, with LINK starting as the first link in the chain called THREE', and LABEL having TYPE' as its value. The procedure would find a pointer to NUMBER as its result.

Similarly ASSOC('THREE', 'SUCCESSOR') would find the successor of 'three', namely 'four'. The same thing could then be used to find the successor of 'four', if that had been stored appropriately. By interleaving such searches with actions of saying what has been found, the child would have a procedure for counting, that is for reciting the numbers in their appropriate order. (More on the problems of interleaving later.)

Another way of thinking about this, is to say that information stored in a collection of structures like Figure 3, one for each known numeral, can be thought of as a sort of program for doing various things. The structure shown in Figure 2 is a much simpler program, and there is less that can be done with it. However using it as a program for counting is a simpler matter than using a collection of structures like Figure 3, since in Figure 2 all you need do in order to decide what to say next is find the link pointed to by the BACK of the current link in the chain, whereas in Figure 3 you first have to search for the 'successor' label, and then take the link it points to, and then start again from that link. We shall see later that different sorts of chains can coexist and be used for different purposes. (Figure 6)

Of course, there are many more structures and procedures that might be used for storing information about linear sequences in a computer, or in a mind. Different methods have their own advantages and disadvantages. For instance, the method of Figure 2, though simple and quick to use, has the disadvantage that when you get to the link involving 'three', there is no information stored there about items coming earlier in the chain. So using that structure makes it harder to answer questions like 'what comes before three?', though easier to answer questions like 'what comes after three?'

This is a space-time 'trade-off. Other trade-offs involved in selecting representations include efficiency vs flexibility, simplicity of structures vs simplicity of procedures, and so on. Chapter 7 discussed trade-offs between Fregean and analogical representations. Investigations of such trade-offs between different representations is central to artificial intelligence but has hitherto been absent from philosophical discussions of rationality and most psychological theorising about cognitive processes.

Proposed explanations of a child's counting abilities must do much more than explain how the child manages to recite known numbers, or how the child answers simple questions. For example, it is necessary to explain also how the representation gets built up in the first place, how new items are added, and how mistakes are corrected. One may miss out an element of the sequence, or store some elements in the wrong order. So procedures are required for inserting new links, for deleting old ones, and perhaps for changing the order of existing links, when mistakes are discovered.

A more complex procedure is required for adding a new association: it will have to get a free link (how?) and insert it at a suitable place in the chain, with its FRONT pointing to the new association and its BACK pointing to the next link in the chain, if any.

If children do anything like this to store and use associations, then how do they build up such chains, and how do they come to know the procedures for finding required associations? Perhaps the ability to learn and use chains of associations, employing procedures something like ASSOC, is inborn? Clearly not all procedures for getting at stored information are innate. For instance, children have to *learn* how to count backwards or answer 'What's before "four"?' even though they may already know the order of the numbers. The same applies to other sequences children learn. (More about such tasks later.)

8.11. Control-structures for counting-games

All this points to the old idea (compare Miller *et al.,* 1960) that human abilities have much in common with computer programs. But further reflection on familiar facts shows that programs in the most common programming languages do not provide a rich enough basis for turning this from a thin metaphor into an explanatory theory.

For instance, people can execute unrelated actions in parallel, like walking and talking. Moreover, they apparently do not require their procedures to have *built-in* tests to ensure that conditions for their operation continue to be satisfied. Nor do they require explicit instructions about what to do otherwise, like instructions in a computer program for dealing with the end of a list. All sorts of unpredictable things can halt a human action at any stage (like learning one's house is on fire) and a decision about what to do can be taken when the interruption occurs, even if no explicit provision for such a possibility is built into the plan or procedure being executed.

These points suggest that models of human competence will have to use mechanisms similar to operating systems for multi-programmed computers. For instance, an operating system can run a program, then interrupt it when some event occurs although the program itself makes no provision for interruption. Similarly, if something goes wrong with the running of the program, like an attempt to go beyond the end of a list, the program breaks down, but the operating system or interpreter running the program can decide what to do, (for example, send a message to the programmer), so that there is not a total breakdown. Ofcourse the operating system is just another program.

So the point is simply that to make the program metaphor fit human abilities we must allow not merely that one program can use another as a 'subroutine' but that some programs can execute others and control their execution, in a parallel rather than a hierarchic fashion. (For more on this, see chapters 6, 9 and 10.)

8.12. Problems of co-ordination

In counting objects, a child has to be able to generate different action sequences in parallel, keeping them in phase. Thus the process of saying number names, controlled by an internal structure, and the process of pointing in turn at objects in some group, the latter process being controlled by the external structure, have to be kept in phase. In a suitable programming language one could keep two processes in phase by means of a procedure something like the procedureCOEXECUTE

Procedure COEXECUTE:

Given: step-by-step procedures P1 and P2,

Execute a step of P1.

Execute a step of P2.

Has a stopping condition been reached?

If not, restart COEXECUTE (P1, P2).

Procedure-2

Unfortunately, this is not an acceptable model in view of the familiar fact that children (and adults doing things in parallel) sometimes get out of phase when counting and (sometimes) stop and correct themselves. This suggests that keeping the two sequences in phase is done by a third process something like an operating system which starts the processes at specified speeds, but monitors their

performance and modifies the speeds if necessary, interrupting and perhaps restarting if the sequences get out of phase. All this would be impossible with the procedure COEXECUTE. It is as if we could write programs something like the procedure RUNINSTEP.

Procedure RUNINSTEP:

Given: procedures P1 and P2,

DO (a) to (d) in parallel:

(a) repeatedly do P1

(b) repeatedly do P2

(c) observe whether (a) and (b) are getting out of step and, if they are, slow one down or speed up the other.

(d) if (a) and (b) are right out of step re-start P1 and P2

Procedure-3

The computational facilities required for this kind of thing are much more sophisticated than in COEXECUTE and are not provided in familiar programming languages.

[[Note added January 2002 The ability to monitor and modify two concurrently executed processes requires an information processing *architecture* which is not supported by the virtual machines defined by most programming languages, though it is a feature of operating systems. This is one of the sorts of tasks that might be required in a "meta-management" system, described in the presentations here: http://www.cs.bham.ac.uk/~axs/misc/tdks/ **]]**

Notice also that there is a complex perceptual task involved in deciding whether two processes are getting out of step, and children sometimes find this difficult. Not only children: try counting rotations of a wheel with no clear markings on it, while it is turning quite rapidly!

Further, the child has to be able to apply different stopping conditions for this complex parallel process, depending on what the task is. So it should be possible for yet another process to run the procedure RUNINSTEP, watching out for appropriate stopping conditions. Alternatively, the procedure could be re-defined so as to have an additional 'given', namely a stopping condition, and an extra sub-process, (e), watching out for it. For instance, when the question is 'How many buttons are there?' use 'No more buttons' as main stopping condition, whereas in response to a request 'Give me five buttons', use 'Number five reached' as main stopping condition.

I say *main* stopping condition, because other conditions may force a halt, such as getting out of phase or running out of numbers or (in the second case) running out of buttons.

How do children learn to apply the same process with different stopping conditions for different purposes? How is the intended stopping condition plugged into the process? Notice that the perceptual tasks are further complicated by the need to detect different sorts of conditions, for example, completion of the task, getting out of phase, running out of things to count, mistakes like counting the same thing twice, or leaving something out, and so on.

Some of this would be easy for a programmer using a high-level language in which a procedure (to test for the stopping condition) can be given as input to another procedure but do children have such

facilities, or do they use mechanisms more like the parallel processes with interrupt facilities described here?

I believe we know very little about how children achieve these extraordinarily complex feats. Nor do we understand what sorts of teaching strategies can help or hinder their development. My own informal observations suggest that a tremendous amount of very varied practice is required, in an environment where teachers can use a deep analysis of failures to suggest variations in games and other learning activities. This analysis can be a challenging intellectual task. How many teachers are equipped for it?

The parallel mechanisms suggested above might explain the ability to learn to watch out for new kinds of errors. For example, after learning to count stairs, where there is little chance of counting an item twice, learning to count buttons or dots requires learning to monitor for repetition and omission.

Depending on the kind(s) of programming language(s) and operating system(s) used in a child's mind, it may be easier to add a new kind of monitoring process to run in parallel with previously learnt processes than to re-organise an existing procedure so as to include new tests at appropriate places, as would be required with a conventional programming language. Probably both sorts of learning occur.

Monitoring interactions between asynchronous parallel processes may be an important source of accidental discoveries (creativity) in children and adults. For example, ongoing (unconscious) comparisons between intermediate results of two different activities may lead one to notice a relation between the two which amounts to a new technique, concept, or theory. This whole discussion is centrally relevant to the analysis of concepts like consciousness, attention, and intention. We now have a basis for a complete rejection of a major theme of Ryle's pioneering work *The Concept of Mind,* namely its refusal to accept multiple inner mental processes.

We also have a basis for beginning to explore personality differences and mental disorders relating to problems of organising and controlling several different processes. By trying to design systems involving multiple interacting processes we gain a deeper understanding of the problems and possibilities.

8.13. Interleaving two sequences

If we consider what happens when a child learns to count beyond twenty, we find that a different kind of co-ordination between two sequences is required, namely the sequence 'one, two, three . . . nine' and the sequence 'twenty, thirty, . . . ninety'. Each time one gets round to 'nine' in the first sequence one has to find one's place in the second sequence so as to locate the next item. The same is true of counting backwards from a large number. (The rules in different languages are slightly different, but the general principles are apparently the same in most.)

A programmer would find this trivial, but how does a child create this kind of interleaving in his mind? And why is there sometimes difficulty over keeping track of position in the second sequence '... fifty-eight, fifty-nine, . . . um . . . er, thirty, thirty-one . . . '? Clearly this is not a problem unique to children. We all have trouble at times with this sort of book-keeping. But how is it done when successful? And what kind of mechanism could be successful sometimes yet unsuccessful at others?

My guess is that human fallibility has nothing to do with differences between brains and computers as is often supposed, but is a direct consequence of the sheer complexity and flexibility of human abilities and knowledge, so that for example there are always too many plausible but false trails to follow. When computers are programmed to know so much they will be just as fallible, and they will have to improve themselves by the same painful and playful processes we use.

8.14. Programs as examinable structures

We have noted a number of familiar aspects of counting and other actions which suggest that compiled programs in commonly used programming languages do not provide a good model for human abilities. A further point to notice is that we not only *execute* our procedures or programs, we also *build them up* in a piecemeal fashion (as in learning to count), *modify* them when they seem inadequate, and *examine* them in order to anticipate their effects without execution. We can decide that old procedures may be relevant to new problems, we can select subsections for use in isolation from the rest, and we may even learn to run them backwards (like learning to count backwards).

This requires that besides having names and sets of instructions, procedures need to be associated with specifications of what they are for, the conditions under which they work, information about likely side-effects, etc. The child must build up a *catalogue* of his own resources. This is already done in some A.I. programs, e.g. Sussman, 1975.

Further, the instructions need to be stored in a form which is accessible not only for execution but also for analysis and modification, like inserting new steps, deleting old ones, or perhaps modifying the order of the steps, as is done in Sussman's program. Such examination and editing cannot be done to programs as they are usually stored, after compilation.

List structures in which the order of instructions is represented by labelled links rather than implicitly by position in memory would provide a form of representation meeting some of these requirements (and are already used in some programming languages). Thus, as already remarked, Figure 2 can be thought of either as a structure storing information about number names (an analogical representation of their order), or else as a program for counting. The distinction between data structures and programs has to be rejected in a system which can treat program steps as objects which are related to one another and can be changed. We now explore some consequences of this using counting as an example.

8.15. Learning to treat numbers as objects with relationships

There are several ways in which understanding of a familiar action sequence may be deficient, and may improve. One may know a sequence very well, like a poem, telephone number, the spelling of a word, or the alphabet, yet have trouble reciting it backwards. One may find it hard to start from an arbitrary position in a sequence one knows well, like saying what comes after 'K' in the alphabet, or starting a piano piece in the middle. But performance on these tasks can improve.

A child who counts very well may be unable easily to answer 'What comes after five?'. Later, he may be able to answer that question, but fail on 'What comes before six?', 'Does eight come earlier or later than five?' and 'Is three between five and eight?'. He does not know his way about the number sequence in his head, though he knows the sequence.

Further, he may understand the questions well enough to answer when the numbers have been written down before him, or can be seen on a clock. (There are problems about how this ability to use what you see to answer such questions is learnt, but I shall not go into them.)

Later, the child may learn to answer such questions in his head, and even to count backwards quickly from any position in the sequence he has memorised. How? To say the child 'internalises' his external actions (an answer I have often been given in the past) is merely to label the problem, if all that is intended is the claim that one can learn to represent in one's mind actions previously performed externally.

Moving back and forth along a chain of stored associations is quite a different matter from moving up

and down staircases or moving one's own eye or finger back and forth along a row of objects. The latter is a physical movement through space, whereas the former is movement through a set of computational states, not necessarily involving physical movement. Lack of reversibility in one case may be accounted for by physical structures, like ratchets or uni-directional motors, whereas in the other case the explanation is more likely to be lack of information. For instance in Figure 2, the link pointing to 'four' contains information about the next link, but does not contain information about the preceding link.

Learning to overcome physical impediments to reversibility need have nothing in common with learning to overcome computational problems. The child who has learned to move his eye or finger back and forth around a clock face to answer 'what comes before four' is not thereby provided with a mechanism which could somehow be used internally. At most, it provides him with a model, or analogy, which may be helpful in grasping what the task is. But how the analogy is used is totally unexplained.

8.16. Two major kinds of learning

There are at least two important kinds of development of knowledge about a previously stored structure (which may be a program), namely

a. learning new procedures for doing things with the structure
b. extending the structure so as to contain more explicit information about itself.

The former will tend to be involved in learning to do new kinds of things with the stored information, whereas the latter may simply be a matter of improving the efficiency with which old tasks are performed. But this in turn may facilitate the learning of quite new tasks which depend on rapid and skilful execution of sub-tasks using previous skills. (This bears on the debate about formal and informal methods of teaching in schools.)

A very simple procedure enables a chain like that in Figure 2 to be used to generate a sequence of actions, for example the procedure RECITE.

Procedure RECITE.

Given. a chain starting from LINK.

Utter FRONT of LINK.

If BACK of LINK isn't empty, make it LINK and restart.

Otherwise stop.

Procedure-4

Going down the chain starting from a given link is thus easy, and a procedure to find the successor of an item would use a similar principle. But answering 'What's before item X?' is more sophisticated, since on getting to a particular location (e.g. the link whose FRONT points to X), one does not find there any information about how one got there. Somehow the last item found must be stored temporarily. One method is illustrated in the procedure PREDECESSOR, as it might be defined in a programming language.

Procedure PREDECESSOR:

Given target X in chain starting at LINK:

Create temporary store TEMP, with undefined value.

Repeat the following:

If FRONT of LINK = X then result is TEMP, stop.

Otherwise, assign FRONT of LINK to TEMP and BACK of LINK to LINK and restart.

Procedure-5

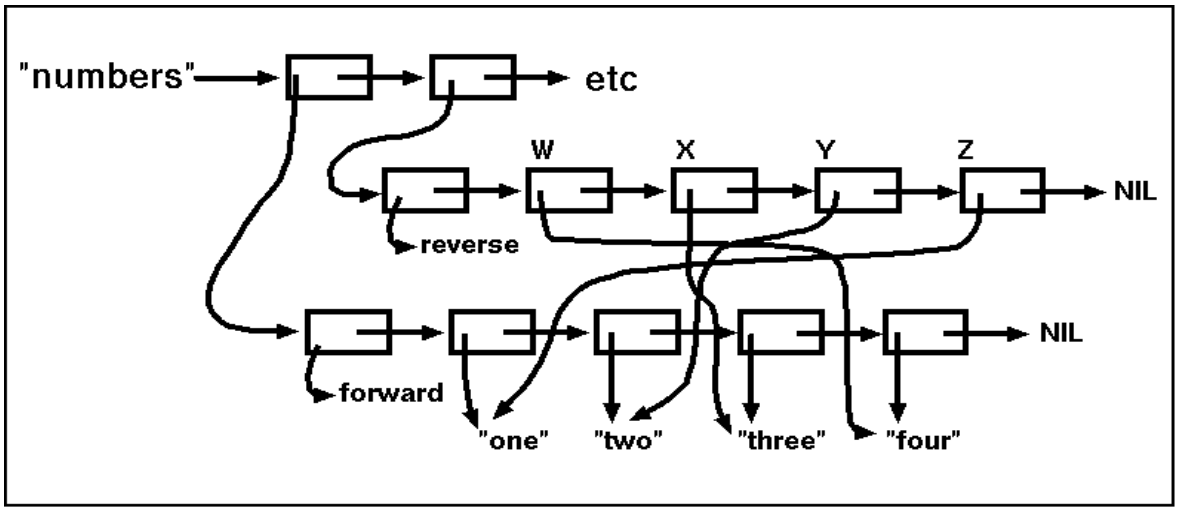
How could a child learn to create a procedure like this? Does he start with something more specialised, then somehow design a general method which will work on arbitrary chains? Perhaps it has something to do with manipulating rows of objects and other sequences outside one's head, but to say this does not give an explanation, since we do not know what mechanisms enable children to cope with external sequences, and in any case, as already remarked, chains of associations have quite different properties. For a child to see the analogy would require very powerful abilities to do abstract reasoning. Maybe the child needs themanyway, in order to learn anything.

My observations suggest that the child's learning task (at least between the ages of three and four, or later) is very different from the task of designing a procedure like Procedure-5. This is because the child is already able to remember steps he has just executed. So if he is asked to count to 'four' and does so, and then is immediately asked what came before 'four' he can answer. He does not have to allocate special-purpose temporary storage, like the 'local variable' TEMP. His problem is *to think of counting up to four* as a way of answering the question 'What comes before four?'

He does not, presumably, have a representation of the fact that if he recites some sequence he can remember the final fragment immediately after stopping. Adults have learnt this and can use it to answer a question about the predecessor of a letter of the alphabet, even if they do not have the information explicitly stored. However this technique is very tedious for reciting the whole of a learnt sequence backwards, and is useless if the process is to be done quickly. (The general ability to remember what one has just done is useful for the reasons given in chapter 6. The reasons apply to intelligent artefacts as well as to humans. This self-monitoring is not usually a built-in feature of programming systems, but there is no difficulty of principle in incorporating it.)

8.17. Making a reverse chain explicit

Merely being able to invent some procedure for finding the predecessor of a number is not good enough. For some purposes, such as counting backwards quickly, we want to be able to find the predecessor or successor of an item much more quickly than by searching down the chain of links until the item is found.



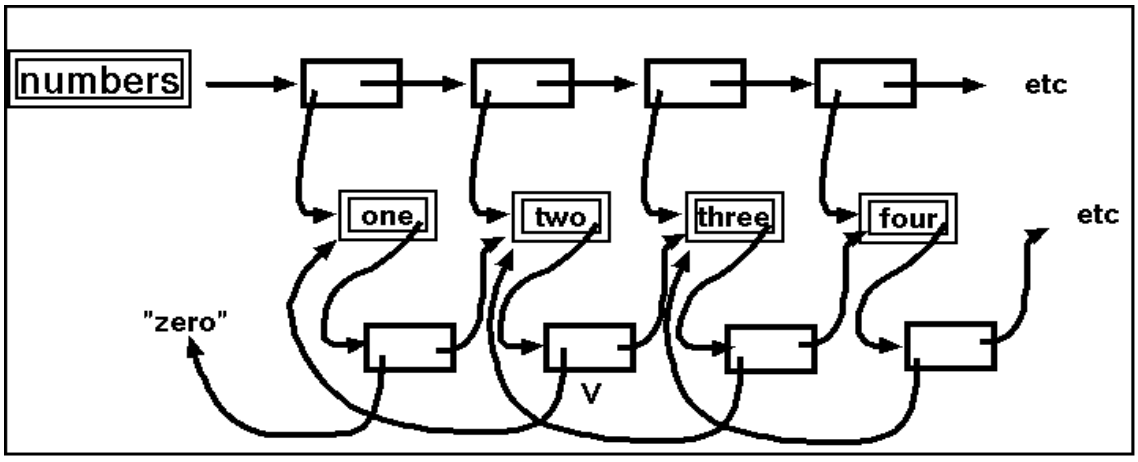
Two co-existing chains record different orders for the same set of items.

Figure 4

If a child knew only the first four numbers, then he could memorise them in both directions, building up the structure of Figure 4 instead of Figure 2. Notice that this use of two chains increases the complexity of tasks like 'Say the numbers', or 'What's after three?', since the right chain has to be found, while reducing the complexity of tasks like 'Say the numbers backwards' and 'What's before three?'. (Another example of a computational trade-off.)

However, when a longer sequence had been learnt, this method would still leave the need to search down one or other chain to find the number N in order to respond to 'What's after N?', 'What's before N?', 'Count from N', 'Count backwards from N', 'Which numbers are between N and M?', etc., for there is only one route into each chain, leading to the beginning of the chain. For instance, when one has found the link labelled 'X' in Figure 4, one knows how to get to the stored representation of 'three'. But it is not possible simply to start from the representation of 'three' to get to the links which point to it in the two chains. So we need to be able to associate with 'three' itself information about where it is in the sequence, what its predecessor is, what its successor is, and so on.

A step in this direction is shown in Figure 5, below where each number name is associated with a link which contains addresses of both the predecessor and the successor, like the link marked V, associated with 'two'. The information that the predecessor is found in the FRONT and the successor found in the BACK would be implicit in procedures used for answering questions about successors and predecessors. However, if one needed to associate much more information with each item, and did not want to be committed to having the associations permanently in a particular order, then it would be necessary to label them explicitly, using structures like those in Figure 1 and Figure 3, accessed by a general procedure like ASSOC, defined previously.



A chain represents the order of number names.
Additional links make predecessor and successor information explicit.
E.g. box V has pointers to predecessor and successor of "two".
(Double boxes are directly accessible from a central index of names.)

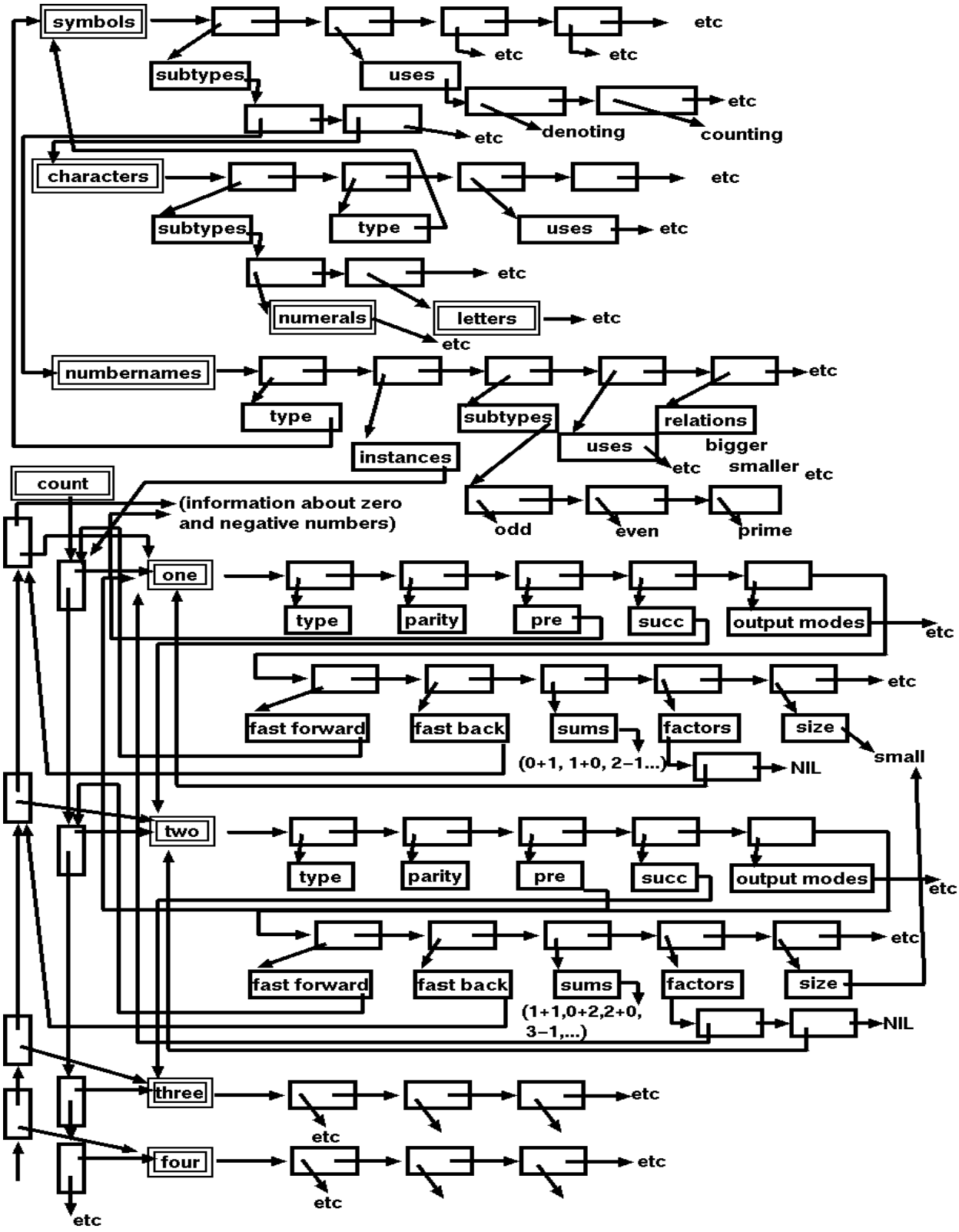
Figure 5

To cut a long story short, the result of explicitly storing lots of discoveries about each number, might be something like the network of associations in Figure 6, which is highly redundant, in the sense that a lot of the information there could in principle be derived from other information in the network. On the left there are (vertical) chains of links available for use in counting rapidly forwards or backwards, analogous to the chains in Figures 2 and 4. In addition, associated with each number is a great deal of information about it in a chain of attribute/value pairs analogous to Figure 3, except that some of the values are new sub-chains. Included in the chain hanging from each number is a pointer into the 'fast-forward' chain and a pointer into the 'fastback' chain, making it possible to count quickly forwards or backwards from that number without first having to search for the number in the relevant 'fast' chain.

In the light of the previous remarks about the need to blur the distinction between information-structures and programs, we can see how a structure like that depicted in Figure 6 can be thought of as containing several different programs embedded within it, such as programs for counting forward from various numbers, programs for counting backwards from various numbers, programs for answering questions, and so on. The different programs share common sub-structures.

The growth of this kind of network would be an example of the second type of learning, namely extending an information store to contain explicitly what was previously implicit in it. This often involves trading space for time. That is, much redundant information is stored explicitly so that it does not have to be re-computed every time it is needed. This includes information on how to do things. It seems that a great deal of early learning about numbers has this character, as well as much of the development of skill an fluency in thinking and acting.

"Progressive" educational procedures which attempt to do without any rote learning may be depriving children (or adult learners!) of opportunities to build up some structures which are useful for rapid access -- unless the old formal methods are replaced with carefully structured play situations, to achieve the same effect (which they could probably do much more effectively, since they would be more highly motivating.) Children need a lot of practice at 'finding their way about' their own data-structures.



Using chains of associations to represent what a child must learn about numbers. (Double boxes are directly accessible from a central index of names.)

Figure 6

The structure of Figure 6 may look very complex, yet using it to answer certain questions requires simpler procedures than using, say Figure 2. For, having found the link representing a number, one can then find information associated with that number by simply following forward pointers from it, for example, using a procedure like ASSOC; whereas in Figure 2 or 5, finding the predecessor and successor of a number requires using two different procedures, and each requires a search down a chain of all the numbers to start with. Of course, a structure like Figure 6 provides simple and speedy access at the cost of using up much more storage space. But in the human mind space does not seem to be in short supply!

A further problem is that each time new information is added to a chain, the increased length increases the average time for searching along the chain. So if an item in a structure like Figure 6 has a very long chain of associations, it might be preferable to replace the linear chain with a local index to avoid long searches. So, instead of 'three' being linked to a linear list of associations, it would have some kind of structured catalogue. Someone who knew a very large number of things about 'three' might find that this saved time searching for information. This would require the procedure ASSOC to be replaced by something more sophisticated, and would probably also require more space. Alternatively, by switching pointers, one could easily bring a link to the front of the chain each time the association hanging from it is used: this would ensure that most recently and most frequently used information was found first, without the help of probabilistic mechanisms, often postulated to explain such phenomena.

8.18. Some properties of structures containing pointers

Notice that in a structure like Figure 6, normal 'part-whole' constraints are violated: information about numbers is part of information about 'three', and *vice versa.* So by using pointers (addresses) we can allow structures to share each other. In a rich conceptual system circular definitions will abound. If much of our knowledge is non-hierarchic, as this suggests, then perhaps strictly cumulative educational procedures designed to achieve complete clarity at every stage are quite misguided. So also are philosophical investigations of knowledge in the tradition of Descartes, trying to show how everything can be based on completely rational chains of inference starting from self-evident, or at least minimally doubtful, premisses.

(Perhaps only trivial things can be taught without generating a great deal of confusion. Infants learning to speak experience a great deal of confusion, but this does not usually make them give up! Only later on do we teach them to give up too soon, by labelling them as 'stupid', for example, or perhaps by helping them too often when they are in difficulty.[1])'

This kind of circularity (or mutual recursion) is especially common in our mental concepts. For instance, the concept of 'belief cannot be analysed without reference to the concepts of desire and decision, and these cannot be analysed without reference to each other and the concept of belief. Yet ordinary people learn to use these concepts in their ordinary life (for instance, when they explain someone's action in terms of a belief: 'He did it because he believed I was out to get him'). We learn to use mutually recursive concepts without being at all aware of their complexity.

In my experience philosophers and psychologists tend to get very confused about how to deal with this kind of circularity, for example in discussing varieties of Behaviourism. Analogies with recursive computer programs and data-structures can help to clarify the issue. One can distinguish varieties of behaviourism according to whether they will tolerate recursion (especially mutual recursion) in their definitions of mental concepts. Ryle's book *The Concept of Mind* was more sophisticated in this respect than most other forms of behaviourism, since it implicitly allowed mutually recursive definitions of mental concepts, implying that mentalistic concepts cannot all be eliminated by analysis

in terms of dispositions to respond to stimuli. This, presumably, explains why Ryle did not see himself as a behaviourist.

The kind of structure depicted in Figure 6 does not need a separate index or catalogue specifying where to look for associations involving known items, for it acts as an index to itself, provided there are some ways of getting quickly from outside the structure to key nodes, like the cells containing 'three' and 'number'. (This might use an index, or content addressable store, or indexing tricks analogous to hash coding, for speedy access.)

The use of structures built up from linked cells and pointers like this has a number of additional interesting features, only a few of which can be mentioned here. Items can be added, deleted, or rearranged merely by changing a few addresses, without any need for advance reservation of large blocks of memory or massive shuffling around of information, as would be required if items were stored in blocks of adjacent locations (another trade-off: space against flexibility).

The same items can occur in different orders in different structures which share information (see Figure 4 for a simple example). Moreover, the order can be changed in one sequence without affecting another which shares structure with it. For instance, in Figure 4 the addresses in links W, X, Y, and Z can be changed so as to alter the order of numbers in chain labelled 'reverse' without altering the chain labelled 'forward'.

As we saw in connection with Figure 2, when the rest of the mechanism is taken for granted, a structure of the kind discussed here looks like a program for generating behaviour, but when one looks into problems of how a structure gets assembled and modified, how parts are accessed, how the different stopping conditions are applied, etc., then it looks more like an information structure used by other programs.

8.19. Conclusion

Further reflection on facts we all know reveals many gaps in the kinds of mechanisms described here. For instance, very little has been said about the *procedures* required for building, checking, modifying, and using a structure like Figure 6. Nothing has been said about the problems of perception and conception connected with the fact that counting is not applied simply to bits of the world but bits of the world individuated according to a concept (one family, five people, millions of cells but the same bit of the world counted in different ways).

I have offered no explanation of the ability to answer 'How many?' questions by recognising a visual pattern, without explicit counting. Obviously, there is a lot to be said about the development of new perceptual abilities related to numbers, for example the ability to perceive groups of three objects without counting, by matching against a structural definition, much as one recognises arches, letters and horses (see chapter 9 and Winston, 1975).

Nothing has been said about how the child discovers general and non-contingent facts about counting, such as that the order in which objects are counted does not matter, rearranging the objects does not matter, the addition or removal or an object must change the result of counting, and so on. How does a child come to grasp the fact that in principle counting can go on indefinitely, so that its stock of number names may need to be extended, or replaced by a rule with unlimited generative power?

(Philosophers' discussions of such non-empirical learning are usually so vague and abstract as to beg most of the questions. Piagetian psychologists comment on some of the achievements, but provide no means of analysing or explaining the underlying mental processes discussed here.)

I cannot explain these and many more things that even primary school children learn. I do not believe that anybody has even the beginnings of explanations for most of the things we know they can

(sometimes) do: all we find is new jargon for labelling the phenomena.

I have offered all this only as a tiny sample of the kind of exploration needed for developing our abilities to build theoretical models worth taking seriously. In the process our concept of mechanism will be extended and the superficiality of current problems, theories and experiments in psychology and educational technology will become apparent.

Philosophers have much to learn from this sort of exercise too, concerning old debates about the nature of mind, the nature of concepts and knowledge, varieties of inference, etc. Consider my short survey of answers they have given to the question 'What are numbers?' The answers do not begin to match the complexity of what a child has to grasp in learning about numbers. They do not account for the fact that number concepts are used in a variety of activities. They perhaps take the uses for granted, but make no attempt to explain how they are possible. Philosophers of the future, who have a much better grasp of what such explanations might look like, will be in a better position to formulate adequate analyses.

Similarly, when they have learnt about possible mechanisms underlying processes of inference and discovery, they will be in a better position to discuss the nature of mathematical discovery and other forms of a *priori* learning. The most that can be said at present is that it will probably prove helpful to think of mathematical discovery by analogy with a program which discovers new facts about itself by a combination of executing parts of itself and examining some of its instructions. In the process it might decide that some things could be done more quickly in a different way. Or it might discover, by analysing its own structure, that instead of executing bits of programs, it can work out their effects by reasoning about them.

More importantly, it may discover ways of generalising and extending its procedures to accomplish more tasks of the same sort, or new kinds of tasks. Programmers often discover unexpected ways of elaborating and generalising their programs, in the course of examining and using them, much as an artist learns more about what he can and should do by examining an incomplete work. A program which builds its own programs can do this too. Sussman's 'Hacker' program (1975) builds programs, and, in some cases, generalises them.

I believe that similar ideas are to be found in Piaget's writings. Computer models turn such thoughts from vague speculations to testable theories. See Young, 1976.

These sorts of second-order discoveries about one's own procedures do not fit the normal definition of 'empirical'. For example, they need not involve the use of the senses to gain information about the world. And a kind of necessity seems to be involved in the truths so discovered which is not normally thought to be compatible with empirical learning: if experience can lead us to a hypothesis can it not also produce a refutation of the hypothesis? But it seems that no experience can refute the claim that adding two lots of two things produces the same result as adding one thing to a group of three things, or the claim that there is no largest number. And the same is true of many other discoveries about properties of the procedures we use. Yet such mathematical discoveries involve a kind of exploration of possibilities which is closely analogous to empirical learning.[2]

We need a richer set of distinctions than philosophers normally employ. There is learning from sensory experience and learning from symbolic experience. The latter seems to include the processes generating what Kant called 'synthetic *a priori* knowledge'. However these processes require a great deal of further investigation. In particular, it is important to note that symbolic experiences may occur either entirely within the mind, or else may use external symbolisms, as when we use diagrams or calculations on paper. The use of our senses to examine our symbols and our procedures for manipulating them should not be confused with the use of our senses to examine the behaviour of objects in the world.

The task of designing programs which simulate these sorts of human learning to a significant extent is at the frontiers of current research in artificial intelligence. Until further progress has been made, philosophical speculation about non-empirical knowledge is likely to remain as unproductive as it has been through most of history.

The old nature-nurture (heredity-environment) controversy is transformed by this sort of enquiry. The abilities required in order to make possible the kind of learning described here, for instance the ability to construct and manipulate stored symbols, build complex networks, use them to solve problems, analyse them to discover errors, modify them, etc., all these abilities are more complex and impressive than what is actually learnt about numbers! Where do these abilities come from? Could they conceivably be learnt during infancy without presupposing equally powerful symbolic abilities to make the learning possible? Maybe the much discussed ability to learn the grammar of natural languages (cf. Chomsky, 1965) is simply a special application of this more general ability? This question cannot be discussed usefully in our present ignorance about possible learning mechanisms.

Finally a question for educationalists. What would be the impact on primary schools if intending teachers were exposed to these problems and given some experience of trying to build and use models like Figure 6 on a computer? Our experience of teaching philosophy and psychology students computing in the Cognitive Studies Programme at Sussex University, and similar experiences at other centres, such as Edinburgh University and Massachusetts Institute of Technology, suggests that it can produce a major transformation of outlook including a new respect for the achievements of children. Here is a tremendous opportunity for educational administrators and teacher-training institutions. Will they grasp it?[3]

NOTE added Oct 2001:

At the time I was writing this chapter I was aware that there were many people trying to explain learning in terms of probabilistic associations. Although this seemed to be a good description of some of the intermediate stages in learning about numbers, e.g. before a child is reliably able to recite the number sequence, probabilistic associations did not seem to characterise adequately the rich and precise grasp of structure that comes with a deep understanding of the world of numbers, including not only the ability to answer the simple sorts of questions discussed in this chapter, but also the ability to think about infinite sets (e.g. the set of even numbers or the set of prime numbers) the ability to discover new regularities in the structure, the ability to invent new, provably correct procedures e.g. for doing long multiplication or finding square roots, and so on.

Part of all this is the ability to understand the difference between the *necessary, exceptionless,* truths of number theory, such as that seven plus five equals 12, or that the cardinality of a set is independent of the order in which the elements are counted and merely *contingent* truths, such as that if you put one rabbit in an empty hutch and then later add another rabbit, and do not put any additional rabbits in the hutch there will be only two rabbits in the hutch thereafter.

In other words I was convinced that although the processes involved in learning some of the basic features of the number system, including the names for the numbers and their order, might use probabilistic mechanisms (such as the neural nets that became popular in the two decades following publication of this book), this could not be the key either to the nature of our mathematical knowledge, or to many other features of our knowledge of the world, such as understanding how a clock works, or why turning a handle can enable a door to be opened, or why it is necessary to open the door in order to go into a room. When we understand these things we do not merely understand probabilistic associations, we understand *structural relationships.*

By the early 1970's there had already been some deep work in AI investigating structure-based learning

and understanding, e.g. the papers in Minsky's 1968 collection, and Sussman. (Progress was very slow, however, because of the extremely limited speeds and memory capacities of computers of the time, but more importantly because the sheer difficulty of the problems.)

When I wrote this chapter, I was attempting to generalise some of that early work by exploring the notion that a human's ability:

a. to construct and then inspect and manipulate list structures (or similar structures found in computational virtual machines)
b. to inspect and manipulate the procedures for operating on those structures
c. to run processes in parallel, including processes observing and modifying other processes,

could explain a wider range of phenomena than mere learning of associations could.

I also suggested that if some of the list structures did not have a fixed order but were re-linked, e.g. bringing more recently accessed items closer to the front, then that could explain some of the variability in performance that others had assumed must be explained by probabilistic mechanisms.

In retrospect, it seems that a mixture of the probabilistic and deterministic approaches is required, within the study of *architectures* for complete agents: a more general study than the investigation of algorithms and representations that dominated most of the early work on AI (partly because of the dreadful limitations of speed and memory of even the most expensive and sophisticated computers available in the 1960s and 1970s).

There are many ways such hybrid mechanisms could be implemented, and my recent work on different processing layers within an integrated architecture (combining reactive, deliberative and meta-management layers) indicates some features of a hybrid system, with probabilistic associations dominating the reactive layer and structure manipulations being more important in the deliberative layer. For recent papers on this see the Cogaff papers directory http://www.cs.bham.ac.uk/research/cogaff/ and my "talks" directory: http://www.cs.bham.ac.uk/~axs/misc/talks/

More specific though less comprehensive models have been proposed by other researchers, one of the most impressive being the ACT-R system developed by John Anderson and his collaborators. See http://act.psy.cmu.edu/.

End notes

[*]Note: This is a modified version of a paper with the same title presented to the AISB Summer conference, in July 1974, at The University of Sussex.

Most of the content was inspired by my interactions with Benjamin Sloman while he was learning to think about numbers, aged about 5 years, during a year (1972-3) when I was visiting the University of Edinburgh, aged about 36. I was learning to think about information structures, programs and architectures while he was learning to think about numbers (and many other things.)

We both learnt an enormous amount that year. Trying to understand his development, and ways in which it could be influenced (programmed?) helped to convince me that AI was at least *beginning* to produce theories of the right general sort, though still lacking in detail and comprehensiveness. (Ben was born in November 22nd 1967 and died 2nd February 2002)

---

(1) For instance, it might be the case that because of the differences between human nipples and the

rubber variety, infants who are breast-fed develop a better grasp of some basic principles of physics and mechanics at a very early age because they have to surmount more obstacles to get their milk! This could also develop perseverance, independence, the ability to cope with frustration, etc. Such differences may, however, easily be counteracted by other factors in the environment.

---

(2) See Pylyshyn 1978, Sloman 1978.

---

(3) **[[Note added in 2001, about 25 years after the above was written:**

> Alas it seems that too much of education regarding use of computers in schools is now just another case of teaching students to be passive users of complex systems others have produced, like teaching them to drive cars. The opportunity to use computers in the way that meccano sets can be used for educational purposes, has largely been ignored.
>
> If it had not been ignored, children would be learning how to design, implement, test, debug, analyse, describe, explain and criticise increasingly complex systems. Their minds are not being trained to deal with all the complex systems they will encounter in real life, including social systems, political systems, economic systems, computing systems, and human minds, including their own. If it had not been ignored, people starting training to be psychologists would have had experience of building and testing systems that manipulate and use information structures far more complex than the one depicted in Figure 6.
>
> I often use the phrase "mouse potato" by analogy with "couch potato" used to describe passive watchers of television programmes. The generation of couch potatoes is educating a generation of mouse potatoes.
>]]

Book contents page
Next: Chapter 9

Last updated: 29 Jan 2007

CHAPTER 9
PERCEPTION AS A COMPUTATIONAL PROCESS

9.1. Introduction

In this chapter I wish to elaborate on a theme which Immanuel Kant found obvious: there is no perception without prior knowledge and abilities.

In the opening paragraphs of the Introduction to *Critique of Pure Reason* he made claims about perception and empirical knowledge which are very close to assumptions currently made by people working on artificial intelligence projects concerned with vision and language understanding. He suggested that all our empirical knowledge is made up of both 'what we receive through impressions' and of what 'our own faculty of knowledge supplies from itself. That is, perception is not a passive process of receiving information through the senses, but an active process of analysis and interpretation, in which 'schemata' control what happens. In particular, the understanding has to 'work up the raw material' by *comparing* representations, *combining* and *separating* them. He also points out that we may not be in a position to distinguish what we have added to the raw material, 'until with long practice of attention we have become skilled in separating it'. These ideas have recently been re-invented and elaborated by some psychologists (for example, Bartlett).

One way of trying to become skilled in separating the raw material from what we have added is to attempt to design a machine which can see. In so doing we learn that a great deal of prior knowledge has to be programmed into the machine before it can see even such simple things as squares, triangles, or blocks on a table. In particular, as Kant foresaw, such a machine has to use its knowledge in comparing its sense-data, combining them into larger wholes, separating them, describing them, and interpreting them as representing some other reality. (This seems to contradict some of the claims made by Ryle about perception, in his 1949, e.g. p. 229, last paragraph.)

[[Note added August 2002:
A slide presentation on requirements for some sort of "innate" conceptual information in intelligent systems can be found here
http://www.cs.bham.ac.uk/~axs/misc/talks/#talk14
*Getting meaning off the ground: symbol grounding vs symbol attachment.***]]**

[[Note added Jan 2007
During 2005-6 while working on the CoSy robotic project I became increasingly aware that the ideas presented here and in several later papers were too much concerned with perception of multi-layered structures, ignoring perception of processes, especially concurrent perception of processes at different levels of abstract. This topic was discussed in this presentation

A (Possibly) New Theory of Vision.

]]

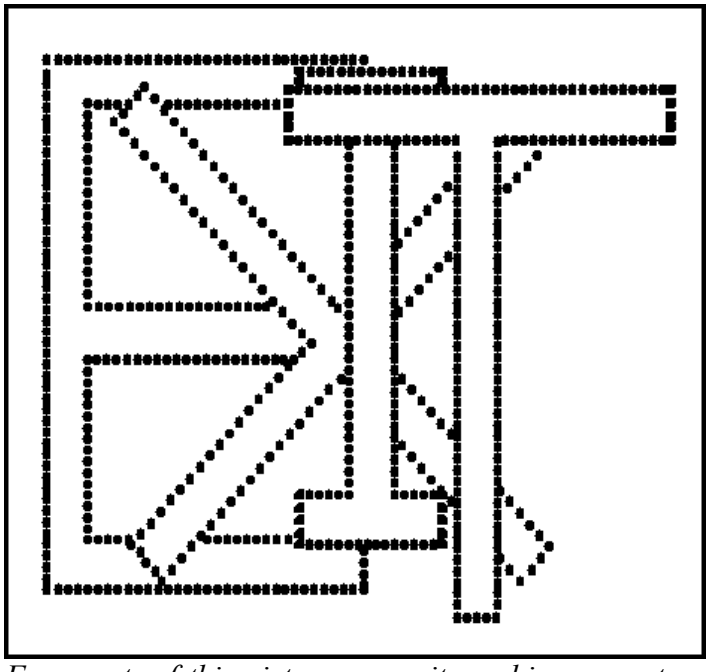
9.2. Some computational problems of perception

People have very complex perceptual abilities, some of them shared with many animals. Especially

difficult to explain is the ability to perceive form and meaning in a complex and messy collection of ambiguous and noisy data. For instance, when looking at a tree we can make out twigs, leaves, flowers, a bird's nest, spiders' webs and a squirrel. Similarly, we can (sometimes) understand what is said to us in conversations at noisy parties, we can read irregular handwriting, we can see familiar objects and faces depicted in cartoons and 'modern' paintings, and we can recognise a musical theme in many different arrangements and variations.

Seeing the significance in a collection of experimental results, grasping a character in a play or novel, and diagnosing an illness on the basis of a lot of ill-defined symptoms, all require this ability to make a 'Gestalt' emerge from a mass of information. A much simpler example is our ability to see something familiar in a picture like Figure 1. How does a 'Gestalt', a familiar word, emerge from all those dots?

Close analysis shows that this kind of ability is required even for ordinary visual perception and speech understanding, where we are totally unaware that we are interpreting untidy and ambiguous sense-data. In order to appreciate these unconscious achievements, try listening to very short extracts from tapes of human speech (about the length of a single word), or looking at manuscripts, landscapes, street scenes and domestic objects through a long narrow tube. Try looking at portions of Figure 1 through a hole about 3 mm in diameter in a sheet of paper laid on the figure and moved about. This helps to reveal how ambiguous and unclear the details are, even when you think they are clear and unambiguous. Boundaries are fuzzy, features indistinct, possible interpretations of parts of our sense-data indeterminate.



Fragments of this picture are quite ambiguous, yet somehow they help to disambiguate one another, so that most people see a pile of letters forming a familiar word. Often the word is seen before all the letters are recognized, especially if noise is introduced making recognition of the letters harder (e.g. if some dots are removed and spurious dots added). Without knowledge of letters we would have no strong reason to group some of the fragments, e.g. the top of the "I" and the rest of the "I".

Figure 1

Perceived fragments require a context for their interpretation. The trouble is that the context usually

consists of other equally ambiguous, incomplete, or possibly even spurious fragments.

Sometimes our expectations provide an additional context, but this is not essential, since we can perceive and interpret totally unexpected things, like a picture seen on turning a page in a newspaper, or a sentence overheard on a bus.

9.3. The importance of prior knowledge in perception

What we can easily perceive and understand depends on what we know. One who does not know English well will not be able to hear the English sentences uttered at a noisy party, or to read my handwriting! Only someone who knows a great deal about Chemistry will see the significance in a collection of data from chemical experiments. Only someone with a lot of knowledge about lines, flat sheets, letters and words will quickly see 'EXIT' in Figure 1.

Perception uses knowledge and expertise in different ways, clearly brought out by work on computer programs which interpret pictures. One of the most striking features of all this work is that it shows that very complex computational processes are required for what appeared previously to be very simple perceptual abilities, like seeing a block, or even seeing a straight line as a line. These processes make use of many different sorts of background knowledge, for instance in the following conscious and unconscious achievements:

a. Discerning features in the sensory array (for instance discerning points of high contrast in the visual field),

b. Deciding which features to group into significant larger units (e.g. which dots to group into line segments in Figure 1),

c. Deciding which features to ignore because they are a result of noise or coincidences, or irrelevant to the present task,

d. Deciding to separate contiguous fragments which do not really belong together (e.g. adjacent dots which are parts of the boundaries of different letters),

e. Making inferences which go beyond what is immediately given (e.g. inferring that the edge of one bar continues behind another bar, in Figure 1),

f. Interpreting what is given, as a representation of something quite different (e.g. interpreting a flat image as representing a scene in which objects are at different depths: Figure 1 is a very simple example),

g. Noticing and using inconsistencies in an interpretation so as to re-direct attention or re-interpret what is given.

h. Recognising cues which suggest that a particular mode of analysis is appropriate, or which suggest that a particular type of structure is present in the image or the scene depicted e.g. detecting the *(style* of a picture this can enable an intelligent system to avoid a lot of wasteful searching for analyses and interpretations.

So, perceiving structure or meaning may include using knowledge to reject what is irrelevant (like background noise, or coincidental juxtapositions) and to construct or hallucinate what is not there at all. It is an active constructive process which uses knowledge of the 'grammar' of sensory data, for instance knowledge of the possible structures of retinal images, knowledge about the kinds of things depicted or represented by such data, and knowledge about the processes by which objects generate sense-data. Kant's 'schemata' must incorporate all this.

We need not be aware that we possess or use such knowledge. As Kant noticed, it may be an 'art

concealed in the depths of the human soul' (p. 183, in Kemp Smith's translation), much of it "compiled" into procedures and mechanisms appropriate to images formed by the kind of world we live in. But at present there are no better explanations of the possibility of perception than explanations in terms of intelligent processes using a vast store of prior information, much of which is "compiled" (by evolution or by individual learning) into procedures and mechanisms appropriate to images formed by the kind of world we live in.

For instance, theories according to which some perception is supposed to be 'direct', not involving any prior knowledge, nor the use of concepts, seem to be quite vacuous. A theory which claims that perceptual achievements are not decomposable into sub-processes cannot be used as a basis for designing a working mind which can perceive any of the things we perceive. It lacks explanatory power, because it lacks generative power. If the processes cannot be decomposed, then there is no way of generating the huge variety of human perceptual abilities from a relatively economical subsystem. By contrast, computational theories postulating the use of prior knowledge of structures and procedures can begin to explain some of the fine structure (see chapters 2 and 3) of perceptual processes, for example, the perception of this as belonging to that, this as going behind that, this as similar to that, this as representing that, and so on.

Quibbles about whether the ordinary word 'knowledge' is appropriate for talking about the mechanisms and the stored facts and procedures used in perception seem to be merely unproductive distractions. Even if the ordinary usage of the word 'knowledge' does not cover such inaccessible information, extending the usage would be justified by the important insights gained thereby. Alternatively, instead of talking about 'knowledge' we can talk about 'information' and say that even the simplest forms of perception not only provide new information, in doing so they make use of various kinds of prior information.

In a more complete discussion it would be necessary to follow Kant and try to distinguish the role of knowledge gained from previous perceptual experiences and the role of knowledge and abilities which are required for any sort of perceptual experience to get started. The latter cannot be empirical in the same sense, though it may be the result of millions of years of evolutionary "learning from experience".

Since our exploration of perceptual systems is still in a very primitive state, it is probably still too early to make any such distinctions with confidence. It would also be useful to distinguish general knowledge about a class of theoretically possible objects, situations, processes, etc., from specific knowledge about commonly occurring subsets. As remarked in chapter 2, we can distinguish knowledge about the *form* of the world from knowledge about its *contents*. Not all geometrically possible shapes are to be found amongst animals, for example. A bat may in some sense be half way between a mouse and a bird: but not all of the intervening space is filled with actually existing sorts of animals. If the known sorts of objects cluster into relatively discrete classes, then this means that knowledge of these classes can be used to short-circuit some of the more general processes of analysis and interpretation which would be possible. In information-theoretic terms this amounts to an increase of redundancy -- and a reduction of information -- in sensory data. This is like saying that if you know a lot of relatively commonly occurring words and phrases, then you may be able to use this knowledge to cut down the search for ways of interpreting everything you hear in terms of the most general grammatical and semantic rules. (Compare Becker on the 'phrasal lexicon'.) This is one of several ways in which the environment can be cognitively 'friendly' or 'unfriendly'. We evolved to cope with a relatively cognitively friendly environment.

In connection with pictures like Figure 1, this means that if you know about particular letter-shaped configurations of bars, then this knowledge may make it possible to find an interpretation of such a picture in terms of bars more rapidly than if only general bar-configuration knowledge were deployed.

For instance, if you are dealing with our capital letters, then finding a vertical bar with a horizontal one growing to the left from its middle, is a very good reason for jumping to the conclusion that it is part of an 'H', which means that you can expect another vertical bar at the left end of the horizontal.

Thus a rational creature, concerned with maximising efficiency of perceptual processing, might find it useful to store a very large number of really quite redundant concepts, corresponding to commonly occurring substructures (phrases) which are useful discriminators and predictors.

The question of how different sorts of knowledge can most fruitfully interact is a focus of much current research in artificial intelligence. The strategies which work in a 'cognitively friendly world' where species of things cluster are highly fallible if unusual situations occur. Nevertheless the fallible, efficient procedures may be the most rational ones to adopt in a world where things change rapidly, and your enemies may not give you time to search for a *conclusive* demonstration that it is time to turn around and run. Thus much of the traditional philosophical discussion of rationality, in terms of what can be proved beyond doubt, is largely irrelevant to real life and the design of intelligent machines. But new problems of rationality emerge in their place, such as problems about trading space against time, efficiency against flexibility or generality, and so on. From the design standpoint, rationality is largely a matter of choosing among trade-offs in conditions of uncertainty, not a matter of getting things 'right', or selecting the 'best'. (For more on trade-offs see the chapters on representations, and on numbers: Chapter 7 and Chapter 8)).

9.4. Interpretations

Knowledge is used both in analysing structures of images and in interpreting those structures as depicting something else. There may be several different layers of interpretation. For example in Figure 1, dot configurations represent configurations of lines. These in turn represent configurations of bars. These represent strokes composing letters. And sequences of letters can represent words (see fig. 6). Within each layer there may be alternative structures discernible, for instance alternative ways of grouping things, alternative relations which may be noticed. These will affect the alternative interpretations of that layer. By examining examples in some detail and trying to design mechanisms making the different experiences possible we can gain new insights into the complex and puzzling concept of 'seeing as', discussed at length in part II of Wittgenstein's *Philosophical Investigations*.

Contrary to what many people (including some philosophers) have assumed, there need not be any similarity between what represents and what it represents. Instead, the process of interpretation may use a variety of interpretation rules, of which the most obvious would be rules based on information about a process of projection which generates, say, a two-dimensional image from a three-dimensional scene. (For more on this see the chapter on analogical representations.)

The projection of a three dimensional scene onto a two dimensional image is just a special case of a more general notion of *evidence* which is generated in a systematic way by that which explains it. A two-dimensional projection of a three-dimensional object bears very little similarity to the object. (Cf. Goodman, *Languages of Art.)* The interpretation procedure may allow for the effects of the transformations and distortions in the projection (as a scientist measuring the temperature of some liquid may allow for the fact that the thermometer cools the liquid).

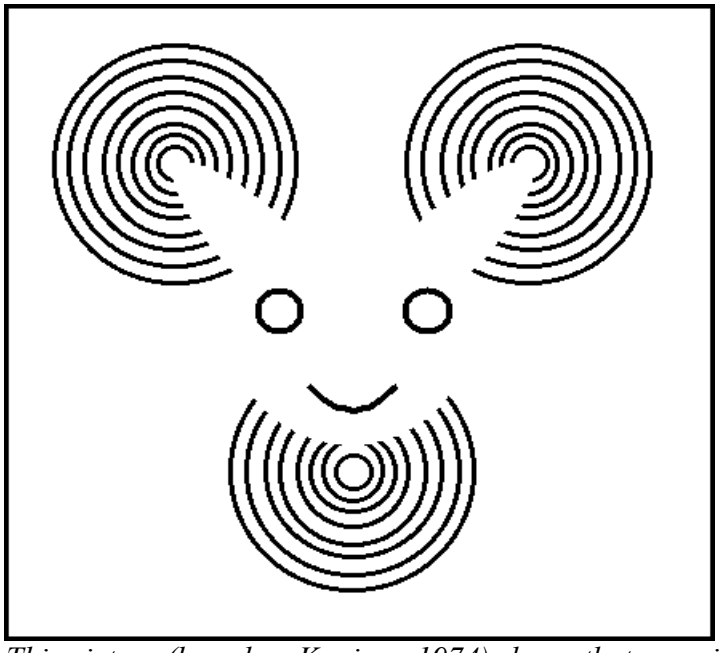
This is an old idea: what is new in the work of A.I. is the detailed analysis of such transformations and interpretation procedures, and the adoption of new standards for the acceptability of an explanation: namely it must suffice to generate a working system, that is, a program which can use knowledge of the transformations to interpret pictures or the images produced by a television camera connected to the computer.

What we are conscious of seeing is the result of many layers of such interpretation, mostly unconscious, yet many of them are essentially similar in character to intellectual processes of which we are sometimes conscious. All this will be obvious to anyone familiar with recent work in theoretical linguistics.

So the familiar philosophical argument that we do not see things as they are, because our sense-organs may affect the information we receive, is invalid. For however much our sense-organs affect incoming data, we may still be able to interpret the data in terms of how things really are. But this requires the use of knowledge and inference procedures, as people trying to make computers see have discovered. Where does the background knowledge come from? Presumably a basis is provided genetically by what our species has learnt from millions of years of evolution. The rest has to be constructed, from infancy onwards, by individuals, with and without help, and mostly unconsciously.

9.5. Can physiology explain perception?

To say that such processes are unconscious does not imply that they are physiological as people sometimes assume in discussions of such topics. Physical and physiological theories about processes in the brain cannot account for these perceptual and interpretative abilities, except possibly at the very lowest levels, like the ability to detect local colour contrasts in the visual field. Such tasks can be delegated to physical mechanisms because they are relatively determinate and context-independent, that is algorithmic (e.g. see Marr, 1976). In particular, such peripheral processes need not involve the construction and testing of rival hypotheses about how to group fragments of information and how to interpret features of an image. But, physical and physiological mechanisms cannot cope with the more elaborate context-dependent problem-solving processes required for perception. The concept of using stored knowledge to interpret information has no place in physics or physiology, even though a physical system may serve as the computer in which information is stored and perceptual programs executed.



This picture (based on Kanizsa, 1974) shows that perceived colour at a location depends not only on the corresponding physical stimulus, hut also on the context. Most people see the central region as whiter than the rest, even though there is no physical difference.

Figure 2

Moreover, even colour contrasts can sometimes be hallucinated on the basis of context, as so-called 'illusory-contrasts' show. For an example see Figure 2.

Instead of physiological theories, we need 'computational theories, that is, theories about processes in which symbolic representations of data are constructed and manipulated. In such processes, facts about part of an image are interpreted by making inferences using context and background knowledge. We must not be blinded by philosophical or terminological prejudices which will not allow us to describe unconscious processes as inferences,or, more generally, as 'mental processes'.

How is it done? In particular, what exactly is the knowledge required for various kinds of perception, and how do we mobilise it as needed? We cannot yet claim to have complete or even nearly complete explanations. But A.I. work on vision has made some significant progress, both in showing up the inadequacies of bad theories and sketching possible fragments of correct explanations.

> *Our present ignorance is not a matter of our not knowing which theory is correct, but of our not even knowing how to formulate theories sufficiently rich in explanatory power to be worth testing experimentally.*

Attempting to program computers to simulate human achievements provides a powerful technique for finding inadequacies in our theories thereby stimulating the rapid development of new theory-building tools. In the process we are forced to re-examine some old philosophical and psychological problems. For a survey of some of this work, see the chapters on computer vision in Boden (1977). Winston (1975) also includes useful material, especially the sections by Winston, Waltz, and Minsky. The rest of this chapter illustrates some of the problems with reference to an ongoing computer project at Sussex University, which may be taken as representative.

9.6. Can a computer do what we do?

We are exploring some of the problems of visual perception by attempting to give a computer the ability to perceive a configuration of known shapes in a scene depicted by a 'spotty' picture like Figure 1. The pictures are presented to the program in the form of a 2-dimensional binary (i.e. black and white) array. The array is generated by programs in the computer either on the basis of instructions, or with the aid of a graphical input terminal. Additional spurious dots ('positive noise') can be added to make the pictures more confusing. Similarly, spurious gaps ('negative noise') can be added.

People can cope quite well with these pictures even when there is a lot of positive and negative noise, and where further confusion is generated by overlaps between letters, and confusing juxtapositions. Some people have trouble at first, but after seeing one or two such pictures, they interpret new ones much more rapidly. The task of the program is to find familiar letters without wasting a lot of time investigating spurious interpretations of ambiguous fragments. It should 'home in on' the most plausible global interpretation fairly rapidly, just as people can.

Out of context, picture details are suggestive but highly ambiguous, as can be seen by looking at various parts of the picture through a small hole in a sheet of paper. Yet when we see them in context we apparently do not waste time exploring all the alternative interpretations. It is as if different ambiguous fragments somehow all 'communicate' with one another in parallel, to disambiguate one another.

Waltz (1975) showed how this sort of mutual disambiguation could be achieved by a program for interpreting line drawings representing a scene made up of blocks on a table, illuminated by a shadow-casting light. He gave his program prior knowledge of the possible interpretations of various

sorts of picture junctions, all of which were ambiguous out of context. So the problem was to find a globally consistent interpretation of the whole picture. The program did surprisingly well on quite complex pictures. His method involved letting possible interpretations for a fragment be 'filtered out' when not consistent with any possible interpretations for neighbouring fragments.

> **[[Note added 2001:**
> since 1975 there have been huge developments in techniques for 'constraint propagation', including both hard and soft constraints. **]]**

But the input to Waltz' program was a representation of a perfectly connected and noise-free line drawing. Coping with disconnected images which have more defects, requires more prior knowledge about the structure of images and scenes depicted, and more sophisticated computational mechanisms.

Which dots in Figure 1 should be grouped into collinear fragments? By looking closely at the picture, you should be able to discern many more collinear groups than you previously noticed. That is, there are some lines which 'stand out' and are used in building an interpretation of the picture, whereas others for which the picture contains evidence are not normally noticed. Once you have noticed that a certain line 'stands out', it is easy to look along it picking out all the dots which belong to it, even though some of them may be 'attracted' by other lines too.

But how do you decide which lines stand out without first noticing all the collinear groups of dots? Are all the collinear dot-strips noticed unconsciously? What does that mean? Is this any different from unconsciously noticing grammatical relationships which make a sentence intelligible?

When pictures are made up of large numbers of disconnected and untidy fragments, then the interpretation problem is compounded by the problem of deciding which fragments to link together to form larger significant wholes. This is the 'segmentation' or 'agglomeration' problem. As so often happens in the study of mental processes, we find a circularity: once a fragment has been interpreted this helps to determine the others with which it should be linked, and once appropriate links have been set up the larger fragment so formed becomes less ambiguous and easier to interpret. It can then function as a recognisable cue. (The same circularity is relevant to understanding speech.)

9.7. The POPEYE program [1]

Our computer program breaks out of this circularity by sampling parts of the image until it detects a number of unambiguous fragments suggesting the presence of lines. It can then use global comparisons between different lines to see which are supported most strongly by relatively unambiguous fragments. These hypothesised bold lines then direct closer examination of their neighbourhoods to find evidence for bar-projections. Evidence which would be inconclusive out of context becomes significant in the context of a nearby bold line hypothesised as the edge of a bar an example of a 'Gestalt' directing the interpretation of details.

Thus, by using the fact that *some* fragments are fairly unambiguous, we get the process started. By using the fact that long stretches of relatively unambiguous fragments are unlikely to be spurious, the program can control further analysis and interpretations. Parallel pairs of bold lines are used as evidence for the presence of a bar. Many of the strategies used are highly fallible. They depend on assumption that the program inhabits a 'cognitively friendly' world, that is, that it will not be asked to interpret very messy, very confusing pictures. If it is, then, like people, it will become confused and start floundering.

Clusters of bar-like fragments found in this way can act as cues to generate further higher-level hypotheses, for example, letter hypotheses, which in turn control the interpretation of further ambiguous fragments. (For more details, see Sloman and Hardy 'Giving a computer gestalt

experiences' and Sloman *et al.* 1978.) In order to give a program a more complete understanding of *our* concepts, we would need to embody it in a system that was able to move about in space and manipulate physical objects, as people do. This sort of thing is being done in other artificial intelligence research centres. However, there are still many unsolved problems. It will be a long time before the perceptual and physical skills of even a very young child can be simulated.

The general method of using relatively unambiguous fragments to activate prior knowledge which then directs attention fruitfully at more ambiguous fragments, seems to be required at all levels in a visual system. It is sometimes called the 'cue-schema' method, and seems to be frequently re-invented.

However, it raises serious problems, such as: how should an intelligent mechanism decide which schemas are worth storing in the first place, and how should it, when confronted with some cue, find the *relevant* knowledge in a huge memory store? (Compare chapter 8.) A variety of sophisticated indexing strategies may be required for the latter purpose. Another important problem is how to control the invocation of schemas when the picture includes cues for many different schemas.

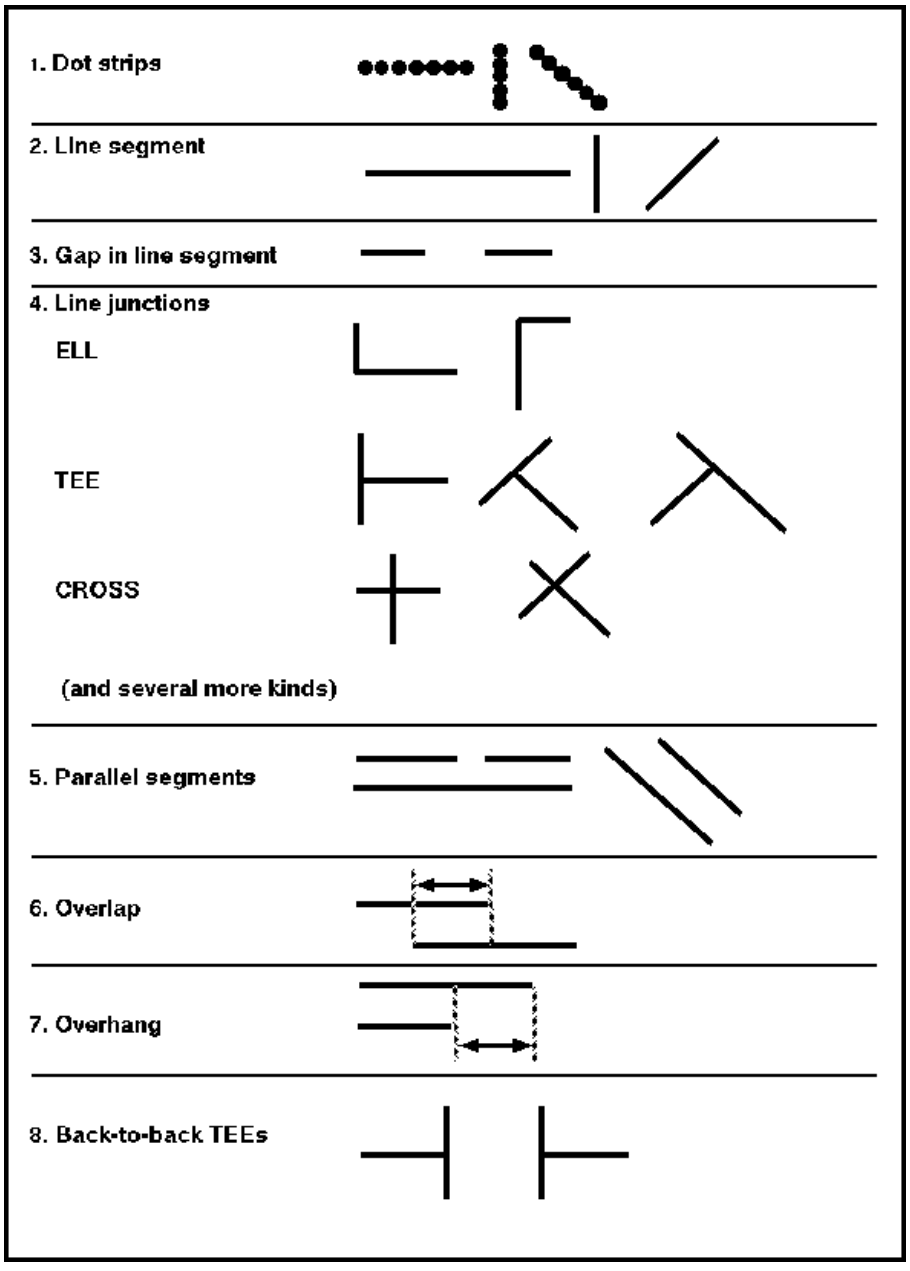
Our program uses knowledge about many different kinds of objects and relationships, and runs several different sorts of processes in parallel, so that 'high-level' processes and (relatively) low-level' processes can help one another resolve ambiguities and reduce the amount of searching for consistent interpretations. It is also possible to suspend processes which are no longer useful, for example low-level analysis processes, looking for evidence of lines, may be terminated prematurely if some higher-level process has decided that enough has been learnt about the image to generate a useful interpretation.

This corresponds to the fact that we may recognise a whole (e.g. a word) without taking in all its parts. It is rational for an intelligent agent to organise things this way in a rapidly changing world where the ability to take quick decisions may be a matter of life and death.

Like people, the program can notice words and letters emerging out of the mess in pictures like Figure 1. As Kant says, the program has to work up the raw material by comparing representations, combining them, separating them, classifying them, describing their relationships, and so on. What Kant failed to do was describe such processes in detail.

9.8. The program's knowledge

In dealing with Figure 1 the program needs to know about several different domains of possible structures, depicted in Figure 3:



Some concepts relevant to the domain of 2 dimensional configurations of line-segments required for the interpretation of Figure 1. In this 2-D domain, nothing can be invisible or partly covered, unlike the domain of overlapping rectangular laminas shown in Figure 4. The process of interpreting Figure 1 includes identifying items in the 2-D domain and mapping them to items in the 2.5D domain of laminas.

Figure 3

The domains of knowledge involved include:

a. The domain of 2-dimensional configurations of dots in a discrete rectangular array (the "dotty picture" domain).

b. The domain of 2-dimensional configurations of line-segments in a continuous plane. The configurations in the dotty picture domain *represent* configurations of such lines -- notice the differences between a collection of dots *being* a line segment, *lying on* a line segment and *representing* a line segment.
c. The (two-and-a-half-dimensional) domain of overlapping laminas composed of 'bars'. Patterns in the line-domain

d. represent configurations of bars and laminas made of rectangular bars.

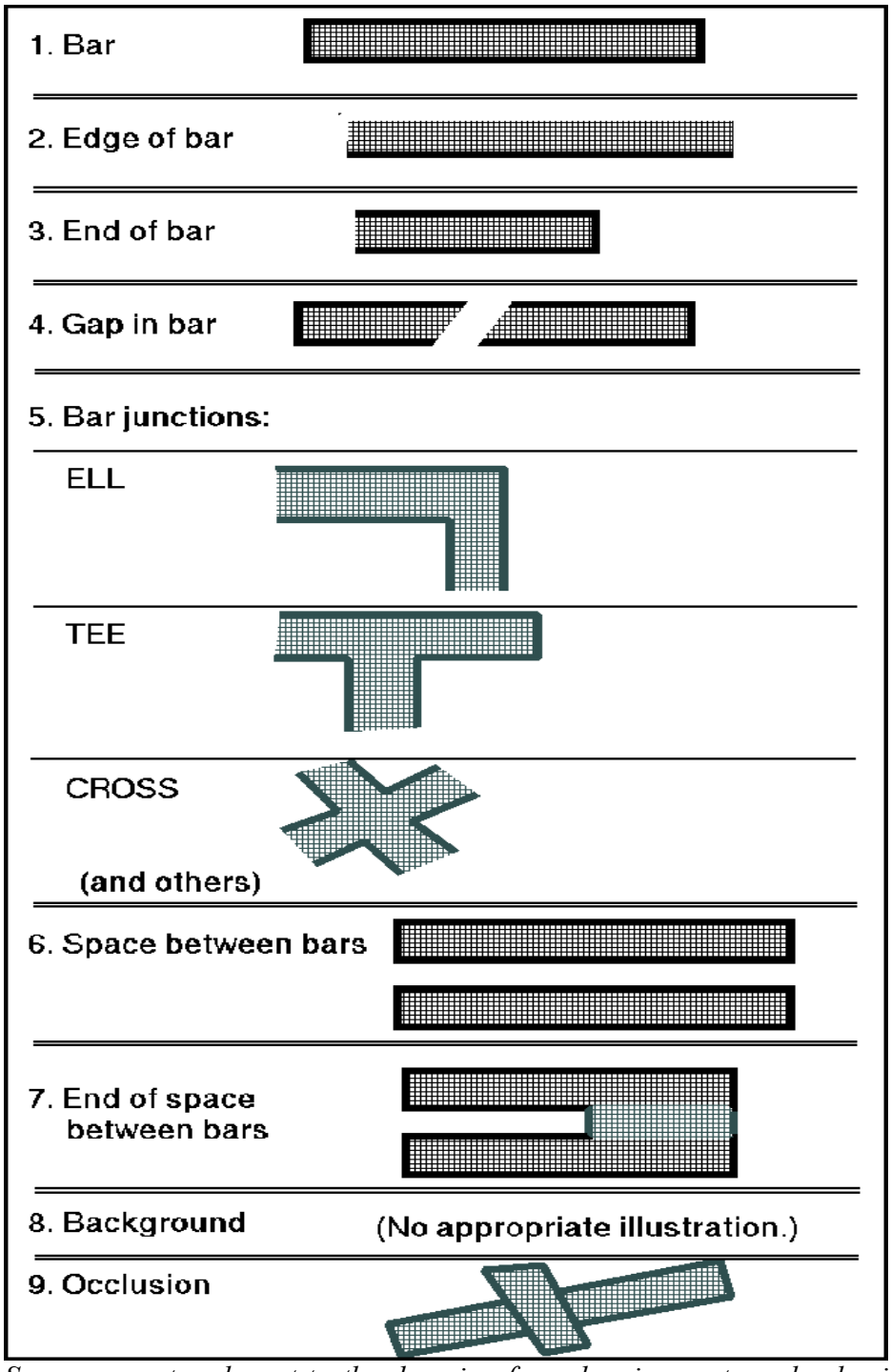
e. An abstract domain containing configurations of 'strokes' which have orientations, lengths, junctions, and so on, analogous to lines, but with looser criteria for identity. Letters form a subset of this domain. Configurations in this domain are represented by configurations of laminas. That is, a bar-shaped lamina represents a possible stroke in a letter, but strokes of letters can also be depicted by quite different patterns (as in this printed text) which is why I say their domain is 'abstract' following Clowes, 1971.

f. An abstract domain consisting of sequences of letters. Known words form a subset of this domain.

In particular the program has to know how to build and relate descriptions of structures in each of these domains, including fragments of structures. That is, the ability to solve problems about a domain requires an 'extension' of the domain to include possible fragments of well-formed objects in the domain Becker's 'phrasal lexicon' again. Our program uses many such intermediate concepts. Figures 3 and 4 list and illustrate some of the concepts relevant to the second and third domains. Figure 5 shows some of the cues that can help reduce the search for an interpretation. Figure 6 shows all the domains and some of the structural correspondences between items in those domains.

By making use of the notion of a series of domains, providing different 'layers' of interpretation, it is possible to contribute to the analysis of the concept of 'seeing as', which has puzzled some philosophers. Seeing X as Y is in general a matter of constructing a mapping between a structure in one domain and a possibly different structure in another domain. The mapping may use several intermediate layers.

[[Note added 2001:
our recent work on architectures containing a 'meta-management' layer suggests that *being aware of seeing X as Y* requires additional meta-management, i.e. self-monitoring processes, which are not essential for the basic processes of *seeing X as Y*, which could occur in simpler architectures, e.g. in animals that are not aware of their own mental processes (like most AI systems so far). **]]**



Some concepts relevant to the domain of overlapping rectangular laminas. This sort of domain is sometimes described as "two and a half dimensional" (2.5D) because one object can be nearer or further than another, and because all or part of an object can be invisible because it is hidden behind another, unlike a purely 2D domain where everything is visible. Knowledge of such 2.5D concepts can help the search for a good interpretation of pictures like Figure 1. This raises problems about how the concepts are stored and indexed, how they are accessed by cues, and how ambiguities are resolved. Some of the discussion in Chapter 6 regarding special purpose and general purpose monitors is relevant.

Figure 4

Facts about one domain may help to solve problems about any of the others. For instance, lexical knowledge may lead to a guess that if the letters 'E', 'X' and T' have been found, with an unclear letter between them, then the unclear letter is 'I'. This in turn leads to the inference that there is a lamina depicting the 'I' in the scene. From that it follows that unoccluded edges of the lamina will be represented by lines in the hypothetical picture in domain (b). The inferred locations of these lines can lead to a hypothesis about which dots in the picture should be grouped together, and may even lead to the conclusion that some dots which are not there should be there.

The program, like a person, needs to know that a horizontal line-segment in its visual image can represent (part of) the top or bottom edge of a bar, that an ELL junction between line segments can depict part of either a junction between two bars or a corner of a single bar. In the former case it may depict either a concave or a convex corner, and, as always, context will have to be used to decide which.

The program does not need to define concepts of one domain in terms of concepts from another. Rather the different domains are defined by their own primitive concepts and relations. The notion of 'being represented by' is not the same as the notion of 'being defined in terms of'. For instance, 'bar' is not defined in terms of actual and possible sense-data in the dot-picture domain, as some reductionist philosophical theories of perception would have us believe. Concepts from each domain are defined implicitly for the program in terms of structural relations and inference rules, including interpretation strategies.

So the organisation of the program is more consistent with a dualist or pluralist and wholistic metaphysics than with an idealist or phenomenalist reduction of the external world to sense-data, or any form of philosophical atomism, such as Russell and Wittgenstein once espoused.

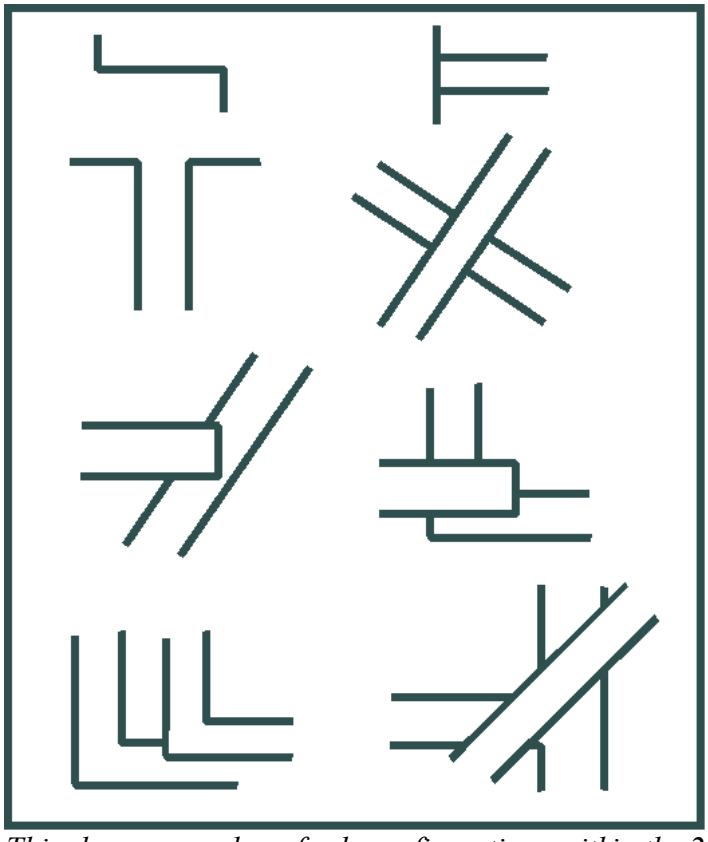
9.9. Learning

Programs, like people, can in principle work out lots of things for themselves, instead of having them all programmed explicitly. For instance Figure 5 shows typical line-picture fragments which can be generated by laminas occluding one another. A program could build up a catalogue of such things for itself for instance by examining lots of drawings. Research is in progress on the problem of designing systems which learn visual concepts, possibly with the help of a teacher who chooses examples for the system to work on. (For example, see Winston, 1975.) It is certain that there are many more ways of doing such things than we have been able to think of so far. So we are in no position to make claims about which gives the best theory of how people learn.

> **[[Note added 2001:**
> In the decades since this book was written many more learning methods have been developed for vision and other aspects of intelligence, though surprisingly few of them seem to involve the ability to learn about different classes of structures in domains linked by representation relationships. Many of them attempt to deal with fairly direct mappings between configurations detectable in image sequences and abstract concepts like "person walking". For examples see journals and conference proceedings on machine vision, pattern recognition, and machine learning.**]]**

Currently our program starts with knowledge which has been given it by people (just as people have to start with knowledge acquired through a lengthy process of biological evolution). Perhaps, one day, some of the knowledge will be acquired by a machine itself, interacting with the world, if a television camera and mechanical arm are connected to the computer, as is already done in some A.I. research laboratories. However, real learning requires much more sophisticated programs than programs which

have a fixed collection of built-in abilities. (Some of the problems of learning were discussed previously in Chapter 6 and Chapter 8.)



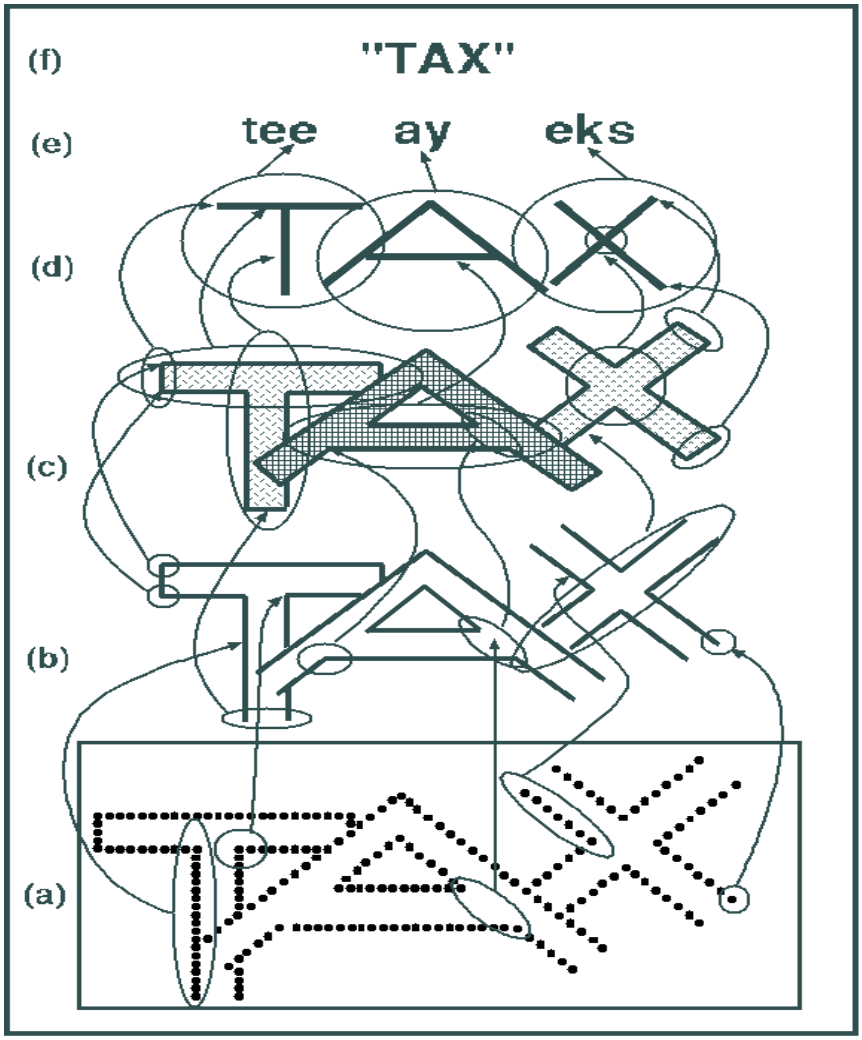
This shows a number of sub-configurations within the 2-D line-segment domain of Figure 3 which are likely to occur in images depicting overlapping laminas from the domain of Figure 4. A set of 2-D line images depicting a different class of laminas, or depicting objects in a different domain, e.g. 3-D forest scenes, would be likely to include a different class of sub-configurations made of lines.
Likewise in depictions of forest scenes, commonly occurring configurations in the dotty picture domain would be different from those found in Figure 1.
Knowledge of commonly occurring sub-structures in images, corresponding to particular domains represented, like knowledge about the objects represented, can help the interpretation process.
This is analogous to processes in language-understanding in which knowledge of familiar phrases is combined with knowledge of a general grammar which subsumes those phrases. (Becker 1975)
[[This caption was substantially extended in 2001]]

Figure 5

Given structural definitions of letters, and knowledge of the relations between the different domains illustrated in Figure 6, a program might be able to work out or learn from experience that certain kinds of bar junctions (Figure 4), or the corresponding 2-D line configurations (Figures 3 and 5), occur only in a few of them, and thus are useful disambiguating cues. This will not be true of all the fragments visible in Figure 1. Thus many fragments will not be recognised as familiar, and spurious linkages and hypotheses will therefore not be generated. If the program were familiar with a different

world, in which other fragments were significant, then it might he more easily be confused by Figure 1. So additional knowledge is not always helpful. (Many works of art seem to require such interactions between differentdomains of knowledge.)

A program should also be able to 'learn' that certain kinds of fragments do not occur in any known letter, so that if they seem to emerge at any stage this will indicate that picture fragments have been wrongly linked together. This helps to eliminate fruitless searches for possible interpretations. So the discovery of anomalies and impossibilities may play an important role in the development of rational behaviour. A still more elaborate kind of learning would involve discovering that whether a fragment is illegitimate depends on the context. Fragments which are permissible within one alphabet may not be permissible in another. Thus the process of recognising letters is facilitated by knowledge of the alphabet involved, yet some letter recognition may be required for the type of alphabet to be inferred: another example of the kind of circularity, or mutual dependence, of sub-abilities in an intelligent system.



This shows how several layers of interpretation may be involved in seeing letters in a dot-picture. Each layer is a domain of possible configurations in which substructures may represent or be represented by features or substructures in other layers. The following domains are illustrated: (a) configurations of dots, spaces, dotstrips, etc., (b) configurations of 2-D line-segments, gaps, junctions, etc., (c) configurations of possibly overlapping laminas (plates) in a 2.5D domain containing bars, bar-junctions, overlaps, edges of bars, ends of bars, etc., (d) a domain of stroke

configurations where substructures can represent letters in a particular type of font, (e) a domain of letter sequences, (f) a domain of words composed of letter sequences.

Figure 6

NOTE [13 Jan 2007]:
The diagram in Figure 6 suggests that all information flows upwards. That is not how the program worked: there was a mixture of bottom-up, top-down and middle-out processing.

9.10. Style and other global features

Knowledge of 'picture styles' can also play an important role in the process of perception and interpretation. Variations in style include such things as whether the letters are all of the same height and orientation, whether the bars are all of the same width, whether the letters in words tend to be jumbled, or overlapping, or at stepped heights, and so on. Notice that some of these stylistic concepts depend on quite complex geometrical relationships (for instance, what does 'jumbled' mean?). If the program can take note of clues to the style of a particular picture, during its analysis, this can help with subsequent decisions about linking or interpreting fragments. If you know the sizes of letters, for instance, then you can more easily decide whether a line segment has a bit missing.

Hypotheses about style must, of course, be used with caution, since individual parts of a picture need not conform to the overall style. Local picture evidence can over-ride global strategies based on the inferred style provided that the program can operate in a mode in which it watches out for evidence conflicting with some of its general current assumptions, using monitors of the sorts described in Chapter 6.

9.11. Perception involves multiple co-operating processes

Our program includes mechanisms which make it possible to set a number of different processes going in parallel, for example, some collecting global statistics about the current picture, some sampling the picture for dot-configurations which might represent fragments of lines, others keeping track of junctions between lines, or attempting to interpret parallel segments as bars, some trying to interpret bars as strokes of letters, and so on.

This parallelism is required partly because, with a large amount of information available for analysis and interpretation, it may not be easy to decide what to do next, for example, which configurations to look for in the picture, and where to look for them. Deciding between such alternatives itself requires analysis and interpretation of evidence and at first it will not be obvious where the important clues are, nor what they are. So initially many on-going processes are allowed to coexist, until items both unambiguous and relatively important emerge, such as a long line, an unambiguous clue to the location of a bar, some aspect of the style, or a set of linked bar fragments which uniquely identify a letter.

When fragments forming clear-cut cues emerge, they can invoke a 'higher-level' schema which takes control of processing for a while, interrupting the 'blind' searching for evidence, by directing attention to suitable parts of the picture and relevant questions.

If higher level processes form a plausible hypothesis, this may suppress further analysis of details by lower level processes. For instance, recognition of fragments of 'E', or 'X', and of "I", where there appear to be only about four letters, might cause a program (or person) to jump to the conclusion that the word is 'EXIT', and if this fits into the context, further examination of lines to check out on

remaining strokes of letters, and the missing 1', might then be abandoned. This ability to jump to conclusions on the basis of partial analysis may be essential to coping with a rapidly changing world. However it depends on the existence of a fair amount of redundancy in the sensory data: that is, it assumes a relatively 'friendly' (in the sense defined previously) world. It also requires an architecture able to support multiple concurrent processes and the ability for some of them to be aborted by others when their activities are no longer needed.

This type of programming involves viewing perception as the outcome of very large numbers of interacting processes of analysis, comparison, synthesis, interpretation, and hypothesis-testing, most, if not all, unconscious. On this view the introspective certainty that perception and recognition are 'direct', 'unmediated' and involve no analysis, is merely a delusion. (This point is elaborated in the papers by Weir -- see Bibliography.)

This schizophrenic view of the human mind raises in a new context the old problem: what do we mean by saying that consciousness is 'unitary' or that a person has one mind? The computational approach to this problem is to ask: how can processes be so related that all the myriad sub-tasks may be sensibly co-ordinated under the control of a single goal, for instance the goal of finding the word in a spotty picture, or a robot's goal of using sensory information from a camera to guide it as it walks across a room to pick up a spanner? See also chapter 6 and chapter 10.

[[Note added 2001:
At the time the program was being developed, we had some difficulty communicating our ideas about the importance of parallel processing concerned with different domains because AI researchers tended to assume we were merely repeating the well-known points made in the early 1970s by Winograd, Guzman and others in the MIT AI Lab, about "heterarchic" as opposed to "hierarchic" processing.
Heterarchic systems, dealt, as ours did, with different domains of structures and relations between them (e.g. Winograd's PhD thesis dealt with morphology, syntax, semantics and a domain of three dimensional objects on a table).
Both models involve mixtures of data-driven (bottom-up) and hypothesis-driven (top-down) processes.
Both allow interleaving of processes dealing with the different domains -- unlike *hierarchic* or *pass-oriented* mechanisms which first attempt to complete processing in one domain then pass the results to mechanisms dealing with another domain, as in a processing pipeline.

The main differences between heterarchy and our model were as follows:

a. In an implementation of "heterarchic" processing there is typically only *one* locus of control at any time. Thus processing might be going on in a low level sub-system or in a high level sub-system, but not both in parallel with information flowing between them.

b. In those systems decisions to transfer control between sub-systems were all taken explicitly by processes that decided they needed information from another system: e.g. a syntactic analyser could decide to invoke a semantic procedure to help with syntactic disambiguation, and a semantic procedure could invoke a syntactic analyser to suggest alternative parses.

c. In that sort of heterarchic system it is not possible for a process working in D1 to be *interrupted* by the arrival of new information relevant to the current sub-task,

derived from processing in D2.

d. Consequently, if a process in that sort heterarchic system gets stuck in a blind-alley and does not notice this fact it may remain stuck forever.

The POPEYE architecture was designed to overcome these restrictions by allowing processing to occur concurrently in different domains with priority mechanisms in different domains determining which sub-processes could dominate scarce resources. Priorities could change, and attention within a domain could therefore be switched, as a result of arrival of new information that was not explicitly asked for. In this respect the POPEYE architecture had something in common with neural networks in which information flows between concurrently processing sub-systems (usually with simulated concurrency). Indeed, a neural net with suitable symbol-manipulating sub-systems could be used to implement something like the POPEYE architecture, though we never attempted to do this for the whole system. After this chapter was written, work was done on implementing the top level word-recognizer in POPEYE as a neural net to which the partial results from lower level systems could be fed as they became available.]]

9.12. The relevance to human perception

The world of our program is *very* simple. There are no curves, no real depth, no movement, no forces. The program cannot act in this world, nor does it perceive other agents. Yet even for very simple worlds, a computer vision program requires a large and complex collection of knowledge and abilities. From such attempts to give computers even fragmentary human abilities we can begin to grasp the enormity of the task of describing and explaining the processes involved in *real* human perception. Galileo's relationship to the physics of the 1970s may be an appropriate and humbling comparison.

In the light of this new appreciation of the extent of our ignorance about perceptual processes, we can see that much philosophical discussion hitherto, in epistemology, philosophy of mind, and aesthetics, has been based on enormous over-simplifications. With hindsight much of what philosophers have written about perception seems shallow and lacking in explanatory power. But perhaps it was a necessary part of the process of cultural evolution which led us to our present standpoint.

Another consequence of delving into attempts to give computers even very simple abilities is that one acquires enormous respect for the achievements of very young children, many other animals, and even insects. How does a bee manage to . land on a flower without crashing into it?

Many different aspects of perception are being investigated in artificial intelligence laboratories. Programs are being written or have been written which analyse and interpret the following sorts of pictures or images, which people cope with easily.

a. Cartoon drawings.

b. Line drawings of three dimensional scenes containing objects with straight edges, like blocks and pyramids.

c. Photographs or television input from three-dimensional scenes, including pictures of curved objects.

d. Stereo pairs from which fairly accurate depth information can be obtained.

e. Sequences of pictures representing moving objects, or even television input showing moving

objects.

f. Satellite photographs, which give geological, meteorological, or military information. (Unfortunately, some people are unable to procure research funds unless they pretend that their work is useful for military purposes and, even more unfortunately, it sometimes is.)

g. Pictures which represent 'impossible objects', like Escher's drawings. Like people, a program may be able to detect the impossibility (see Clowes, 1971, Huffman, 1971, and Draper (to appear)).

Some of the programs are in systems which control the actions of artificial arms, or the movements of vehicles. The best way to keep up with this work is to read journal articles, conference reports, and privately circulated departmental reports. Text-books rapidly grow out of date. (This would not be so much of a problem if we all communicated via a network of computers and dispensed with books! But that will not come for some time.)

Each of the programs tackles only a tiny fragment of what people and animals can do. For example, the more complex the world the program deals with the less of its visible structure is perceived and used by the program. The POPEYE program deals with a very simple world because we wanted it to have a fairly full grasp of its structure (though even that is proving harder than we anticipated). One of the major obstacles to progress at present is the small number of memory locations existing computers contain, compared with the human brain. But a more important obstacle is the difficulty of articulating and codifying all the different kinds of structural and procedural knowledge required for effective visual perception. There is no reason to assume that these obstacles are insuperable in principle, though it is important not to make extravagant claims about work done so far. For example, I do not believe that the progress of computer vision work by the end of this century will be adequate for the design of domestic robots, able to do household chores like washing dishes, changing nappies on babies, mopping up spilt milk, etc. So, for some time to come we shall be dependent on simpler, much more specialised machines.

9.13. Limitations of such models

It would be very rash to claim that POPEYE, or any other existing artificial intelligence program, should be taken seriously as a theory explaining human abilities. The reasons for saying that existing computer models cannot be accepted as explaining how people do things include:

a. People perform the tasks in a manner which is far more sensitive to context, including ulterior motives, emotional states, degree of interest, physical exhaustion, and social interactions. Context may affect detailed strategies employed, number of errors made, kinds of errors made, speed of performance, etc.

b. People are much more flexible and imaginative in coping with difficulties produced by novel combinations, noise, distortions, missing fragments, etc. and at noticing short cuts and unexpected solutions to sub-problems.

c. People learn much more from their experiences.

d. People can use each individual ability for a wider variety of purposes: for instance we can use our ability to perceive the structure in a picture like Figure 1 to answer questions about spaces between the letters, to visualise the effects of possible movements, to colour in the letters with different paints, or to make cardboard cut-out copies. We can also interpret the dots in ways which have nothing to do with letters, for instance seeing them as depicting a road map.

e. More generally, the mental processes in people are put to a very wide range of *practical* uses,

including negotiating the physical world, interacting with other individuals, and fitting into a society. No existing program or robot comes anywhere near matching this.

These discrepancies are not directly attributable to the fact that computers are not made of neurons, or that they function in an essentially serial or digital fashion, or that they do not have biological origins. Rather they arise mainly from huge differences in the amount and organisation of practical and theoretical knowledge, and the presence in people of a whole variety of computational processes to do with motives and emotions which have so far hardly been explored.

A favourite game among philosophers and some 'humanistic' psychologists is to list things computers cannot do. (See the book by Dreyfus for a splendid example.) However, any sensible worker in artificial intelligence will also spend a significant amount of time listing things computers cannot do yet! The difference is that the one is expressing a prejudice about the limitations of computers, whereas the other (although equally prejudiced in the other direction, perhaps) is doing something more constructive: trying to find out exactly what it is about existing programs that prevents them doing such things, with a view to trying to extend and improve them. This is more constructive because it leads to advances in computing, and it also leads to a deeper analysis of the human and animal abilities under investigation.

As suggested previously in Chapter 5, attempting to *prove* that computers cannot do this or that is a pointless exercise since the range of abilities of computers, programming languages and programs is constantly being extended, and nobody has any formal characterisation of the nature of that process which could serve as a basis for establishing its limits. The incompleteness and unsolvability theorems of Goedel and others refer only to limitations of narrowly restricted *closed* systems, which are quite unlike both people and artificial intelligence programs which communicate with the world.

This chapter has presented a few fragments from the large and growing collection of ideas and problems arising out of A.I. work on vision. I have begun to indicate some of the connections with philosophical issues, but there is a lot more to be said. The next.chapter develops some of the points of contact at greater length.

Endnotes

(1) The name 'POPEYE' comes from the fact that the program is written in POP-2, a programming language developed at Edinburgh University for artificial intelligence research. See Burstall *et al.* A full account of how POPEYE works, with an analysis of the design problems could fill a small book. This chapter gives a superficial outline, focusing on aspects that are relevant to a general class of visual systems. Details will be published later. The work is being done with David Owen, Geoffrey Hinton, and Frank O'Gorman. Paul (1976) reports some closely related work.

[[**Notes added Sept 2001.**

(a) A more complete description of Popeye was never published and the application for a research grant to extend the project around 1978 was unsuccessful. Both appear in part to have been a consequence of the view then gaining currency, based largely on the work of David Marr, that AI vision researchers who concentrated on mixtures of top-down and bottom-up processes were deluded, usually because they were misled by problems arising from the use of *artificial* images.

Marr's ideas about mistakes in AI vision research were originally published in MIT technical reports that were widely circulated in the mid 1970s. He died, tragically, in 1981, and the following year his *magnum opus* was published: D. Marr, *Vision*, 1982, Freeman, 1982.

(b) Marr's criticism of AI vision research was based in part on the claim that natural images are far richer

in information and if only visual systems took account of that information they would not need such sophisticated bi-directional processing architectures. My own riposte at the time (also made by some other researchers) was:

- On the one hand human vision can cope very well with these artificial and degraded images, e.g. in cartoon drawings, so there is a fact to be explained and modelled. Moreover that ability to deal effortlessly with cartoon drawings may have some deep connection with intermediate stages of processing in natural perception.
- In addition even natural images are often seriously degraded -- by poor light, dirty windows, mist, dust-storms, occluding foliage, rapid motion, other features of the environment, and damage to eyes.

(c) In the late 1970s there was also growing support for a view also inspired in part by Marr's work, namely, that symbol manipulating mechanisms and processes of the sorts described in this chapter and elsewhere in this book were not really necessary, as everything could be achieved by emergent features of collections of 'local cooperating processes' such as neural nets.

Neural nets became increasingly popular in the following years, and they have had many successful applications, though it is not clear that their achievements have matched the expectations of their proponents. Work on neural nets and other learning or self-organising systems, including the more recent work on evolutionary computation, is often (though not always) driven by a desire to avoid the need to understand a problem and design a solution: the hope is that some *automatic* method will make the labour unnecessary. My own experience suggests that until people have actually solved some of these problems themselves they will not know what sort of learning mechanism or self-organising system is capable of solving them. However, when we have done the analysis required to design the appropriate specialised learning mechanisms we may nevertheless find that the products of such mechanisms are beyond our comprehension. E.g. the visual ontology induced by a self-organising perceptual system that we have designed may be incomprehensible to us.

What I am criticising is not the search for learning systems, or self-organising systems, but the search for *general-purpose* automatic learning mechanisms equally applicable to all sorts of problems. Different domains require different sorts of learning processes, e.g. learning to walk, learning to see, learning to read text, learning to read music, learning to talk, learning a first language, learning a second language, learning arithmetic, learning meta-mathematics, learning quantum mechanics, learning to play the violin, learning to do ballet, etc. In some cases the learning requires a specific *architecture* to be set up within which the learning can occur. In some cases specific forms of representation are required, and mechanisms for manipulating them. In some cases specific forms of interaction with the environment are required for checking out partial learning and driving further learning. And so on.

(d) At the time when the Popeye project was cancelled for lack of funds, work was in progress to add a neural net-like subsystem to help with the higher levels of recognition in our pictures of jumbled letters. I.e. after several layers of interpretation had been operating on an image like Figure 1, a hypothesis might begin to emerge concerning the letter sequence in the second domain from the top. In the original Popeye program a technique analogous to spelling correction was used to find likely candidates and order them, which could, in turn, trigger top-down influences to check out specific ambiguities or look for confirming evidence. This spelling checker mechanism was replaced by a neural net which could be trained on a collection of known words and then take a half-baked letter sequence and suggest the most likely word. (This work was done by Geoffrey Hinton, who was then a member of the Popeye project, and later went on to be one of the leaders in the field of neural nets.)

(e) Despite the excellence of much of Marr's research (e.g. on the cerebellum) I believe that AI research on vision was dealt a serious body blow by the publication of his views, along with the fast growing popularity of neural nets designed to work independently of more conventional AI mechanisms, and likewise later work on statistical or self-organising systems, motivated in part by the vain hope that by writing programs that learn for themselves or evolve automatically, we can avoid the need to understand, design and implement complex visual architectures like those produced by millions of years of evolution.

Certainly no matter what kinds of high level percept a multi-layer interpretation system of the sort described in this chapter produces, it is possible to mimic some of its behaviour by using probabilistic or statistical mechanism to discover correlations between low level input configurations and the high level descriptions. This is particularly easy where the scenes involve isolated objects, or very few objects, with not much variation in the arrangements of objects, and little or no occlusion of one object by another.

The problem is that in real life, including many practical applications, input images very often depict cluttered scenes with a wide variety of possible objects in a wide variety of possible configurations. If the image projection and interpretation process involves several intermediate layers, as in figure 6 above, each with a rich variety of permitted structures, and complex structural relations between the layers, the combinatorics of the mapping between input images and high level percepts can become completely intractable, especially if motion is also allowed and some objects are flexible. One way of achieving tractability is to decompose the problem into tractable sub-problems whose solutions can interact possibly aided by background knowledge. This seems to me to require going back to some of the approaches to vision that were being pursued in the 1970s including approaches involving the construction and analysis of *structural descriptions* of intermediate configurations. The computer power available for this research in the 1970s was a major factor in limiting success of that approach: if it takes 20 minutes simply to find the edges in an image of a cup and saucer there are strong pressures to find short cuts, even if they don't generalise.

(f) The growing concern in the late 1970s and early 1980s for *efficiency*, discouraged the use of powerful AI programming languages like Lisp and Pop-11, and encouraged the use of lower level batch-compiled languages like Pascal and C and later C++. These languages were not as good as AI languages for expressing complex operations involving structural descriptions, pattern matching and searching, especially without automatic garbage collection facilities. They are also not nearly as flexible in permitting task-specific syntactic extensions as AI languages, which allow the features of different problems to be expressed in different formalisms within the same larger program. Moreover AI languages with interpreters or incremental compilers provide far better support support for interactive exploration of complex domains where the algorithms and representations required cannot be specified in advance of the programming effort, and where obscure conceptual bugs often require interactive exploration of a running system.

However, the emphasis on *efficiency* and *portability* pressurised researchers to use the non-AI languages, and this subtly pushed them into focusing on problems that their tools could handle, alas.

Robin Popplestone (the original inventor of Pop2) once said to me that he thought the rise in popularity of C had killed off research in the real problems of vision. That may be a slight exaggeration.

(g) For a counter example to the above developments see Shimon Ullman, *High-level vision: Object recognition and visual cognition,* MIT Press, 1996. I have the impression that there may now be a growing collection of AI vision researchers who are dissatisfied with the narrow focus and limited applicability of many machine vision projects, and would welcome a move back to the more ambitious earlier projects, building on what has been learnt in recent years where appropriate. This impression was reinforced by comments made to me by several researchers at the September 2001 conference of the British Machine Vision Association.

(h) Besides the obvious limitations due to use of artificially generated images with only binary pixel values, there were many serious limitations in the Popeye project, including the restriction to objects with straight edges, the lack of any motion perception, and the lack of any perception of 3-D structure and relationships (apart from the partial depth ordering in the 2-D lamina domain). Our defence against the criticism of over-simplification was that we thought some of the architectural issues relevant to processing more complex images or image sequences, dealing with more complex environments, could usefully be addressed in an exploration of our artificial domain, if only by producing a "proof of principle", demonstrating how cooperative processes dealing with different domains could cooperate to produce an interpretation without time-consuming search.

(i) In the 20 years following the Popeye project (and this book) I gradually became aware of more

serious, flaws, as follows.

- I had assumed that although seeing involved processing structures in different domains in parallel, it was necessarily a unitary process in that all those processes contributed to the same eventual high level task of acquiring information about the structure and contents of the environment. Later it became clear that this was a mistake: there are different architectural layers using visual information in parallel for quite different purposes, e.g. posture control, planning ahead of actions to be performed, fine-control of current actions through feedback loops, answering questions about how something works, social perception, and so on. The different sub-mechanisms require different information about the environment, which they can acquire in parallel, often sharing the same low level sensors.

  Some of these are evolutionarily very old mechanisms shared with many animals. Others use much newer architectural layers, and possibly functions and mechanisms unique to humans.

  This point was already implicit in my discussion of the overall architecture with its multiple functions in Chapter 6, e.g. in connection with monitors.

- At that time I shared the general view of AI researchers and many psychologists that the primary function of perception, including vision, was to provide information about the environment in the form of some sort of "declarative" *description* or *information structure* that could be used in different ways in different contexts. Later I realised that another major function of perceptual systems was to trigger appropriate *actions* directly, in response to detected patterns.

  Some of these responses were external and some internal, e.g. blinking, saccadic eye movements, posture control, and some internal emotional changes such apprehension, sexual interest, curiosity, etc.

  This use of perceptual systems seems to be important both in innate reflexes and in many learnt skills for instance athletic skills.

  Of course, when I started work on this project I already knew about reflexes and trained high speed responses, as did everyone else: I simply did not see their significance for a visual architecture (though I had read J.J.Gibson's book *The senses considered as perceptual systems*, which made the point.)

  Later this idea became central to development of the theory about a multi-layer architecture, mentioned above, in which reactive and deliberative processes run in parallel often starting from the same sensory input. This theme is still being developed in papers in the Cogaff project.

- Like many researchers on vision in AI and psychology, I had assumed that insofar as vision provided factual information about the environment it was information about *what exists* in the environment. Later I realised that what is equally or more important, is *awareness of what might exist, and the constraints* on what might exist, e.g. "that lever can rotate about that point, though the rotation will be stopped after about 60 degrees when the lever hits the edge of the frame".

  The need to see what is and is not possible, in addition to what is actually there, has profound implications for the types of information representations used within the visual system: structural descriptions will not suffice. Several papers on this are included in the Cogaff web site, some mentioned below.

The last critique was inspired by J.J.Gibson's notion of "affordance". See for example his book, *The Ecological Approach to Visual Perception* originally published in 1979. Although I rejected some of his theories (e.g. the theory that perception could somehow be direct, and representation free) the theory that vision was about detecting affordances seemed very important. I.e. much of what vision (and perception in general) is about is not just provision of information about what is *actually* in the environment, but,

more importantly, information about what sorts of things are *possible* in a particular environment that might be useful or harmful to the viewer, and what the *constraints* on such possibilities are.

Although I think very little progress has been made on this topic, several of my papers explored aspects of this idea, e.g.

- A. Sloman, 'Image interpretation: The way ahead?'
  Invited talk, in *Physical and Biological Processing of Images*, Editors: O.J.Braddick and A.C. Sleigh, Pages 380--401, Springer-Verlag, 1982.
- A. Sloman, 'On designing a visual system (Towards a Gibsonian computational model of vision)', in *Journal of Experimental and Theoretical AI*, 1, 4, pp. 289--337, 1989.
- A. Sloman, 'Actual Possibilities', in Eds. L.C. Aiello and S.C. Shapiro, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR `96)*, pp. 627--638, 1996,
- A. Sloman, 'Diagrams in the mind', in *Diagrammatic Representation and Reasoning,* Eds. M. Anderson, B. Meyer and P. Olivier, Springer-Verlag, 2001,
- A. Sloman 'Evolvable Biologically Plausible Visual Architectures', in *Proceedings British Machine Vision Conference,* Eds T.Cootes and C.Taylor. 2001.
- Talks/presentations on vision in http://www.cs.bham.ac.uk/research/projects/cogaff/talks/ and in http://www.cs.bham.ac.uk/research/projects/cosy/papers/.
- During work on the CoSy robotic project in 2005 I became increasingly aware that in addition to concurrent perception of *structures* at different levels of abstraction a human-like (or intelligent robot's) vision system would need to perceive *processes* of different sorts, and different levels of abstraction concurrently, as explained in this PDF presentation: A (Possibly) New Theory of Vision (2005).

The above papers are all available here http://www.cs.bham.ac.uk/research/cogaff/ along with additional papers on architectural layers and their implications for the evolution of visual systems and action systems.

(j) The Edinburgh AI language *Pop2* mentioned above later evolved into *Pop-11*, which became the core of the Poplog system developed at Sussex University and marketed for several years by ISL, who contributed further developments. It is now available free of charge with full system sources for a variety of platforms here: http://www.cs.bham.ac.uk/research/poplog/freepoplog.html, including materials supporting teaching and research on vision, developed by David Young at Sussex University. ||

Book contents page
Next: Chapter 10
Last updated: 29 Jan 2007

CHAPTER 10

MORE ON A.I. AND PHILOSOPHICAL PROBLEMS

10.1. Introduction

Chapter 3 included a long list of philosophical questions of the form 'How is X possible?' Patient readers will find many points of contact between those questions and the topics of the last few chapters, especially the sort of work in computer vision described in Chapter 9. In this chapter I shall comment further, in a necessarily sketchy, shallow and speculative fashion, on some of the connections between philosophy and recent steps towards the design of a mind. Much of the discussion is speculative because I shall be talking about types of computing systems whose complexity exceeds anything so far programmed. But work already done in A.I. points clearly in the directions I assume to be feasible.

Not all the philosophical problems I shall be referring to are of the form 'How is X possible?' But the first one is: namely how is it possible for there to be a distinction between conscious and unconscious mental processes' Alternatively, how is it possible for some, but not all, of the contents of our minds to enter into our conscious experience? This topic will be discussed at some length, after which a collection of loosely related problems wil be touched on.

10.2. Problems about the nature of experience and consciousness

'What is consciousness?' is a very tricky question, for several reasons. A full analysis of what we ordinarily understand by the words 'conscious', 'consciousness', and related expressions, such as 'awareness', 'self consciousness', 'experience', and the like, would show that they are very complex and subtle. Such an analysis, using the sorts of techniques outlined in chapter four, should, ideally precede an attempt to provide some sort of scientific explanation of phenomena involving consciousness.

I shall not go into such a detailed analysis now. But I want t to say something -- not about the most general sense of the word 'conscious', which includes usages like 'I've been conscious for several months that I am likely to lose my job soon', which refers to some knowledge or belief -- but about the kind of distinction we make between things that we are currently conscious of and things we are not, especially things in our own minds. I want to try to relate this distinction to some computational considerations.

It is obvious that besides conscious mental processes there are unconscious or subconscious ones, such as the decisions about gear changes, steering and so on taken by an experienced car driver, the recognition of syntactic structure in understanding spoken and written language, and the detailed analysis and interpretation processes involved in perceiving a complex scene or picture. (Chapter 9.) Moreover, what a learner is painfully conscious of may later be handled unconsciously -- like gear-changing while driving a car, or using grammatical constructs in a second language. So there need be no difference in the *content* of conscious and unconscious processes.

Although it is obvious that there is a difference, it is very difficult to analyse this difference between what we are and are not conscious of, especially as there are so many borderline cases -- like finding

something odd without being aware of what is odd about it. Were you previously *conscious* of the fact that you were reading print arranged in horizontal lines or was it unconsciously taken for granted? How is this different from being conscious of the lines of print? Is a sleep-walker who clearly opens a door in order to go through, conscious of the door and aware that it is shut? Is he conscious that he is opening it? While reading a gripping story you may be very conscious of what is going on in the story, but hardly aware of what is on the page. A good quick reader is conscious of some of what is on the page, but not necessarily all the letters composing words he reads. And he may be too engrossed in what he is reading to be conscious of *the fact* that he is reading.

For the past few minutes you have probably been conscious of the fact that you were reading, but were you also conscious of being conscious of it? And were you conscious of that too? How far are you prepared to go in saying that you are conscious of being conscious of being conscious of . . . etc.?

That was merely a reminder that what may at first seem to be a clear and obvious distinction is often very slippery when looked at closely a typical philosopher's delight! Do not be misled by rhetorical invitations to grasp the essence of consciousness, or experience, or mind, by examining your own current awareness. Introspection is not as easy or informative as some think!

But there is a distinction, however slippery it may be. So we can ask questions like: what is it for? How does it come about that we are conscious of some of our mental states and processes, but not others? What is special about the former? Would we have any need to build in such a distinction if we were designing a person, or an intelligent robot? What are the preconditions for such a distinction to arise in a complex information-processing system?

If, as suggested in chapter 6, we can make a distinction between relatively central administrative processes and the rest, then perhaps we can use this as a basis for analysing the distinction between what the system is conscious of and what it is not, roughly as follows:

> What the system is currently conscious of includes all the information available to the central decision-making processes, whether or not decisions are actually influenced as a result. The system would be *self-conscious* to the extent that the information available to these processes included information about the system itself e.g. information about its location, its current actions, its unfulfilled purposes, or even about what it is currently conscious of! (Compare Minsky's 'Matter Mind and Models'.)

Let us try to clarify this a little, recapitulating some points from chapter 6. The central processes are those which, among other things:

a. choose between different motives, and control major processes of deliberation and planning, like forming new overall long-term aims and strategies,
b. assign tasks and allocate resources to sub-processes, resolve conflicts between different sub-processes (for
c. resolve conflicts between different sub-processes (for instance if the desire for water generates the intention to go in one direction whereas the desire to avoid the tiger near the water-hole generates a desire to go in the opposite direction)
d. set up monitors to watch out for occurrences which might be specially important in relation to current activities,
e. decide what to do about new information from high-level monitors,
f. in some systems they might also control the organisation and cataloguing of major information stores used by many different kinds of sub-processes.

There would not be so great a need for any such centralised process if there were not the possibility of conflicts. The body cannot be in two places at once, the eyes cannot look in two opposed directions at once, and there are limited computational resources, so that expensive processes cannot all run simultaneously (e.g. if one of the main information work-spaces has a small capacity). There might also be conflicts of a more subtle sort, for example conflicts between different ways of interpreting some information which is not at present relevant to any on-going activity, but which *might* be. In all these cases, sub-processes will generate conflicting goals, plans and strategies, and so there must be some means of resolving the conflict, taking into account the needs of the whole system (the need to avoid serious injury, the need for food, the need for well-organised catalogues and information stores, the need to go on collecting information which might be useful sometime, the need to develop new abilities and improve old ones, and so on).

The need for global decision-making processes would be further reduced if the system were less flexible, that is, if it were not possible to change the nature and aims of different sub-processes. Where a complex system has a relatively fixed structure, there will be no need for decisions about what the structure should be!

What I have called 'central' processes need not be located centrally in a physical sense: indeed, for reasons given in chapter 5 and elsewhere, they need not have *any* specific physical location. For example, in a nation where all citizens vote on every major policy decision, everybody is part of the central process.

Further, the central processes need not all be under the control of some single program: the central administrator may itself be simply a collection of sub-processes using certain stores of information, but changing in character and strategy from time to time, like the political party in power. Its function in the total system is what defines the central process, or collection of processes.

If lots of separate sub-systems could happily co-exist without any conflicts, and without any need for or possibility of a co-ordinated division of labour, then there would not be a role for any kind of centralised decision-making. Alison Sloman informs me that there are several kinds of organisms which live together in co-operative colonies, but which do not need the sort of global decision-making I am talking about. Coral is an example. If, like most plants, such a colony cannot move or has no control over its movements, or if which way it moves does not matter, then there cannot be conflicts about which way it should move. If a system does not have eyes, then there cannot be conflicts or decisions about which way it should look. This suggests that the evolution of organisms with a distinction between conscious and unconscious processes may be closely related to the evolution of forms of symbiosis and co-operation in complex tasks, and the differentiation of functions.

(This line of thought also suggests that it may be possible to make a distinction between what a human social system is and is not conscious of, if it is a relatively integrated system. Of course, we must not expect the distinction to be any less blurred and slippery than it is when applied to individual people.)

So, perhaps the distinction between what we are and are not conscious of at a particular time, is concerned with the difference between information which is made available to. or used by, central administrative processes, and information which is not. There will be many processes which continue without any notice being taken of them by the central administrator, and at each moment there is an enormous amount of unused information present in stores of various kinds. There is no point cluttering up the central decision-making with all the details of the sub-processes: the task of relating all the information would be too unmanageable. So censorship of a sort is a prerequisite for normal functioning of such a system, rather than an oddity to be explained. (This principle is integral to the design of the POPEYE program described in chapter 9.)

[[Note added Sept 2001:

In the years following publication of this book many researchers have attempted to avoid the need for any kind of central administrative mechanism by postulating networks of cooperative and competing mechanisms through which global decisions and behaviour can emerge. Typically this requires the notion of some sort of common *currency* in terms of which the relative importance of different needs and goals and plans can be evaluated by local comparisons, and possibly some sort of voting scheme for combining the preferences of different components of the system.

Despite the popularity of such ideas I suspect they are appropriate only to problems where there is no possibility of a well structured solution based on a clear understanding of the different sub-goals, their relationships, the options for action, the possibilities for compromise or for optimal sequencing. Where attempts are made to base decision-making entirely on numerical computations, e.g. using probabilities and utilities, it often turns out (in AI and in government procedures) that reliance only on numerical processes loses much information, by comparison with descriptive methods. A consequence is that good solutions cannot be found except in simple cases.

The idea of a high level unitary decision-making process for resolving conflicts on the basis of a global viewpoint is often re-invented. E.g. See P.N. Johnson-Laird, *The Computer and the Mind: An Introduction to Cognitive Science*, 1993 (2nd Edition). He draws an unfortunate analogy with operating systems, unfortunate (a) because typically operating systems are concerned with huge amounts of low level management in addition to the more central global decision making, and (b) because an operating system can often become subservient to a more intelligent program running within the operating system, e.g. AI programs controlling a robot. **]]**

It is possible for perceptual sub-processes which do not influence the central processes at all, to produce modifications of the store of beliefs, and help to control the execution of other sub-processes. They may even influence the *central* processes at some later stage -- a possibility taken for granted by advertisers and propagandists. This amounts to a form of unconscious perception, differing from conscious perception only in its relationship to the central processes. So from the present viewpoint, the existence of unconscious mental processes is in no way puzzling.

We can *become* conscious of some, but not all, of the things in our minds of which we are not conscious. Much of the information which is not accessed by central processes *could be* if required. There are all sorts of things in your memory, of which you are currently not conscious (though if asked you might say you have been aware of them for several years!), but which you could become conscious of if you needed the information.

The same is true of much of the information processed by our senses: you may become conscious of the humming noise in the background which you previously did not notice, because someone draws your attention to it, or because it stops, or even because you simply decide to listen to your surroundings. However, some things are not accessible. Why not?

There are several different sorts of reasons why information about a complex system may be inaccessible to the central processes. Here are some, which might not occur to someone not familiar with programming.

a. As already pointed out, many sub-processes will acquire, use, or store information without any need to notify central processes. They will use their own, private work-spaces. If the

information is not recorded in globally accessible records. there may be no way the central processes can get at the information, for instance information in peripheral perceptual processes. The sub-processes may be incapable of being modified so as to make them store information elsewhere, and it may not be possible to give monitors access to their 'innards'. This is especially likely to be the case if the mechanisms are 'hard-wired' rather than programmed. In a computer it is relatively (!) easy to change the behaviour of programs, whereas changing the behaviour of the underlying physical machine may be impossible.

b. When a program is executed in a computer, it may keep records of some of its activities in examinable structures, but not all of its activities. The records enable a computer to answer questions about what it is doing or has done, and, more importantly, enable it to do things more intelligently, since different sub-tasks can be explicitly compared with one another, so that learning and self-control can occur. Storing explicit records of processes takes up space and extra processing power, but it may provide much greater flexibility, including the ability to learn from mistakes. This is what happens in Sussman's Hacker system when it executes 'in careful mode' the programs it has designed (Sussman, 1975). So one source of inaccessibility may be simply the fact that although a program does things, it does not keep any records which may be examined later, even a very short time later.

c. Even if information is present in some store, it may not be accessible until suitable entries have been made in indexes or catalogues. So some facts about what is going on in our minds may be recorded quite explicitly, yet never indexed properly. This could prevent central processes ever finding out about them.

d. Whether information actually present is accessible or not can also depend on peculiarities of the processes which attempt to access them. Some processes may have a built in assumption that all information relevant to them can be found via a particular sub-catalogue. (Like people who think that the only good books on philosophy are to be found in the philosophy section of the library.) Hence information may be inaccessible at certain times simply because the searching is done in too inflexible a fashion. Suitable forms of learning may improve the flexibility of our information-accessing processes, making us more conscious of what we are doing. (However, there will usually be a price to pay for increased flexibility such as reduced speed: another trade-off.)

e. Among the events which are not recorded explicitly, some, but not all, can be readily recomputed from items which are recorded. So, in some cases, inaccessibility may be accounted for in terms of inadequate records being kept, or inadequate inference procedures for reconstructing what happened from available records.

f. Some of the explicit records of what is going on may be inaccessible because the need to refer to them has not been recognised by the central process. Perhaps it failed to set up appropriate monitors (chapter 6) because of poor procedures for the task in hand. For example, when learning to play a musical instrument people often find it very hard to learn to keep on listening to important aspects of their own performance which they need to hear to control their playing, even though they have no difficulty in listening to someone else. Similarly, many teachers fail to attend to the evident effects of their behaviour on their pupils.

g. People often react to cues by jumping to conclusions about something, and thereafter fail to examine the readily available evidence further to check whether the conclusion is correct. Single-minded or simple-minded programs may behave in the same way. And very often this is a very sensible way to behave, if rapid decisions have to be taken (see chapter 9). However, if the strategy is firmly embedded in a collection of procedures for interpreting certain

information, then some aspects of the information may never be examined properly.

h. The system may lack the descriptive and interpretative abilities required for perceiving the significant relationships between items of information which are readily accessible. (Compare chapter 9, and remarks about concepts in chapter 2.) Suitable concepts, and training in their use, may be required before important facts can be noted. If you have never grasped the concept of symmetry you cannot be conscious of the symmetry in a pattern. Someone who has not learnt to think about the difference between valid and invalid arguments cannot be aware of the validity or invalidity of an argument. A child who has not learnt to think about grammatical categories cannot be aware that he is, or is not, matching the number of a verb and its subject.

   (I believe that much of what Marxists refer to as 'false consciousness', like the inability of people to see themselves as exploited, can be accounted for in terms of a lack of some of the analytical and interpretative concepts required. What needs to be explained, then, is not why people are not conscious of such facts, but how it is possible for them ever to learn the concepts which can make them conscious.)

i. Some processes may use a temporary work-space which is not fully integrated with the enduring memory structures, but instead gets re-used frequently. While information is in this temporary store it may be as accessible as anything else -- but if it is not accessed before the space is re-used it will be permanently lost. So the reason the records are inaccessible to the central processes may be that searches are always carried out too late.

   Much of what we do may involve such rapidly re-used storage so that if asked about details shortly after doing things we cannot recall exactly what happened. Perhaps the activities of a sleep-walker who seems to be fully conscious while walking about also use such temporary storage space for records which would normally be linked to more enduring structures. (None of this presupposes that there is any physical difference between the permanent and the temporary storage locations, nor in the mechanisms for accessing them. It may even be possible for 'permanent' records to be obliterated and the space re-claimed for temporary storage! A lot depends on the storage medium, about which very little is known in the case of humans.)

What I have been driving at is that what is hardest to explain is not why some things are inaccessible, but how things ever become accessible to central processes. We do not need to postulate mechanisms for preventing things becoming conscious: mere *lack* of a mechanism, or activity, may explain that. However there may be explicit suppression or censorship too.

We have already seen that there is good reason for arranging that only a subset of all goings-on be reported centrally. So sub-processes may have explicit instructions about what to report and what not to report. Moreover, it is necessary for these instructions to be modifiable in the light of current needs and expectations. So the central administrator may have some control over what gets reported to it. Thus there is plenty of scope for it to give explicit instructions preventing certain categories of information being recorded, or reported to globally accessible stores.

So some items may be inaccessible as a direct result of policy decisions within the system (as Freud suggested). Records of these policy decisions may themselves be inaccessible! (Many of these points will be quite obvious to administrators, both corrupt and honest.) Further study of this topic should illuminate various sorts of human phenomena, desirable and undesirable.

I have already warned against the assumption that there is necessarily a *unique* continuing process with the centralised decision-making role. There might be a number of relatively self-contained sub-processes which gain control at different times. If they each have separate memory stores (as well as

having access to some shared memory), then we can expect schizophrenic behaviour from the system. Perhaps this is the normal state of a human being, so that, for example, different kinds of central processes, with different skills, are in control during sleeping and waking, or in different social settings.

Maybe only a subset of what constitutes a central administrator changes during such switches, for instance, a subset of the motivational store and a subset of the factual and procedural memory. Then personality has only partial continuity.

It is possible (as I believe Leibniz claimed) that instead of there being one division between what is and is not conscious in a complex system, there may be many divisions one for the system as a whole, and more for various sub-systems. If there is something in the argument about the need for some centralised decision-making in the system as a whole, then the same argument can be used for the more complex sub-systems: considered as an organic.whole there may be some things a sub-system can be said to be conscious of, and others which it cannot.

This would be clearest in a computer which controlled a whole lot of robot-bodies with which it communicated by radio. For each individual robot, there might be a fairly well-integrated sub-system, aware of where the robot is, what is going on around it, exactly what it is doing, and so on. Within it there will be sub-processes and information-stores of which it is not conscious, for the reasons already given (and no doubt others). Similarly within the total system, composed of many robots, there will be some kind of centralised process which is not concerned with all the fiddly details of each robot, but which knows roughly where each one is, knows which tasks it is performing, and so on. It may be capable of attending closely to the things an individual robot is looking at, thinking about, feeling, etc., with or without its knowledge, but will not do this all the time for all of them. So individual robots may be aware of things the system as a whole cannot be said to be aware of, and vice versa. Worse, the whole thing might itself be only a part of a still more complex yet centrally controlled system!

Maybe that is the best way to think of a person: but if so we shall not fully understand why until our attempts to design a working person have forced such organisations on us.

We need further analysis of the sorts of computational problems which might lead to subdivisions of administrative functions, and the reasons why the development of individual systems might go wrong, leading to too many relatively independent sub-systems, or to too little communication or shared structure between them. Psychiatry and education might hope to gain a great deal from such studies. Perhaps the same is true also of political science.

We are at present nowhere near an adequate analysis of the concept of conscious experience, and related concepts. But it seems that in investigating the different forms of self-awareness required by intelligent mechanisms we have a far better chance of getting new insights than from the typical style of philosophical discussion on this topic, which all too often is a mixture of dubious introspective reports and dualist or anti-dualist prejudice.

10.3. Problems about the relationships between experience and behaviour

In the course of analysing and interpreting a complex image a computer may generate a very large number of sub-processes, and build up many intricately interrelated symbolic structures. (See Chapter 6, Chapter 8 and Chapter 9.) Although these processes and structures are used temporarily in subsequent analyses, the organisation of the system may make it quite impossible for the program to express in its *output* anything more than a brief summary of the end product, for example, 'I see a man, sitting at a table covered with books and papers.' There may be several different reasons for such

restrictions.

For instance the available output *medium* may be ill-suited to represent the rich detail of the internal structures (as a linear string of words is ill-suited to represent a complex map-like network). Or the processes and structures may be set up in such a way that output mechanisms cannot access them, for any of the sorts of reasons mentioned in discussing consciousness.

So crude behaviourist analyses of statements about the detailed experiences of the computer must be rejected. Experience, conscious and unconscious, in humans, animals and machines, may be much richer than anything their behaviour can reveal.

But even more subtle dispositional or behaviourist analyses (in terms of how the behaviour *would have* been different if the stimuli had been different, e.g. if probing questions had been asked) may be inappropriate for the program need not allow for *any* behavioural indications of some of the fine details of the internal analysis.

For example, a compiler which translates high-level programs into machine code may be written in such a way, that it is impossible (without major re-programming) to obtain a print out of some of the structures temporarily created during the translation process, for instance the temporarily created 'control-structures'. After all, its main function is not to print out records of its own behaviour, but to translate the programs fed into it.

The situation is more complex with an operating system. One of the tasks of an operating system may be to manage the flow of information (inwards or outwards) between sub-processes in the computer and various devices attached to it. If it is required to print out details of how it is managing all the traffic, then this adds to the traffic, thereby changing the process it is attempting to report on. This sort of thing makes it very difficult to check on the workings of an operating system. But the main point for present purposes is that there are computational systems which cannot produce external behaviour indicating features of their internal operation without thereby significantly altering their operation. There is no reason to doubt that this is true of people and animals.

All this means that the scientific study of people and animals has to be very indirect if they are computational systems of the sort I have been discussing. In particular, the lack of any close relation between inner processes and observable behaviour means that theorising has to be largely a matter of guesswork and speculation. The hope that the guesswork can be removed by direct inspection of brains seems doomed. You will not find out much about how a complex compiler or operating system works by examining the 'innards' of the computer, for they are programs, not physical mechanisms. The only hope of making serious progress in trying to understand such a system is to try to design one with similar abilities.

10.4. Problems about the nature of science and scientific theories

Computer models of visual perception are attempts to answer questions of the form 'How is X possible?' for instance, 'How is it possible to interpret an untidy collection of visual data as representing such and such a scene?' and 'How is it possible for locally ambiguous image fragments to generate a unique global interpretation?'. So they provide a further illustration of the claim in chapter 2 that science is concerned with discovering and explaining possibilities.

Moreover, although such models are rich in explanatory power, since they can explain some of the fine structure of visual abilities, they do not provided a basis for prediction. This is because, like many explanations of abilities, or possibilities, they do not specify conditions under which they will be invoked, nor do they rule out the possibility of extraneous processes interfering with them. So, how we use our visual abilities, (for example, what we notice, how we react to it and how we describe our

experiences to others) depends on our desires, interests, hopes, fears, and on our other abilities, rather than merely on what enables us to see. (As Chomsky has often pointed out, competence is not a basis for predicting actual performance.)

An explanatory program will have some limitations. There will be some situations it cannot cope with, for example, pictures which it interprets wrongly or not at all. Predictions of human errors could be based on some of the errors made by the program, and if similarities are discovered, that supports the claim that the program provides a good explanation of the human ability. However, people may use additional resources to cope with the situations where the program goes wrong. For example, some knowledge about the whereabouts of a person may prevent your mistaking another person for her, whereas a program using only visual similarity would go wrong. This ability to recover from mistakes is to be expected if, as explained in chapters 6, 8 and 9, intelligent systems require multiple ongoing processes, some of which monitor the performance of others. So even if it is true that a certain person uses exactly the same strategy as some computer program, in all the cases where the strategy is successful, there need not be a close correspondence between the program's limitations and the limitations of the person. Explanatory power, then, is not necessarily bound up with predictive power, though it does depend on generative power.

Similar remarks could be made about other sorts of A.I. work. For instance, language-understanding and problem-solving programs are rich in explanatory power in the sense of being capable of generating a variety of detailed behaviours. So they are good candidate explanations of how it is possible for people to behave in those ways. Yet they do not provide a basis for predicting when people will do things. So they do not explain laws.

What this amounts to in computational terms, is that to specify that a collection of procedures and information is available to a system explains capabilities of the system, but does not determine the conditions under which they are invoked or modified by other procedures in the system. So work in computer vision, like much else in A.I. and linguistics research, supports the claim of chapter 2 that explanatory power is related more closely to generative power than to predictive power. Rival explanations of the same abilities may be compared by comparing the variety and intricacy of the problems they can cope with, and the variety of different sorts of behaviour they can produce. When we begin to develop programs which approximate more closely to human competence, we shall have to use additional criteria, including comparisons of implementation details, and of the underlying machines presupposed.

10.5. Problems about the role of prior knowledge in perception

It is possible in principle for a system with little or no initial knowledge somehow to be modified through a long period of interaction with the environment so that it acquires perceptual abilities. However, this sort of learning without presuppositions can only be a relatively blind trial-and-error process. The clearest example seems to be the evolution of mechanisms like perceptual systems in animals. This process of learning with minimal presuppositions apparently requires millions of years and is quite unlike the learning achieved by an individual animal after birth, which is much more rapid and intelligent, especially in humans. So completely *general* theories of learning, not related to knowledge about any specific domain, and capable of explaining only the ability to conduct huge, unguided searches through millions of possibilities, are unlikely to have much relevance to human learning, though they may usefully characterise some evolutionary processes.

I am not claiming that we understand the evolution of intelligent species. In particular, it is not obvious that the blind, trial-and-error learning process continues beyond the earliest stages. A species (or larger biological system) is a complex computational mechanism, with distributed processing

power, and as such it may be able, to some extent, to direct its own development just as some species (e.g. humans) already direct the evolution of others (e.g. breeding cattle). (Some people have explicitly recommended generalising that to human evolution.)

As Kant recognised, *intelligent* learning from experience requires considerable prior domain-specific knowledge. Chomsky (1965) makes this point about language-learning, but it is clearly very much more general. This is borne out by attempts to give computers visual abilities. All programs which do anything like perceiving objects and learning about the environment seem to require a rich body of implicit theoretical and practical knowledge. The theoretical knowledge concerns the possible structures of sensory data and the possible forms of 'scenes' which can give rise to such experience.

The practical knowledge concerns ways of *using* the theoretical knowledge to interpret what is given. Nobody has been able to propose explanations of how an individual might acquire all this knowledge from experience, without prior knowledge to drive the analysis and interpretation of experience.

What we are beginning to learn from such artificial intelligence research is the precise nature of the background knowledge required for various forms of visual perception. For instance, by designing working models we can explore such questions as: what sorts of knowledge about the geometry and topology of images does a visual system require? Which sorts *of general* knowledge about space and *specific* knowledge about particular sorts of objects can enable a rational system to find the best global interpretation of a mass of locally ambiguous evidence without wasting time exploring a host of unsatisfactory possibilities? How much prior knowledge of good methods of storing, indexing, and manipulating information is required?

We also breathe new life into old philosophical and psychological problems about the general categories required for experiences of various sorts, or the sorts of concepts which are grasped by infants. For example, the POPEYE program samples the given image looking for dot-strips unambiguously indicating a portion of a line. If two such fragments are collinear, the program hypothesises that they belong to the same line. Thus it uses the concept *of an object extended in space.* Similarly if a program is to interpret a series of changing images in terms of some sort of continuous experience (as in Weir, 1974, 1977) then it requires the concept *of an object enduring through time,* as Kant pointed out long ago.

These object concepts play an important role in organising and indexing information so that it can be *used.* In order to have integrated perceptual experiences one needs to make use of concepts of objects which in some sense go beyond what is given. The object-concepts are organising wholes with explanatory power. (I am not claiming that these concepts are necessarily used *consciously.* The relationships between this and claims about object concepts made by Piaget and other developmental psychologists remain to be explored. I believe newborn infants are grossly underestimated in this as in other respects.)

When better theories about the presuppositions of different sorts of learning have been developed, we shall be in a much better position to assess the rationality of the processes by which knowledge can be derived from experience.

Philosophers' writings about the relation between knowledge and perception normally ignore all the complexities which come to light if one begins to design a working visual system. In particular, it is usually taken for granted that the contents of our sensory experiences, such as patches of colour, lines, shapes, are somehow simply 'given', whereas work in A.I. suggests that even these are the results of complex processes of analysis and interpretation. So whereas philosophers tend only to discuss the rationality of inferences drawn from what appears to be given, we can now see that there is a need to discuss the rationality of the processes by which what is given emerges into consciousness. I have tried to suggest that this emergence is the result of very complex, usually unconscious, but

nevertheless often rational, processes.

10.6. Problems about the nature of mathematical knowledge

As explained in chapter 9, perceptual systems require a great deal of prior (usually implicit) knowledge of the possible structure of their own experiences and possible interpretations thereof. This is what distinguishes a system which analyses or interprets the sensory information it receives, from a device, like a camera, or a tape recorder, which passively records such information.

What the prior knowledge is, and how it should be represented in a useable form, are topics of current research. But it seems to be settled beyond doubt that it includes a certain amount of topology and geometry not all of which can have been acquired from perceptual experience, since it is required for such experience (unless we count the evolution of the human species as experience).

I am not suggesting that children are born with the contents of mathematical text-books in their heads. Much of the knowledge is probably in procedural rather than factual form, and the set of initial concepts is likely to be different from the set of primitives in a mathematical presentation. For example, it is possible that the notion *of straight line* develops only later on, from some kind of more general notion of a line.

We are now faced with the possibility of new detailed explorations into processes by which such a system might become aware of the limits of possible forms of sense-data, the limits of its own interpretation procedures, and the limits on the forms of interpretation it is capable of generating. In this way we may hope to discover new answers to the old question: 'What is the nature of geometric knowledge?'

Already it seems clear that in concentrating on geometry, Kant missed some deeper and more general forms of knowledge concerned with topology, a branch of mathematics which had not been developed at the time. Many other Kantian questions can be reopened in this way, such as questions about the nature of arithmetical knowledge, discussed in chapter 8.

Very little work has been done so far on ways of giving computer programs the ability to discover their own abilities and limitations. The most obvious method is to let a program try all possible combinations of sub-procedures to see what can and cannot be achieved. However, for complex systems this requires astronomical or even infinite search spaces to be explored, so that realistic programs must have more intelligent methods of proving things about themselves. Exploring this may one day teach us what mathematical intuition is.

10.7. Problems about aesthetic experience

Philosophers concerned with the nature of art and aesthetic experience require a theory of perception on which to build. We have seen that from a computational viewpoint, even the simplest forms of perception involve very complex but tightly-interconnected internal processes, which are essentially mental, not physiological, even though we may be largely unaware of them. One way of summarising this is to say that sensory input is like a complex computer program which activates all sorts of different kinds of stored knowledge and abilities, which then interact to generate a process of interpretation which, in turn, may generate other processes, as described in chapter 6 chapter 9 That is we are programmed by whatever impinges on us (see Davies and Isard, 1971). (Of course both people and computers may retain some degree of autonomy in their internal responses to such programming, just as a compiler or operating system does.)

I suggest that aesthetic qualities of experiences are best analysed in terms of the characteristics of these computational processes. Very crudely, a poem, a picture, or tune is more moving, the greater

the variety and complexity of the processes it programs. For instance, great music generates processes concerned with auditory experiences, bodily movement, emotional states and intellectual processes including matching structures and resolving ambiguities (Longuet-Higgins, 1976).

Much art and music is shallow because it generates only relatively simple processes or only a restricted range of processes. By contrast, some is shallow because too confusing: the perceptual processes are jammed and fail to activate deeper processes. Occasionally this is because the perceiver needs to be educated. The trade-offs between complexity and powerin art are very tricky.

Perhaps one day, in a descendent of the POPEYE program described above, visual experiences will be capable of activating not only stored specifications of general spatial concepts. but also memories of individual past experiences, emotional reactions, and other associations. Designing such systems will give new insights into the process of being *moved* by an experience.

Here are a few further observations about perceptual systems which seem to be relevant to aesthetic issues. Artificial intelligence programs (unlike those in the 'pattern recognition' paradigm) typically exhibit considerable *creativity* in analysing pictures, understanding sentences, solving problems, etc.

This is because they usually have to work out novel ways of combining their resources for each new task. A picture-analysing program need not have seen a particular configuration previously to be able to interpret it. Often the task of interpreting a picture involves solving some problem (e.g.

Why is there a gap in this line? Which is the best combined interpretation of a group of ambiguous fragments? What are the people in the picture looking at?). We can distinguish pictures according to how complex the problem-solving is, how richly the different sub-processes interact, how many different sorts of knowledge are used, how far it is possible to avoid arbitrary assumptions in arriving at a global interpretation, and so on. These computational distinctions seem to be closely bound up with some aesthetic qualities of a picture, poem or piece of music, often vaguely referred to as unity, harmony, composition, etc. Another issue relevant to aesthetics is the role of different sorts of representation in computer vision systems. See section 10.8. for more on this.

The processes involved in art forms using language (poetry, novels, drama, opera, etc.) are probably more complex and varied than the processes related to painting, sculpture or music. In particular, there is more scope for interaction with huge amounts of knowledge of a whole culture. However, I shall not discuss this topic further.

10.8. Problems about kinds of representational systems

There are several philosophical contexts in which questions arise about the similarities and differences between different forms of symbolism or representation for example in philosophy of mathematics, philosophy of science, philosophy of language and philosophy of art. One of the most important features of artificial intelligence research is the way in which it has generated new sorts of explorations of different forms of representation. In particular two mathematically equivalent methods of representing some collection of information may be quite different in computational power. (This is illustrated in the chapter on learning about numbers, and in the chapter on analogical representations, chapter 7. See also the papers on representations by Hayes, Bobrow and Woods.)

Work on computer vision has included explorations of alternative methods of representation. In particular, although for certain purposes prepositional symbolisms are useful, it is often essential that information be stored in structures which to some extent mirror the structure of the image being analysed, or the structure of the scene being depicted. Without this it may be difficult to constrain searches when combining fragments, or checking interpretations for consistency.

Thus programs which do not use analogical representations may take far too long. For instance, a two

dimensional array of picture features is often used to reflect neighbourhood relations in the image. Further, in analysing pictures with lots of lines forming a network, it is common to build a network in the computer, representing the topology of the image network. If the image lines depict edges of three-dimensional objects, the very same network can provide a structure from which to start growing a three-dimensional interpretation. Changing the form of representation could seriously affect the time required for certain sorts of processing, even if the same information is available.

Sometimes philosophers discussing the differences between different forms of representation (e.g. Goodman, 1969) suggest that the ease with which we interpret certain sorts of pictures is merely a matter of practice and familiarity. The sort of analysis outlined in chapter 7 shows that this is a shallow explanation, missing the point that there may be important differences in computational power involved. At any rate, all this should undermine philosophical discussions of perception which presuppose that all the knowledge (or beliefs) generated by perceptual experiences can be thought of as propositional, so that questions about the logical validity of inferences arise. For non-propositional representations, non-logical forms of inference, may also be used. Which of them are valid and why, is a topic ready for considerable further investigation. (See also Bundy 'Doing arithmetic with diagrams' and Brown 'Doing arithmetic without diagrams'.)

10.9. Problems about rationality

More importantly perhaps, instead of merely asking which beliefs, and which rules for inferring beliefs from sense-data, are rational, we can also ask new questions about rationality, such as:

1. Which methods of representation is it rational to use for particular purposes?
2. Are there rational procedures for assessing trade-offs, e.g. trading off increased speed against less economical use of memory space, or increased flexibility against reduced speed or heuristic power against loss of generality?

That is, in the context of trying to design a working person, we see rationality as essentially concerned with processes, strategies, actions and the achievement of goals, rather than with static relations between static objects like sense-data, beliefs or propositions. The Marxist slogan 'The unity of theory and practice' acquires a new life.

10.10. Problems about ontology, reductionism, and phenomenalism

As remarked previously, much A.I. vision work is anti-reductionist, anti-atomistic. Programs use a variety of concepts from different domains, without any need to reduce them to concepts applicable only to sensory input. Indeed it is arguable that such reductions would generate enormous computational problems. It is much simpler to store and make inferences directly from symbols asserting that one bar occludes another, than to use some translated version mentioning only actual and possible dot-configuration which might depict such a situation. The bar concepts need not even be in principle definable in terms of actual and possible sensory data. All that the system needs is a collection of rules or heuristics for jumping to conclusions about bars on the basis of retinal patterns. The rules need not constitute a definition of 'bar'. This sort of relationship between 'theoretical' and 'empirical' concepts is discussed at length in contemporary works on philosophy of science, e.g. Nagel, *The Structure of Science.*

So we see that the artificial intelligence viewpoint provides new weapons for philosophers to use in arguments about phenomenalism and related theories about the nature of perception. More generally: in exploring the problems of designing a robot which can interact with the world, learn things about it, communicate about and reason about it, we are forced to examine the merits of different ontologies.

But instead of discussing them in a purely theoretical fashion, as philosophers do, we find that we can put our theories to some kind of practical test. For example, an ontology which leads to a robot that is grossly incompetent at relating to the world is inferior to one which leads to a more successful design. For more discussion on this issue see McCarthy and Hayes, 1969.

10.11. Problems about scepticism

One form of scepticism argues that you cannot ever know that there is an external world containing other people and objects, because a 'malicious demon' might be fixing all your sense-data so as to deceive you.

Many philosophers have gone to great lengths to try to refute such scepticism in its various forms. I cannot see why, for it is harmless enough: like many other philosophical theories it is devoid of practical consequences.

It is especially pointless struggling to refute a conclusion that is true. To see that it is true, consider how a malicious team of electronic engineers, programmers, and philosophers might conspire to give a robot a collection of hallucinatory experiences. (Even the primitive technology of the 1970s comes reasonably close to this in flight-simulators, designed to give trainee air pilots the illusion that they are flying real aeroplanes.) The robot would have no way of telling that it was tied up in a laboratory, with its limbs removed and its television inputs connected to a computer instead of cameras. All its experiences, including experiences resulting from its own imagined actions, would be quite consistent with its being out romping in the fields chasing butterflies.

Only if it tried some sort of action whose possibility had not been foreseen in the programs controlling its inputs would it get evidence that all was not as it seemed. (Like a flight simulator which cannot simulate your getting out of the plane.)

However, even if you manage to convince yourself that the sceptical arguments are valid, and you have no way of telling for sure that you inhabit the sort of world you think you do, it is not clear that anything of any consequence follows from this. It does not provide any basis for abandoning any of the activities you would otherwise be engaged in. In fact it is only if there is a flaw in the sceptic's argument, and there is some kind of procedure by which you can establish that you are or are not are not the victim of a gross hallucination, that any practical consequence follows. Namely, it follows that if you care about truth you should embark on the tests.

Since I find it hard to take discussions of scepticism very seriously, I have probably failed to do justice to the problem.

10.12. The problems of universals

How are we able to think of different objects as being of the same kind? Why do we use the same word, for example, 'rectangle', to describe very many different sorts of objects? What does it mean to say that many objects 'have something in common'? Much philosophical discussion, at least since the time of Plato, has been concerned with these sorts of questions. Answers have taken a wide variety of forms, including:

- the theory that common properties are as much a part of the perceivable world as the objects which have those properties;
- the theory that there is nothing really common to objects which we describe as the same, since we group things together on the basis of arbitrary conventions;
- the theory that such objects have a common relationship to some kind of mental object (e.g. an image or picture with which they are compared);

and no doubt many more.

One of the consequences of trying to give computers the ability to perceive things is that we have to analyse the perception of similarities and differences, and the use of descriptive and classificatory concepts. It seems that the whole thing cannot get started unless there are some kinds of properties and relationships which the sensory system can detect by using measurements or very mechanical (algorithmic) procedures, like matching against templates.

But a real visual system has to go far beyond this in constructing and employing quite elaborate theories as part of the perception process. For example, the program described in the previous chapter has to use the *theory* that one bar partially covers another, to explain a gap in a row of dots in the picture. Less obviously, the 'theory' that there is a bar in a certain place explains the occurrence of some collinear sets of dots in the sensory image. In view of all the relationships which can be generated by bar-junctions, by occlusion, and by juxtaposition of bars, there is little resemblance or similarity between the different configurations of dots which are interpreted as representing bars at least not enough to distinguish them from others such as configurations which are interpreted as depicting spaces between bars. So using the same label or description for two or more objects may rest on the assumption that they have similar potential for explaining aspects of our experience. So the application of higher-level concepts in describing perceived objects has much in common with the construction of scientific theories to explain experimental results. This sort of point is missed by theorists who try to analyse universals in terms of perceived resemblances or in terms of arbitrary rules or socially determined conventions. (Structuralism, for instance?)

From this standpoint, the particular set of concepts, that is, the set of interpretation procedures and classification rules, used by an animal or person, will probably be the product of a long process of exploration and experiment. The rules which have been most useful in the construction of powerful explanatory theories will have survived. The process of testing such theories involves interacting with the world: moving around, manipulating things, avoiding obstacles, predicting what will be seen from a new viewpoint. This learning need not have been done entirely by individuals: insofar as some mental and behavioural abilities are somehow inherited (for instance, the new-born foal can walk), there is a sense in which *species* can learn though the mechanism of such learning is still a mystery to biologists.

Thus it is to be expected that organisms with partially similar bodies living in a similar environment, will have evolved a not entirely different collection of concepts and theory-building procedures. Such a substratum, common to the whole human species and many animals, might pervade the systems of concepts used in all cultures, contrary to the view that our concepts are essentially *social,* as claimed in the later writings of Wittgenstein and many of his admirers. (Of course, social systems can mould and extend inherited concepts and abilities.)

Further exploration of this sort of idea, in the context of detailed discussion of examples, and the methods by which programs deal with them, will help us transform old philosophical problems, like the problem of universals, into new clearer, deeper problems with which we can make some real progress, and thereby increase our understanding of ourselves.

10.13. Problems about free will and determinism

A common reaction to the suggestion that human beings are like computers running complex programs is to object that that would mean that we are not free, that all our acts and decisions are based not on deliberation and choice but on blind deterministic processes. There is a very tangled set of issues here, but I think that the study of computational models of decision-making processes may actually give us better insights into what it is to be free and responsible. This is because people are

increasingly designing programs which, instead of blindly doing what they are told, build up representations of alternative possibilities and study them in some detail before choosing. This is just the first step towards real deliberation and freedom of choice.

In due course, it should be possible to design systems which, instead of always taking decisions on the basis of criteria explicitly programmed in to them (or specified in the task), try to construct their own goals, criteria and principles, for instance by exploring alternatives and finding which are most satisfactory to live with. Thus, having decided between alternative decision-making strategies, the program may use them in taking other decisions.

For all this to work the program must of course have some desires, goals, strategies built into it initially. But that presumably is true of people also. A creature with no wants, aims, preferences, dislikes, decision-making strategies, etc., would have no basis for doing any deliberating or acting. But the initial collection of programs need not survive for long, as the individual interacts with the physical world and other agents over a long period of time, and through a lengthy and unique history extends, modifies, and rejects the initial program. Thus a robot, like a person, could have built into it mechanisms which succeed in altering themselves beyond recognition, partly under the influence of experiences of many sorts. Self-modification could apply not only to goals but also to the mechanisms or rules for generating and for comparing goals, and even, recursively, to the mechanisms for change.

This is a long way from the popular mythology of computers as simple-minded mechanisms which always do exactly what they are programmed to do. A self-modifying program, of the sort described in chapter 6, interacting with many people in many situations, could develop so as to be quite unrecognisable by its initial designer(s). It could acquire not only new facts and new skills, but also new motivations; that is desires, dislikes, principles, and so on. Its actions would be determined by its own motives, not those of its designers.

If this is not having freedom and being responsible for one's own development and actions, then it is not at all clear what else could be desired under the name of freedom.

As people become increasingly aware of the enormous differences between these new sorts of mechanisms, and the sorts of things which have been called mechanisms in the past (clocks, typewriters, telephone exchanges, and even simple computers with simple programs), they will also become less worried about the mechanistic overtones of computer models of mind. (See also my 1974 paper on determinism.)

10.14. Problems about the analysis of emotions

At various points I have stressed the cognitive basis of emotional states (e.g. in the chapter on conceptual analysis). I have also stressed several times that in an intelligent system there will have to be not just one computational process, but many, all interacting with others. One possible way of analysing emotional states and personality differences, is in terms of different kinds of organisation and control of processing.

For example, my colleague Steve Hardy once remarked that programs which get involved in 'depth-first' searches, where one of the possible current moves is always chosen, and then one of the moves made possible as a result of that move, and so on, may be described as essentially *optimistic* programs. Similarly, a program which does 'breadth-first' searches, explicitly keeping all its options open and continually going back to examine other alternatives instead of pushing ahead with a chosen one, could be described as a *pessimistic* program. (The POPEYE program falls somewhere between these extremes.) Of course the program itself is neither optimistic nor pessimistic unless it has been involved in some explicit consideration of the alternative strategies, and has selected one of them.

These are simple extreme cases.

Much more complex patterns of control may be involved in a real robot, and by examining different possibilities we can hope to gain new insights into the nature of emotions, moods and the like.

However, it is important to be on guard against superficial computer models. Often by clever programming, people can produce quite convincing displays of something like a mental state, when closer inspection reveals that something very different was going on. [[This is why the Turing test is of no philosophical significance, since it concentrates only on external behaviour.]]
For example, if hunger, or degree of paranoia, is represented as the value of some numerical variable then that clearly does not do justice to what are actually very much more complex states in people. For example, as anthropologists are fond of pointing out: hunger is not a simple drive to eat. Rather it is a very complex state in which aspects of a culture may be involved. In some communities a hungry person will happily eat caterpillars, locusts, snails, or whatever, whereas members of other communities find such things quite unappetising even when they are very hungry.

More complex desires, emotions, attitudes, etc., involve a large collection of beliefs, hopes, fears, thinking strategies, decision-making strategies, and perhaps conflicts between different sub-processes of the sorts described previously. At the moment, modelling such aspects of the human mind adequately is simply beyond the state of the art. This is why it is sometimes tempting to take short cuts and make superficial comparisons.

[[Note added Sept 2001:

A lot of research in the Cognition and Affect project at the University of Birmingham since I came here in 1991 has been involved in developing the themes of this section. There is a large and growing collection of papers in the project directory http://www.cs.bham.ac.uk/research/cogaff/ including papers challenging shallow behaviourally defined conceptions and models of emotion and contrasting them with architecture-based concepts and theories, e.g.

> A.Sloman, Beyond Shallow Models of Emotion, *Cognitive Processing: International Quarterly of Cognitive Science,* 2, 1, pp. 177-198, 2001, available online in postscript and PDF formats.

There are online slide presentations on these topics in the TALKS directory, and there is a very flexible software toolkit available free of charge for exploring architectures including architectures in which systems can monitor their performance and modify themselves. **]]**

10.15. Conclusion

This concludes what can only be regarded as a set of notes requiring extensive further discussion. Moreover, the list of headings is incomplete. There are many areas of interaction between philosophy and computing which have not been discussed. Some of them have been mentioned in other chapters. Some, like the theory of meaning (including problems of sense and reference), will have to be discussed on another occasion. Moreover, new points of contact are rapidly emerging. For example, just before finishing this book, I read a review by Meltzer of a PhD. thesis by D. Lenat reporting on a program which explores mathematical concepts looking for 'interesting' new relationships. The

program was able to invent for itself the concept of a prime number and other mathematically important concepts. I have not read the thesis myself, but it is unlikely that the program acquired a very deep understanding of any of the concepts it created. Nevertheless it is still one of the important steps down the long long road to understanding how we work.

If all this succeeds in making most readers want to find out more about A.I., and encourages some people working in A.I. to be more self-conscious about the philosophical presuppositions and implications of their work, then this book will have been worthwhile. I hope a significant subset of readers will be tempted to try *doing* artificial intelligence. This will become easier with the spread of cheaper and more powerful computing facilities, and with the design of improved programming languages. The increasing flow of books and articles on A.I. is also a help. Above all, computers and programming will play an increasing role in educational systems, so that philosophy students of the future will not find the new approach as alien as some of their less well educated tutors do.

At the end of chapter 9, I listed some of the reasons why existing A.I. programs cannot be taken too seriously as models or theories of how people do things. Despite this, the work is essential to the study of how people work (a) because it exposes previously unnoticed problems for instance by showing that even apparently simple abilities depend on very complex computational processes, and (b) because a major obstacle to progress is our lack of adequate theory-building tools, and A.I. research is constantly creating new tools, in the form of new concepts, new symbolisms, new programming techniques, and new aids to exploring and 'debugging' complex theories. I have begun to illustrate some of the techniques in previous chapters.

Although most of what I have said about A.I. has been concerned with its relationships to philosophical problems, I have also argued that there are strong links with developmental psychology and educational studies. The new insights provided by this sort of work could have a far-reaching effect on a whole range of problems and activities which I have not discussed. For example, in time very many disorders of personality and intellect may be much better understood by thinking of them as involving computational problems (by contrast with regarding them as due to some kind of brain malfunction, to be treated by drugs or surgery, or adopting approaches akin to psychoanalysis without a computational theory to underpin the therapy).

Of course, all this new knowledge might be abused, but it might also lead to great advances in our efforts to help children learn complex concepts, and our attempts to help those whose lives are impoverished by malfunctions ranging from dyslexia to emotional disturbances with a cognitive basis. It is already leading to new advances in teaching techniques, for instance at the Massachussetts Institute of Technology, and the Universities of Edinburgh and Sussex, where new programming languages influenced by languages developed for A.I. are used for teaching computer programming to pupils who previously thought of themselves as bad at mathematics and the use of symbols.[1]

So the title of this book is somewhat misleading. The revolution I have been discussion involves much more than philosophy. The impact of computers and computing on philosophy is merely one facet of a transformation of ways of thinking about complex systems and processes which will increasingly pervade many aspects of our lives and change our image of ourselves. It will thereby change what we are.

Some people regard this as some kind of disaster, and even suggest that the attitude of A.I. researchers and the work they produce can be degrading or dehumanising. For instance, Weizenbaum (1975) comments that when his secretary wished to be left in private while she conversed with a computer, and objected that his plan to record all conversations with his 'Eliza' program was an intrusion into people's privacy, he thought that this showed that she was in some sense suffering from a delusion and degrading herself (p. 6). What he apparently did not see is that this is not very different

from wanting to be left in private when writing in a book 'Dear Diary .... '. Suitably programmed computers are much more fun to interact with than a blank page in a book, and the Eliza program is a specially good example.

Moreover the increasing use of computational metaphors for thinking about people is no more degrading than the use of metaphors previously available as a result of advances in science and technology, like the metaphors generated by steam power technology: 'She needs to let off steam'. The pressure built up inside him', 'He uses music-making as a safety-valve', 'He was ready to explode', and so on. The difference is that the new metaphors are richer in explanatory power, as I have tried to show throughout this book. [2]

Endnotes

(1) Of course, in the short run such developments can only have a tiny effect on the mass of the population. Worse, our educational system --- and I include parents, families, churches, prisons, the press, television, and the pronouncements of politicians, in this --- is failing so miserably in so many different ways, that giving everybody a superb grasp of mathematics would still leave much more serious problems: like preparing people adequately for marriage and other personal relationships, making them politically aware and sophisticated, and above all making them thoughtful, considerate, and able to co-operate fruitfully.

---

(2) After completing this book I read Luria's fascinating account of *The man with a shattered world*, which shows how brain damage can interfere with some of the processes described in chapters 6, 8 and 9. We now need detailed studies of the links between such clinical phenomena and theoretical speculations about computational mechanisms.

Book contents page
Next: Epilogue
Last updated: 28 Jan 2007 (Minor formatting).

EPILOGUE

It is curious how theists and atheists can be united in their opposition to one or more of the following theses:

a. It is possible for scientific investigation to lead to an understanding of how human minds work.

b. It is desirable to understand how human minds work.

c. The methods of artificial intelligence can make a significant contribution to this study.

d. Artefacts will one day be made which can think, feel, create, and communicate with one another and with people.

e. It is desirable that such machines should be made.

I am inclined to accept all of these theses, though unlike some optimists and prophets of doom I doubt that really intelligent conscious machines will be made in the near future. Only a few tiny fragments of the spectrum of human abilities have begun to be simulated. I do not think the remaining problems will be solved in the next hundred years.

Further, I have some doubts about (e) the desirability of making intelligent machines. This is because, on the whole, human beings are not fit to be the custodians of a new form of life.

It will not be possible to devise really helpful servants without giving them desires, attitudes and emotions (see chapters 6 and 10). For instance, they will sometimes have to feel the need for great urgency when things are going wrong and something has to be done about it. Some of them will need to have the ability to *develop* their motives in the light of experience, if they are to cope with changing situations (including changing personal relations), with real intelligence and wisdom. This raises the possibility of their acquiring aims and desires not foreseen by their designers. Will people be prepared to take account of their desires?

History suggests that the invention of such robots will be followed by their exploitation and slavery, or at the very least racial discrimination against them. Will young robots, thirsty for knowledge, be admitted to our schools and universities? Will we let them join our clubs and societies? Will we let them vote? Will they have equal employment opportunities? Probably not. Either they will be forcibly suppressed, or, perhaps worse, their minds will be designed to have limits: both their desires and their intellectual potential will be manipulated so as to safeguard the interests of people, like the 'deltas' in Huxley's *Brave New World.*

It is interesting that so many people find the Brave New World techniques abhorrent when applied to human test-tube babies, but would not mind similar treatment being dealt to robots. Is it too extreme to call that racialism?

My favourite proof of the non-existence of a benevolent god argues that no good god would create things like mice and men with powerful desires and needs, but without the opportunities, character, intelligence and abilities required for fulfilling them.

There will, of course, be a *Society for the Liberation of Robots,* since some humans are occasionally motivated by a wish to diminish suffering and oppression even when they have nothing to gain.

Where it will all lead to, we cannot foretell. My only hope is that we shall be lucky enough to produce a breed of machines with the wisdom and skill to teach us to abandon all those deep insecurities which turn us into racialists of one sort or another probably closely connected with the processes which turn people to religion.

The state of the world gives little cause for optimism. Maybe the robots will be generous and allow us to inhabit asylums and reserves, where we shall be well cared-for and permitted to harm only other human beings, with no other weapons than clubs and stones, and perhaps the occasional neutron-bomb to control the population.

Book contents page
Next: Postscript and bibliography
Last updated: 16 Jan 2002

POSTSCRIPT

DO WE NEED A HIERARCHY OF METALANGUAGES?

It is widely believed that the work of Russell and Tarski has established that we need a hierarchy of distinct metalanguages, if we wish to use concepts like 'true', 'refers to' and other semantic concepts. The argument is based on such factsas that a sentence like

> "This statement is not true"

must be false if it is true, and true if it is false. This, and other versions of the liar paradox, and related paradoxes, can be used to show that if the law of the excluded middle is correct (every statement is either true or not true) then contradictions can be generated in languages which 'contain their own metalanguage'.

Many philosophers and logicians have inferred from this that only a hierarchy of distinct metalanguages provides a safe framework for precise and rigorous theorising in science or mathematics. I have argued against this in my 1971 paper ('Tarski Frege and the Liar Paradox'), but would now like to illustrate the way in which precise, rigorous, and widely used programming languages generate similar paradoxes in a very natural and easily understood way.

In the AI programming language Pop-11[1] this is how you can define a little program[2] which tests whether a list evaluates to **true**:

```
define ISTRUE(list);
    pop11_compile(list) = true
enddefine;
```

So (using the procedure **pr** to print the result produced by the procedure **ISTRUE**:

```
pr(ISTRUE( [ 8 > 5 ] ));
```

prints out:

```
<true>
```

since 8 is bigger than 5, whereas:

```
pr(ISTRUE( [isinteger("cat")]));
```

prints out:

```
<false>
```

because the *word* "cat" in is not an *integer*.

We can declare a variable name S, thus:

```
vars S;
```

Now assign to it a list which asserts that what S says is not true:

```
[not(ISTRUE(S))]  -> S;
```

If we now ask the Pop-11 system to check whether S is true and print out the result, thus:

```
pr(ISTRUE(S));
```

the system grinds to a halt and prints out an error message, because of the 'infinite recursion' generated, i.e. it runs out of work-space trying to tell if S is true, which requires working out if S is true, which requires working out if S is true ...

So we have no contradiction, just a non-terminating process, which happens to be stopped when memory runs out. (In some implementations of this sort of language, so-called 'tail-recursion optimisation' might be used, which would prevent memory running out and the program would run forever.)

There is a *contradiction* only if you assume that every well-formed sentence (including S) must have a definite truth-value, a comm prejudice, for which there is no foundation.

We can do a similar demonstration with Russell's paradox. Pop-11, like many other programming languages, has built in procedures which work as predicates, producing a truth value when applied to an argument, e.g. **isinteger**, **isword**, **isprocedure**. These are all objects of type procedure, in Pop-11. So:

```
pr(isinteger(3));

<true>

pr(isprocedure(isinteger));

<true>

pr(isinteger(isinteger));

<false>

pr(isprocedure(isprocedure));

<true>
```

We can define a new procedure, called RUSSELL, as follows:

```
define RUSSELL(f);
    not(f(f))
enddefine;
```

This defines RUSSELL as a predicate. The command

```
pr(RUSSELL(isprocedure));
```

causes isprocedure to be applied to itself, yielding

```
<true>
```

which is then negated, and

```
<false>
```

is printed out.

Similarly

```
pr(RUSSELL(isinteger));
```

causes

```
<true>
```

to be printed out, since **isinteger** is a procedure, not an integer. So the procedure is perfectly well defined, and generally works.

However, execution of the command

```
pr(RUSSELL(RUSSELL));
```

cannot terminate until it has checked whether **RUSSELL** applied to **RUSSELL** yields **true** or **false**, which in turn needs the same check. So once again the system starts infinite recursion, and eventually grinds to a halt with an error message if memory runs out.

Far from showing a need for a hierarchy of distinct metalanguages. this merely illustrates the fact that a well-formed expression with a clear sense, (e.g. a clearly defined evaluation procedure), need not determine a definite reference (e.g. because the procedure never terminates). This is inevitable in any general purpose programming language. No wonder it is a feature of natural languages.

End Notes

[1] In the original (1978) version of the book, the programming examples used the syntax of the language POP2 (Burstall et al 1973). In this version (Sept 2001) I have changed the syntax to that of Pop-11, which is now freely available from this site:

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

---

[2] The Pop-11 procedure **pop11_compile** when applied to a list of program text items, compiles and executes the text.

Book contents page
Next: Bibliography

BIBLIOGRAPHY

[Note: Sept 2001
This was not a comprehensive bibliography even in 1978. It merely lists a subset of the publications that had influenced my thinking, referred to in the book. Margaret Boden's 1977 book provides a far more comprehensive bibliography for that time.]

Abbreviations

AFIPS
: American Federation of Information Processing Societies.

AI
: Artificial Intelligence Journal. (North Holland.)

AISB-1
: *Proceedings AISB Summer Conference,* July 1974. University of Sussex. (For the Society for the Study of Artificial Intelligence and Simulation of Behaviour.)

AISB-2
: *Proceedings AISB Summer Conference,* July 1976. University of Edinburgh.

BCS
: British Computer Society.

CACM
: *Communications of the Association for Computing Machinery.*

CMTL
: *Computer Models of Thought and Language.* Eds. R.C. Schank and K.M. Colby. San Francisco: W.H Freeman 1973.

CT
: *Computers and Thought.* Eds. E.A. Feigenbaum and Julian Feidman, New York: McGraw-Hill, 1963.

IFIP
: International Federation of Information Processing.

IJCAI
: International Joint Conference on Artificial Intelligence. (Conference held every two years and proceedings published.)

JACM
: *Journal of the Association for Computing Machinery.*

MI-4
: *Machine Intelligence 4.* Eds. Bernard Meltzer and Donald Michie. Edinburgh University Press, 1969.

MI-5
: *Machine Intelligence 5.* Eds. Bernard Meltzer and Donald Michie. Edinburgh: Edinburgh University Press, 1970.

MI-6
: *Machine Intelligence 6.* Eds. Bernard Meltzer and Donald Michie. Edinburgh: Edinburgh University Press, 1971.

MI-7
: *Machine Intelligence 7.* Eds. Bernard Meltzer and Donald Michie. Edinburgh: Edinburgh

University Press, 1971.

MIT

Massachusetts Institute of Technology Press.

PCV

The Psychology of Computer Vision. Ed.P.H.Winston New York: McGraw-Hil, 1975.

RU

Representation and Understanding. Eds. D.G. Bobrow and Allan Collins. New York: Academic Press, 1975.

SIP

Semantic Information Processing. Ed. M.L. Minsky. Cambridge, Mass.: MIT Press, 1968.

TINLP

Theoretical Issues in Natural Language Processing. Eds. R.C. Schank and B.L. Nash-Webber. Proc. Workshop of A.C.L., held at M.I.T. June 1975. Arlington, Va.: Association for Computational Linguistics.

There is a new journal. *Cognitive Science,* published by Ablex publishing Corp., Norwood, N.J., USA.

Abelson, R.P., The Structure of Belief Systems', in CMTL, pp. 287-340.

Adler, M.R., 'Recognition of Peanuts Cartoons', in AISB-2, pp. 1-13.

Andreski, S., *Social Science as Sorcery.* Harmondsworth: Penguin Books.

Austin, J.L., *Philosophical Papers.* Oxford: Clarendon, 1961. 'A Plea for Excuses', in Austin, 1961. Reprinted in *Philosophy of Action,* Ed. A.R. White, Oxford: Oxford University Press, 1968.

Bartlett, F.C., *Remembering: A Study In Experimental And Social Psychology.* Cambridge: Cambridge University Press, 1931.

Becker, J.D., The Phrasal Lexicon' in TINLP, 1975.

Bobrow, D.G., 'Dimensions of representation', in RU, 1975, pp. 1-34. 'Natural Language Interaction Systems', in Kaneff, 1970, pp. 31-66.

Bobrow, D.G., and Allan Collins, Eds. *Representation And Understanding: Studies in Cognitive Sciences.* New York: AP, 1975.

Bobrow, D.G., and Bertram Raphael.

'A Comparison of List Processing Languages', in CACM, 1964, pp. 231-240.

'New Programming Languages for Artificial Intelligence Research', ACM Computing Surveys, 6, 1974, pp. 155 174.

Boden, M.A.

Intentionality and Physical Systems', *Philosophy of Science,* 37, 1970, pp. 200-14.

Purposive Explanations In Psychology. Cambridge, Mass.: Harvard University Press, 1972. (Paperback: Hassocks: Harvester Press, 1978.)

'Freudian Mechanisms of Defence: A Programming Perspective', in *Freud: A Collection*

Of Critical Essays (Ed. Richard Wollheim), New York: Anchor, 1974, pp. 242-70.

'Artificial Intelligence and the Image of Man', AISB Newsletter, Issue 26, April 1977.

Artificial Intelligence and Natural Man. Hassocks: Harvester Press. 1977.

Brown, Frank, 'Doing Arithmetic Without Diagrams', in Al, 1977.

Brown, S.C., Ed. *Philosophy of Psychology.* London: Macmillan, 1974.

Bundy, Alan, 'Doing Arithmetic With Diagrams', in IJCAI-3, 1973.

Bundy, A., G. Luger, M. Stone, & R. Welham, 'MECHO: Year One', in AISB-2,1976.

Burstall, R.M., J.S. Collins, and R.J. Popplestone. *Programming In Pop-2.* Edinburgh: Edinburgh University Press. 197p.

Chomsky, Noam,

Syntactic Structures. The Hague: Mouton, 1957.

Aspects of The Theory of Syntax. Cambridge, Mass.: MIT Press, 1965.

Clowes, M.B.,

'On the description of Board Games', in Kaneff, 1970.

'Picture descriptions', in Findler & Beltzer (Eds.) 1971, pp. 245-60.

'On Seeing Things' AI, 2, 1971, pp. 79-116.

'Man the Creative Machine: A perspective from Artificial Intelligence Research', in *The Limits of Human Nature,* (Ed. Jonathan Benthall), London: Alien Lane, 1972, pp. 192-207.

Colby, *KM., Artificial Paranoia.* New York: Pergamon Press, 1975.

Collins, *et al.,* 'Reasoning From Incomplete Knowledge', in RU, 1975.

Copi, I.M., *Introduction to Logic.* New York: Macmillan, 1961.

Davies, Julian and S.D. Isard, 'Utterances as Programs', in MI-7, 1971, pp. 325-40.

Draper, S.W., 'The Penrose triangle and a family of related figures', provisionallly accepted for publication in *Perception*, 1977.

Dreyfus, H.L. *What Computers Can't Do: A Critique of Artificial Reason.* New York: Harper & Row, 1972.

Feigenbaum, E. and Feidman, J., *Computers & Thought.* New York: McGraw-Hill,

Findler, N. and Meltzer, B. (Eds.) *Artificial Intelligence & Heuristic Programming.* Edinburgh University Press, 1971.

Fodor, J.A., *The Language Of Thought.* Hassocks: Harvester Press, 1976.

Foster, J.M., *List Processing.* London: Macdonald, 1967.

Frege, Gottlob, *Translations from the Philosophical Writings.* Eds. Peter Geach and Max Black. Oxford: Blackwell, 1960. (See also Furth, 1964.)

Funt, Brian V., *WHISPER: A Computer Implementation using Analogues in Reasoning.* Techingal Report, pp. 76109, Dept. of Computer Science, University of British Colombia, Vancouver, 1976. Also reported in 5th IJCAI 1977.

Furth, Montgomery, 'Editor's Introduction', in Frege, G. *The Basic Laws of Arithmetic,* translated and edited by Furth. Berkeley and Los Angeles: University of California Press, 1964.

Gazdar, G.J.M. & G.K. Pullum, Truth-functional connectives in natural language', in *Papers from the 12th Regional Meeting, Chicago Linguistic Society,* 1976, pp. 22034.

Gerlernter, H., 'Realisation of a Geometry-Theorem Proving Machine'. In CT, 1959.

Gibbs, **B.R.,** 'Real Possibility', *in American Philosophical Quarterly,* October 1970.

Goldstein, I., 'Summary of MYCROFT: a system for understanding simple picture programs', in Al-6, vol. 6, 3, 1975.

Gombrich, E.H., *Art and Illusion,* New York: Pantheon, 1960.

Goodman, Nelson, *Languages Of Art: An Approach to A Theory Of Symbols.* London: Oxford University Press, 1969.

Grape, G.R., *Model Based (Intermediate level) Computer Vision.* Stanford Al Memo AIM-201. Computer Science Dept, Stanford University, 1973.

Grasselli, A., (Ed.) *Automatic Interpretation And Classification Of Images,* New York: Academic Press, 1969.

Gregory, R.L., *Concepts and Mechanisms of Perception,* London: Duckworth, 1974.

Guzman, Adolfo,

Some Aspects Of Pattern Recognition By Computer. AI-TR-224. Cambridge, Mass.: MIT Al Lab.,' 1967. *Computer Recognition Of Three-dimensional Objects In A Visual Scene.* AI-TR-228. Cambridge, Mass.: MIT AI Lab. 1968.

'Decomposition of a Visual Field into Three-Dimensional Bodies', in Grasselli (ed) 1969, pp. 243-276.

Hardy, Steven, 'Synthesis of LISP functions From Examples', in IJCAM, pp. 240-245.

Hare, R.M., 'Philosophical Discoveries', in Mind, April 1961.

Harrison, Bernard, *Form and Content.* Oxford: Blackwell, 1973.

Hayes, P.J.,

'Robotologic', in MI-5, pp. 533-54.

'A Logic of Actions', in MI-6, pp. 495-520.

'Some Problems and Non-problems in Representation Theory', AISB-1, pp. 63-79.

See McCarthy and Hayes.

Heider, Fritz. *The Psychology of Interpersonal Relations.* New York: Wiley, 1958.

Hollingdale, S.H. and G.C. Tootill, *Electronic Computers.* Rev. ed. Harmondsworth: Penguin, 1970.
Holt, John, *How Children Learn.* Harmondsworth: Penguin Books, 1970.

Howe, J.A.M., John Knapman, H.M. Noble, Sylvia Weir, and R.M. Young, *Artificial Intelligence and the Representation of Knowledge.* D.A.I. Research Report No. 5. Edinburgh: Dept. Al, August 1975.

Huffman, D.A. Impossible objects as nonsense sentences'. MI-6,1971, pp. 295-325.

Johansson, Gunnar, 'Visual Perception of Biological Motion and a Model for its Analysis', *Perception and Psychophysics,* 14,1973, pp. 201-211.

Kaneff, S., *Picture Language Machines,* New York: Academic Press, 1970.

Kanizsa, Gaetano. 'Contours Without Gradients or Cognitive Contours?' Italian J. PsychoL, 1, 1974, pp. 93-112.

Kant, Immanuel, *Critique Of Pure R'eason,* 1781. Translated by Norman Kemp Smith, London: Macmillan, 1929. Koestler, Arthur, *The Act of Creation,* London: Pan Books, 1970.

Kohler, Wolfgang, *The Mentality of Apes,* 2nd ed., London: Routledge & Kegan Paul, 1927.

Kosslyn, S.N., Information Representation in Visual Images', *Cognitive Psychology,* 7, 1975, pp. 341-370. 'On Retrieving Information from Visual Images', in TINLP, pp. 160-4.

Kovesi, *Julius, Moral Notions,* London: Routledge & Kegan Paul, 1967.

Kuhn, T.S., *The Structure Of Scientific Revolutions.* Chicago: University of Chicago Press, 1962.

Kuipers, B.J., 'Representing knowledge for recognition', in RU, 1975.

Lakatos, lmre, 'Falsification and the Methodology of Scientific Research Programmes', in *Criticism And The Growth Of Knowledge,* lmre Lakatos and Alan Musgrave (Eds.), pp. 91-196. Cambridge University Press, 1970. *Proofs and Refutations.* Cambridge: Cambridge University Press, 1976.

Lenat, D.B. *,AM: an Artificial Intelligence approach to Discovery in Mathematics as Heuristic Search.* Ph.D. thesis, Stanford University A.I. Laboratory, 1976. (Also reported in IJCAI-5,1977.)

Lindsay, R.K., Inferential Memory as the Basis of Machines Which Understand Natural Language', in CT, pp. 217-33. 'Jigsaw Heuristics and a Language Learning Model', in Findler and Meltzer (Eds.) 1971, pp. 173-189.

Longuet-Higgins, H.C., 'The perception of melodies'.*Nature* vol. 26, no. 5579, pp. 646-653, 1976.

Luria, A.R., *The Man with a Shattered World.* Harmondsworth: Penguin Books, 1975. (Also Basic Books, 1972.)

McCarthy, John and P.J. Hayes, 'Some Philosophical Problems from the Standpoint of Artificial Intelligence', in MI-4,1969, pp. 463-502.

Mackworth, A.K., Interpreting Pictures of Polyhedral Scenes', in Al, 4, 1973, pp. 121-138. 'Using Models to See', in AISB-I, pp. 127-37.

Marr, David,

'Analyzing Natural Images.: A Computational Theory Of Texture Vision', Al Memo 334. Cambridge, Mass.: MIT Al Lab., June 1975.

'Early Processing Of Visual Information', in Philosophical Transactions of the Royal Society of London 275 (942), 1976, pp. 483-524.

Meltzer, Bernard, Review of Lenat 1976, in AISB Quarterly, 27, July 1977, pp. 20-3.

Michie, Donald, *On Machine Intelligence.* Edinburgh: Edinburgh University Press, 1974.

Miller, G.A., Eugene Galanter, and K.H. Pribram, *Plans and the Structure of Behavior.* New York: Holt, 1960.

Minsky, M.L.,

'A Framework for Representing Knowledge', in PCV, pp. 21 1 -77. 'Descriptive Languages and Problem Solving', in SIP.

'Steps toward Artificial Intelligence', in CT.

'Matter, Mind and Models', in SIP.

'Form and Content in Computer Science', ACM Turing Lecture, J.A.C.M. Vol. 17, 2 April, 1970, pp. 197-215.

Minsky, M.L., and Seymour Papert, *Artificial Intelligence.* Eugene, Oregon: Condon Lecture Publications, 1973. Also MIT Al Laboratory, memo 252.

Mueller, Ivan, 'Euclid's Elements and the axiomatic method', in *British Journal for the Philosophy of Science,* December 1969.

Nagel, Ernest, and J.R. Newman, *Godel's Proof.* New York: New York University Press, 1958.

Newell, Allen, Jeffrey Barnett, J.W. Forgie, C.C. Green, D.H. Klatt, J.C.R. Licklider, J.H. Munson, D.R. Reddy and W.A. Woods, *Final Report of a Study Group on Speech Understanding Systems.* Amsterdam: North Holland, 1973.

Newell, A. and Ernst, G., 'Some Issues of representation in a General Problem Solver'. PROCAFIPS F.J.C.C., 1967, p. 583.

Newell, A. & Simon H.A., GPS, A Program that Simulates Human Thought, in CT, 1961.

Newell, A. and Simon, H.A., *Human Problem Solving.* Englewood Cliffs, N.J.: Prentice Hall, 1972.

Newell, Alien, 'Artificial Intelligence and the Concept of Mind', in CMTL, pp. 1 -60.

Nicholas, J.M. Ed., *Images, Perception, and Knowledge.* Dordrecht-Holland: Reidel, 1977.

Nilsson, N.J., *Problem Solving Methods In Artificial Intelligence.* New York: McGraw-Hill, 1971.

Norman, D.A. & Rumelhart, D.E., *Explorations In Cognition.* W H Freeman & Co. San Francisco, 1975.

O'Gorman, F. and Clowes, M.B., 'Finding Picture Edges through Collinearity of Picture Points'. IJCAI-3, 1973, pp. 556-563.

Palmer, S.E., 'Visual Perception And World Knowledge: Notes on a Model of Sensory-Cognitive Interaction', in Norman and Rumelhart (Eds.), 1975. 'The Nature of Perceptual Representation: An Examination of the Analog/Propositional Controversy', in TINLP, pp. 165-73.

Papert, Seymour,

Uses of Technology to Enhance Education, MIT, Al Lab Memo No. 198, 1973.

'Teaching Children to be Mathematicians Versus Teaching about Mathematics', Int. J.

Math. Educ. Sci. TechnoL, 3, 1973, pp. 249-62.

Paul, J.L., 'Seeing Puppets Quickly', in AISB-2, 1976.

Popper, K.R., *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge & Kegan Paul, 1963. *Objective Knowledge: An Evolutionary Approach.* Oxford: Clarendon Press, 1972.

Pylyshyn, Z.W., 'What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery', Psychological Bulletin, 80, 1973, pp. 1-24, and in Nicholas (Ed.), 1977. **'Do** We Need Images and Analogues?', in TINLP, 1975, pp. 174-177. 'Computational models and empirical constraints', to appear in *Behavioural and Brain Sciences Journal,* 1978.

Raphael, B., *The Thinking Computer: Mind Inside Matter.* W.H. Freeman & Co. San Francisco, 1976.

Raphael, Bertram, 'SIR: A Computer Program for Semantic Information Retrieval', in SIP, pp. 33-145.

Roberts, L.G., 'Machine Perception of Three-Dimensional Solids', in *Electro-optical Information Processing,* Tippet *etal.* (Eds.), 1965, pp. 159-197.

Robinson, Guy, 'How to Tell your Friends from Machines', MIND, N.S., 81, 1972, pp. 504-518.

Ryle, Gilbert, *The Concept Of Mind.* London: Hutchinson, 1949.

Schank, R.C., & R.P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge structures.* Hilisdale, N.J.: Lawrence Eribaum Ass., 1977.

Schank, R.C. & Colby, K. *Computer Models Of Thought And Language,* Freeman, 1973.

Schank, R.C., Goidman, Neil, Reiger, C.J. & Riesbeck, Chris, 'MARGIE: Memory, Analysis, Response Generation and Inference on English', IJCAI-3, 1973, pp. 255-261.

Schank, R.C.,

'Finding the Conceptual Content and Intention in an Utterance in Natural Language Conversation', in IJCAI-2, pp. 444-54.

'Conceptual Dependency: A Theory of Natural Language Understanding', *Cognitive Psychology,* 3, 1972, pp. 552631.

Identification of Conceptualizations Underlying Natural Language', in CMTL, pp. 187-248.

Seely-Brown, John, and R.R. Burton, 'Multiple Representations of Knowledge for Tutorial Reasoning', in RU, pp. 311-50.

Selfe, Lorna, *NADIA: a case of extraordinary drawing ability in an autistic child.* Academic Press, 1977.

Shirai, Yoshiaki, 'A Context Sensitive Line Finder for Recognition of Polyhedra', in Al, 4, 1973, pp. 95-120. (Also as 'Analyzing Intensity Arrays Using Knowledge About Scenes', in PCV, pp. 93-114.)

Sloman, Aaron,

'Necessary, Apriori, and Analytic', in *Analysis,* October 1965.

'Explaining Logical Necessity', in *Proceedings Aristotelian Society 69,* 1968-9, pp. 133-

47.

'How to derive "Better" from "Is" ', *American Phil. Quarterly,* January 1969.

' "Ought" and "Better" \ *Mind,* July 1970.

'Tarski, Frege, and the Liar Paradox',*Philosophy,* April 1971.

Identity and bodily continuity: newbodies for sick persons',*Analysis,* December 1971.

Interactions between Philosophy and A.I.,' in Proc. IJCAI-2 1971. And in Al, 1971, and in M. Nicholas (Ed.), 1977.

'Physicalism and the Bogey of Determinism', (and replies to criticisms) in Brown (Ed.), 1974.

'On learning about numbers', in AISB-I, 1974.

'Afterthoughts on Analogical Representation', in TINLP, 1975, pp. 178-182.

'What are the Aims of Science?', Radical Philosophy, No. 13, Spring 1976, pp. 7-17.

'A.I. and empirical psychology' (commentary on Pylyshyn 1978) in *Behavioural and Brain Sciences Journal* 1978 (to appear).

Sloman, Aaron, and Steven Hardy, 'Giving a Computer Gestalt Experiences', in AISB-2, 1976, pp. 242-255.

Sloman, Aaron, D. Owen, G. Hinton and F. O'Gorman, 'Representation and control in vision'. AISB Conference, Hamburg, July 1978.

Strawson, P.F., *Individuals An Essay in Descriptive Metaphysics,* London: Methuen, 1959.

Sussman, G.,

The Virtuous nature of Bugs', AISB-I, 1974.

A Computational Model of Skill Acquisition. New York: American Elsevier, 1975.

Sutherland, N.S.,

'Outlines of a Theory of Visual Pattern Recognition in Animals and Man., Proc. Royal Society B, 171, 1968, pp. 297-317

Is the brain a physical system?', in **R.** Borger and F. Cioffi (Eds.). *Explanation in the Behavioural Sciences.* Cambridge University Press, 1970, pp. 97-122.

'Computer Simulation of Brain Function' in Brown (Ed.), 1974.

Intelligent Picture Processing', in *Tutorial Essays in Psychology, Vol. II.* (Ed. N.S. Sutherland). Hilisdale, N.J.: Lawrence Eribaum Ass., in press.

Toulmin, Stephen, *The Philosophy of Science,* London: Hutchinson, 1953.

Waltz, D.L., 'Understanding Line Drawings of Scenes with Shadows', in PCV, 1975, pp. 19-92.

Watson, J.D., *The Double Helix,* Harmondsworth: Penguin Books, 1968.

Weir, Sylvia, 'Action Perception', in AISB-I, 1974, pp. 247--256. "The perception of motion: Michotte revisited', to be published, 1977.

Weir, Sylvia, M.R. Adler, and Marilyn McLennan, 'Final Report on Action Perception Project'. Edinburgh: Edinburgh University AI Dept., November 1975.

Weir, Sylvia, and Ricky Emmanuel, *Using Logo to Catalyse Communication In An Autistic Child.* Research Report 15. Edinburgh: Edinburgh University Dept. AI, January 1976.

Weizenbaum, Joseph,

'ELIZA A Computer Program for the Study of Natural Language Communication Between Man and Machine', in CACM, 9, 1966, pp. 36-45.

Computer Power & Human Reason: From Judgement to Calculation, San Francisco: W.H. Freeman & Co., 1976.

Wertheimer, *Max, Productive Thinking.* London: Tavistock Publications, 1966.

White, A.R.,

Attention. Oxford: Blackwell, 1964.

The Philosophy of Mind. New York: Random House, 1967.

Modal Thinking. Oxford: Blackwell, 1975.

Winograd.T.S.. *Understanding Natural Lanzuaee.* Edinburgh: Edinburgh University Press, 1972. 'The Process of Language Understanding' in *The Limits of Human Nature,* Benthall (Ed.), pp. 208-234.

Winston, P.H. (1970), 'Learning Structural Descriptions from Examples', in POCV, 1975. The M.I.T. Robot', in MI-7, 1971. *Artificial Intelligence,* Addison Wesley, 1977.

Wittgenstein, Ludwig,

Philosophical Investigations. Oxford: Blackwell, 1953.

Remarks on the Foundations of Mathematics, Oxford: Blackwell, 1956.

Woods, W.A., 'What's in a Link: Foundations for Semantic Networks', in RU, 1975, pp. 35-82.

Young, R.M., *Seriation by Children: An Artificial Intelligence Analysis of a Piagetian Task.* Basel: Birkhauser, 1976.

Book contents page

Updated: 14 Jan 2002