Views into the Chinese Room

New Essays on Searle and Artificial Intelligence

EDITED BY

John Preston and Mark Bishop

CLARENDON PRESS · OXFORD

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Singapore Taipei Tokyo Toronto

with an associated company in Berlin

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

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

First published 2002

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging in Publication Data

Views into the Chinese room : new essays on Searle and artificial intelligence / edited by
John Preston and Mark Bishop.
p. cm.
Includes indexes.
1. Artificial intelligence. 2. Machine learning. 3. Searle, John R. I. Preston, John, 1957–
II. Bishop, Mark, 1962–
Q335.5 .V54 2002 006.3'5–dc21 2002066233
ISBN 0–19–825057–6
ISBN 0–19–925277–7 (Pbk.)

1 3 5 7 9 10 8 6 4 2

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd, Guildford & King's Lynn

ACKNOWLEDGEMENTS

The following material has previously been published, and is reprinted here with permission.

Material from pages 283–6 of Ned Block's 'The Computer Model of the Mind', which first appeared in Daniel N. Osherson and Edward E. Smith (eds.), *Thinking: An Invitation to Cognitive Science, Volume 3* (Cambridge, Mass.: MIT Press, 1990), by permission of MIT Press. All rights reserved.

Material from pages 21–30 of Roger Penrose's *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press, 1989), © Oxford University Press 1989, and pages 40–4 of his *Shadows of the Mind: A Search for the Missing Science of Consciousness* (Oxford: Oxford University Press, 1994), © Oxford University Press 1994, both by permission of Oxford University Press.

Material from Roger Penrose's 'On Understanding Understanding', *International Studies in the Philosophy of Science*, 11 (1997), 7–10, by permission of Carfax Publishing Company, Abingdon, England.

of evolutionary psychology, philosophical issues in artificial life and evolutionary robotics, and the nature and plausibility of the idea that genes code for phenotypic traits. His first book, *Reconstructing the Cognitive World: The Next Step*, is forthcoming from MIT Press.

Terry Winograd is Professor of Computer Science at Stanford University, California. He was an early researcher in artificial intelligence, focusing on natural language understanding. The seeming intelligence of his program SHRDLU led to optimism that computers could indeed understand language. He later moved away from artificial intelligence, arguing in a book co-authored with Fernando Flores (*Understanding Computers and Cognition* (Addison-Wesley, 1987)) that mainstream AI research was grounded on inadequate and misleading philosophical assumptions about knowledge. For the last decade he has written and done research on human–computer interaction.

1

Introduction

John Preston

Cognitive Science

In the mid-1970s one of the USA's best-known philanthropic organizations, the Alfred P. Sloan Foundation, invested substantial funds in a programme designed to stimulate progress in a burgeoning cross-disciplinary study of the nature and workings of the mind: 'cognitive science'. Although, with hindsight, it can be traced back to the 1950s, cognitive science came to public recognition (and was dubbed by the psychologist Christopher Longuet-Higgins) only in the early 1970s. It comprises a constellation of disciplines (the core members being psychology, linguistics, artificial intelligence, and neuroscience) which currently attempts to explain cognitive phenomena (thinking, reasoning, intelligence, perception, learning, understanding, belief, knowledge, memory, etc.) on the basis of hypotheses about the kinds of information-processing which support them. Motivated and underpinned by a certain philosophical perspective, the constellation subsequently broadened to include parts of or approaches to related fields like anthropology, archaeology, and sociology.

The University of California at Berkeley was one of the main beneficiaries of the Sloan Foundation's programme, as part of which prominent researchers were funded to travel around the country, lecturing at universities. How one of these researchers, a philosopher from UC Berkeley, came to be thought of as

I am grateful to John Searle, Andrew Hodges, Jack Copeland, and Mark Bishop for comments on draft versions of this introduction. Remaining errors it contains should not, of course, be laid at their door. My work on this material was also supported by a research fellowship from the Leverhulme Trust, to whom I am also grateful.

supplying the best-developed and most pointed threat to a core component of cognitive science is the story we have to tell. Although there is still tremendous controversy over its success, there is some consensus over the import of this 'Chinese Room argument' (CRA), which John Searle first published in a paper entitled 'Minds, Brains, and Programs' (Searle 1980*a*). The argument turns on an easily understood thought-experiment which mobilizes readily available intuitions. If sound, it undermines the official self-image of artificial intelligence (AI), one of the supposed foundations of much contemporary cognitive science. It may well also be contemporary philosophy's best-known argument.

Before we get around to the argument itself, though, we need to have some concepts at our disposal, and some history in place.

Turing Machines

Computer science, and the disciplines it made possible, such as AI, have a foundation in the work of the British mathematician Alan Turing.[1] His 1936 paper 'On Computable Numbers, with an Application to the *Entscheidungsproblem*', a founding document of computer science, is remarkable both for its *theoretical* and its *practical* implications. Before it, there was no real theory of computation, yet it made possible the development (about a decade later) of the first modern stored-program electronic computers, which, in turn, was one of the main preconditions for the rise of cognitive science. The fundamental concepts introduced and explored in the paper can be explained in a relatively simple way.

Turing's paper, published even before he had completed his Ph.D., is about the *decision problem* posed by the German mathematician David Hilbert. The issue concerned *formal systems*, that is, mathematical systems in which the methods of constructing mathematical statements, as well as the assumptions and principles used in proving theorems, are governed by explicit and precise rules. In 1928 Hilbert had queried whether there is an 'effective' (or 'mechanical') method that can determine, of any given statement in a formal system, whether or not it is provable in that system. Turing, setting out to show rigorously that there is not (that mathematics is *undecidable*, in this technical sense), required a precise, convincing, and general definition of an effective or mechanical method. The concept he came up with, now known as the *Turing machine*, is the basic concept of theoretical computer science. The proposal now known as *Turing's thesis* says that whenever there is an effective or mechanical method for calculating the values of a mathematical function, that function can be computed by a Turing machine. The relatively informal concept of an effective or mechanical method can thus be *replaced* by the precise concept of a Turing machine.

[1] Andrew Hodges's biography of Turing (Hodges 1983) is a wonderful source of material on Turing's thought, as well as his life.

It's crucial to what follows that Turing's idea of such an idealized 'machine' was explicitly modelled on that of a *person* performing a computation, and therefore presupposes the adequacy of his own analysis of such activity. When he talks about 'computers' in this 1936 paper, Turing means *humans who compute* (since, of course, there were no 'computers' in the modern colloquial sense at that time). He begins his analysis by pointing out that computation could always be done by writing symbols from a finite alphabet onto a paper tape uniformly divided into square frames. He then says:

> The behaviour of the computer at any moment is determined by the symbols which he is observing, and his 'state of mind' at that moment. We may suppose that there is a bound *B* to the number of symbols or squares which the computer can observe at one moment. If he wishes to observe more, he must use successive observations. We will also suppose that the number of states of mind which need to be taken into account is finite . . . Let us imagine the operations performed by the computer to be split up into 'simple operations' which are so elementary that it is not easy to imagine them further divided. Every such operation consists of some change of the physical system consisting of the computer and his tape. We know the state of the system if we know the sequence of symbols on the tape, which of these are observed by the computer . . . and the state of mind of the computer. (Turing 1936, in Davis 1965: 136)

In the final step of his analysis, Turing proposes that we can avoid these references to the computing person's 'states of mind' altogether by supposing that he or she writes at every single step of the computation a 'note of instructions' which explains exactly how the computation should be continued. These notes of instructions take the place of the computer's 'states of mind', therefore 'the state of progress of the computation at any stage is completely determined by the note of instructions and the symbols on the tape' (ibid. 139–40).

In proposing that 'we may now construct a machine to do the work of this computer' (p. 137) Turing forges the connection between person and machine:

> We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions . . . which will be called '*m*-configurations'. The machine is supplied with a 'tape' (the analogue of paper)

> running through it, and divided into sections (called 'squares') each capable of bearing a 'symbol'. At any moment there is just one square . . . which is 'in the machine'. We may call this square the 'scanned square'. The symbol on the scanned square may be called 'the scanned symbol'. The 'scanned symbol' is the only one of which the machine is, so to speak, 'directly aware' . . . The possible behaviour of the machine at any moment is determined by the *m*-configuration and the scanned symbol. This pair will be called the 'configuration'. (ibid. 117)

Any particular Turing machine can assume any one of a finite number of different internal 'configurations' at a given time. The activity of such machines consists in serially ordered, discrete steps, each step being completely determined by the machine's current *m*-configuration, the contents of its tape, and the square currently being scanned. A tabulation of all the possible configurations of a given Turing machine, together with a specification of the behaviour that follows from each configuration for each input, forms its 'machine table', which we now know as this kind of computer's *program*.

Following this analysis, the simplest kind of Turing machine is usually introduced and pictured as a physical device, consisting of:

> A tape, indefinitely long in both directions, on which symbols are printed, divided into square frames of the same size, each of which can contain at any one time not more than one symbol,

and

> a movable head that (a) prints discrete symbols, drawn from a finite alphabet, onto the tape, (b) erases one symbol at a time from the tape, and (c) reads, or identifies the contents of each square of the tape, one square at a time.

Anything that has this articulation, that is, a 'memory' for discrete symbols, and a way of serially accessing and changing its contents according to a program, counts as a Turing machine. So although Turing machines are often thought of as machines in the colloquial sense (physical devices whose parts operate according to the laws of mechanics), what's essential to them is their *functioning* or *operation*. Anything that operates *as* a Turing machine would *be* a Turing machine. In particular, a person who works like this to produce the results of computations would, by definition, be (operating as) a Turing machine.

Although the basic operations of a Turing machine form an extremely limited set, Turing claims that it includes 'all those which are used in the computation of a number' (Turing 1936–7: 118). So Turing machines are of interest because they illustrate what could be done by machines, in principle, given unlimited time and storage capacity.

One of Turing's most significant achievements, 'Turing's theorem', was a proof that there are Turing machines which can compute whatever any other Turing machine can compute. The devices whose conceivability it establishes, *universal* Turing machines, are of even more theoretical importance than special-purpose Turing machines. Turing realized that a coded description of any given Turing machine could be written onto a section of the tape of another such machine, which would then emulate the behaviour of the first machine, computing exactly what its progenitor would have computed when supplied with the same input. The possibility of universal Turing machines means that the 'engineering' problem, of making separate Turing machines for each kind of computational task, is replaced by what Turing later called 'the office work of "programming" the universal machine to do these jobs' (Turing 1948: 7). On the one hand, no physically existing computer can have the resources of a universal Turing machine (unlimited memory, perfect reliability, etc.). On the other, nobody would *use* such devices for computing, since their simple architectures make their operations far more unwieldy than any practical computing machine. Nevertheless, they are still thought of as prototypes, idealized models, of the kinds of digital computers we are familiar with, since these are computationally equivalent to Turing machines with finite tapes. Turing's idea of putting the program of a special-purpose machine into the memory of another, universal machine, the idea of a *stored-program computer*, was crucial in the development of contemporary machines.

The Church–Turing Thesis

But what exactly *can* Turing machines do? Here, controversy lurks. Turing more than once said that 'logical computing machines' (the term he later used for what we now call Turing machines) can do 'anything that could be described as "rule of thumb" or "purely mechanical"' (ibid.). Working independently and without knowledge of each other, Turing and Alonzo Church, the American logician who later supervised his Ph.D. thesis at Princeton, were both concerned with the concept of an effective or mechanical method, a notion used in mathematics to indicate a class of mathematical *functions* (or results, or problems) that can be computed (attained, solved) in a mechanical way ('by following fixed rules', as it's sometimes put). In late 1933, Church suggested identifying this informal notion with the mathematically precise concept of *lambda-definability*. His subsequent proof of the equivalence between this concept and the mathematical notion of a *recursive*

function led to his first public identification of the effectively calculable functions with the recursive functions (this is now known as 'Church's thesis').

Around the same time, Turing independently presented another formally precise replacement for the concept of an effective or mechanical method: *computability by Turing machine*. He showed that, as long as we accept Turing's thesis, Hilbert's question is settled in the negative, proving that the predicate calculus is undecidable (i.e. that no Turing machine can determine, in a finite number of steps, whether or not an arbitrary formula of the calculus is provable). Turing delayed publication of his 1936 paper in order to show, in an appendix, that these two apparently different replacement concepts are actually equivalent (i.e. they pick out the very same set of mathematical functions). So the resulting *Church–Turing thesis* says that this set of recursive functions, those functions computable by Turing machines, contains every function whose values can be obtained by an effective or mechanical method.

Its denomination ('thesis') betrays the fact that, unlike Turing's *theorem*, the Church–Turing thesis isn't regarded as proven. However, the unforeseeable but proven equivalence of all existing attempts to analyse the notion of an effectively calculable function[2] is often regarded as strong evidence in its favour. Aside from this issue of its *status*, though, the exact *nature* of the thesis is also unclear. Is it a definition (and therefore a kind of convention, as Church himself thought), a proposal, or a conjecture? If the latter, as most commentators nowadays assume, is it a mathematical conjecture or an empirical one? Does the foundation of the concept of computability in human abilities mean that the thesis is intended to cover only phenomena in our physical world?

Even more important than the nature of the thesis, perhaps, is the matter of its implications. It's no exaggeration to say that the Church–Turing thesis has constituted the fundamental inspiration behind AI, the reason for thinking that electronic digital computers *must* be capable of (at least) human-level intelligence. Cognitive scientists have generally taken the Church–Turing thesis to mean that any function that can be computed can be computed by a Turing machine. This would mean that, as long as we ignore or abstract away from resource-limitations, anything the human brain can do (any function it can compute) could also be done (computed) by an electronic digital computer. Cognitive processes, no matter how intelligent, must be decomposable into routines whose primitive steps can all be executed by a machine.

[2] In terms of recursiveness, Turing computability, lambda-definability, Markov algorithm computability, etc.

The Turing Test

The computer age's first serious attempt to give a criterion for mentality and an important goal for AI forms the core of Turing's best-known paper. 'Computing Machinery and Intelligence' (Turing 1950) is surely the most famous, most widely read and reprinted, and the most influential article ever to have been published in a philosophy journal. The criterion Turing argued for involved the simulation of behaviour, specifically, linguistic behaviour. He called it the 'imitation game', but I shall refer to it as it's now known: the *Turing Test*. Because it's so well-known, a sketch will suffice here.

Turing famously rejected the question 'Can machines think?', deeming it 'too meaningless to deserve discussion' (Turing 1950: 442).[3] He therefore proposed to recast it as a question about a game in which a computer programmer has to render it impossible for a human 'interrogator' to tell, from typewritten output alone, whether that output is generated by a human or by a machine. Turing and his defenders then insist that if a machine cannot be distinguished from a human being under these conditions we must credit it with intelligence.

One of the more persistent complaints about the Test is that it is unacceptably *behaviouristic*, or that it proffers an *operational definition* of intelligent thought. Turing himself did subscribe to an account of learning which bears some hallmarks of psychological 'behaviourism'. According to such an approach to psychology, the discipline is suitably scientific only if it confines itself to the study of observable features of bodily motion, described in a restricted and colourless vocabulary. The heyday of psychological behaviourism preceded that of cognitive science, but the two bear an uneasy relationship to one another. Cognitive scientists think of themselves as having gone decisively beyond behaviourism by virtue of relinquishing its underlying conceptions of the mind, science, and psychological explanation. It would be ironic if their best test for mentality was in thrall to the view they thought of themselves as overthrowing.

In philosophy, behaviourism was the view that the meaning of statements about an organism's mental or psychological phenomena can be given wholly in terms of testable statements about its observable physical features and motions. This 'logical behaviourism' had important defenders from the 1930s to the 1950s, although it was challenged by, evolved towards, and was eventually

[3] Reprinted in Boden (1990: 49). From here referred to as B49. See, however, later remarks Turing made, recorded in Copeland (2000).

superseded in popularity by the materialist identity thesis, according to which minds are brains, and mental phenomena type-identical with neural phenomena. This view in its turn, however, had the limelight snatched from it by the perspective which still dominates the philosophy of mind: *functionalism*.

As for operational definitions, 'operationalism' is a philosophy of science according to which theoretical terms should be characterized in terms of the operations needed to verify that they apply. It's linked with behaviourist psychology because any psychological theory all of whose terms are defined in this way would perforce be acceptable to a behaviourist, since its terms would be characterized in terms of properties of, or relations between, stretches of observable phenomena. But the widespread view that Turing put forward, in the imitation game, an operational definition of thought or intelligence finds no support in what he actually said, and cannot be right, since such definitions state logically necessary and sufficient conditions of what they define, whereas Turing explicitly offered the Test as a sufficient condition only.

Turing Machine Functionalism

In the early 1960s, to skip ahead a little, the American philosopher Hilary Putnam, having already made contributions to the foundations of computer science in several papers on a form of Hilbert's decision problem and computational proof procedures for quantification theory, used the concept of a Turing machine (suitably liberalized) in order to state what came to be known as the *functionalist* view of the mind.

Functionalism is driven by the feeling that mentality is a matter of *functioning* rather than *substance*. Whether a creature is made out of carbon compounds, or (more generally) biological stuff, or (more generally still) even out of physical stuff (if there's any conceivable alternative) is of consequence only in so far as these substrates *constrain* their operation. What matters, as far as mentality goes, is not matter, what a thing is made of, but functioning, how the thing *works*, and what its *capacities* are. But functionalism, in this most general sense, is nowadays almost always accompanied by a naturalistic metaphysical thesis according to which mental phenomena are individuated in terms of their causal roles. From this derives the doctrine of *multiple realizability*, according to which mental phenomena can be credited to anything having states with the appropriate causal roles.

Since functionalism suggests that mental concepts are functional concepts, cognitive scientists explicitly or implicitly committed to some version of

functionalism often suppose that *simulating* mental phenomena amounts to *duplicating* them. Specifically, if what's essential to mental phenomena is their causal roles, and if those roles can be simulated in computer programs, there's no reason why computers running those programs shouldn't be credited with the mental phenomena in question.

Putnam originally proposed an analogy between Turing machines and humans. Just as there are two possible descriptions of the behaviour of a given Turing machine: the *engineer's* structural description of its hardware, and the *logician's* or *computer scientist's* 'machine table' (what we would now call its *program*), so there are two possible descriptions of human psychology, the *physiological* description (corresponding to the engineer's structural description), and

> a more abstract description of human mental processes, in terms of 'mental states' . . . and 'impressions' . . . —a description which would specify the laws controlling the order in which the states succeeded one another, and the relation to verbalization. This description, which would be the analogue of a 'machine table', it was in fact the program of classical psychology to provide! (Putnam 1960: 373)

But Putnam soon came to think that the correspondence between human and Turing machine was more than a mere analogy. By 1967 he was affirming that humans *are* what he called 'probabilistic automata'. (This differs from a Turing machine mainly in that the transitions between the configurations of a probabilistic automaton are allowed to differ in the probability that they will take place, rather than being deterministic.) Turing machine functionalism is then (roughly) the hypothesis that all systems capable of having any given psychological state are probabilistic automata.

Functionalists soon left behind this early form of their view, having been persuaded that humans, unlike probabilistic automata, can be in more than one (psychological) state, and can perform more than one (mental) operation, at a time. Crucial to this transition was the realization that a human mind couldn't be any kind of Turing machine because the *serial* processing that a single Turing machine is capable of cannot possibly reproduce, within the constraints provided by the human brain, the sorts of psychological abilities people display. Even fans of serial processing now insist that architectures comprising massively parallel *collections* of serial processors are needed. But a broader kind of functionalism according to which psychological states are individuated not just by their machine table states but by their causal connections with sensory inputs, with one another, and with behavioural outputs, has come to dominate the philosophy of mind and now lies at the heart of cognitive science.

Computationalism

Within contemporary cognitive science this perspective is closely associated with *computationalism*, whose most prominent and important philosophical champion is Jerry Fodor. He explicitly traces it back to Turing (although he sometimes warns readers that his history is idealized, not scholarly), and it's instructive to see how.

According to Fodor, cognitive science in general (and AI in particular) is primarily about *intelligent thought*. Central to its agenda are questions like 'How can people go from one true thought to another?', 'How can thought processes be *coherent*, or *intelligent*, or *rational*?'. Turing answered a related question, namely, 'Given that a state of an organism (or a system) has *semantic* features (e.g. truth), how could its state transitions preserve or respect those features?'. By doing so, he bequeathed to cognitive science not just *a* model but the *only* sensible model we have been able to come up with, of what intelligent (i.e. rational) thought processes could be. His model says that thought processes involve (perhaps even consist in) the computational transformation of mental representations. Mental representations are symbols, having both semantic and *syntactic* features. Thinking *is* essentially a matter of manipulating mental symbols. And *intelligent* or *rational* thought is a matter of *preserving truth* in inferences, moving from symbols (premises) to symbols (conclusions) in such a way that if one's premises are true, one's conclusions will also be true.

Turing, in providing a syntactical theory of computation, thereby provided a syntactical theory of intelligence. What he showed, according to Fodor, was how to construct a device (namely, a Turing machine) that can process symbols, purely in virtue of respecting their syntactic features, in a way that ensures that none of their semantic features will be violated. The state-transitions or processes of such a device can respect or preserve the 'content' of its states. For example, it can recognize (to put things loosely) that an inference from 'P & Q' to 'P' is valid (truth-preserving), regardless of what statements the component non-logical symbols stand for. And it does this by exploiting *parallels* between syntactic and semantic features. This is supposed to be the basic idea of the branch of logic known as *proof theory*, which deals with formal systems:

> The basic question of cognitive science is, How could a mechanism be rational? The serious answer to that question is owing to Turing, namely: it could be rational by being a sort of proof-theoretic device, that is, by being a mechanism that has representational capacities—mental states that represent states of the world—and that can operate on these mental states by virtue of its syntactical properties. The basic idea in cognitive science is the idea of proof theory, that is, that you can simulate semantic relations—in particular, semantic relations among thoughts—by syntactical processes. That is what Turing suggested, and that is what we have all been doing in one or the other area of mental processing. (Fodor 1995: 88)

For this kind of cognitive science, then, human cognition is a brain process in which symbols (which have both syntactic and semantic features) are manipulated in virtue of their syntactic features (only). The semantic features ('content') of the symbols, although irrelevant to their processing, are nevertheless preserved by it.

Artificial Intelligence

Artificial intelligence, claims about which are the focus of the arguments discussed in this volume, began in earnest in the mid-1950s. An account of its history isn't practicable here,[4] so a quick sketch of its origins and development must suffice for our purposes.

Margaret Boden has plausibly traced the inception of AI to a 1943 paper by Warren McCulloch and Walter Pitts (McCulloch and Pitts 1943). Having proposed a 'correspondence' between the physiological relations among neurons and the logical relations among propositions, they showed that certain kinds of networks of artificial neurons can compute whatever functions a Turing machine can compute, and took this to be what they called a 'psychological justification' of the Church–Turing thesis. Since then, AI has mainly comprised two rather different (albeit intertwined) research programmes.

The founders of what is now thought of as the 'classical', 'symbolic' approach to AI, including John McCarthy, Herbert Simon and Allen Newell, and Marvin Minsky, were more familiar with, and influenced by, McCulloch's work than Turing's. Early research of this kind concentrated on a small number of problem-domains, and its working focus changed quite rapidly. Programs designed to prove theorems in areas of mathematics such as logic, geometry, and algebra rubbed shoulders with games-playing programs, devoted to draughts (checkers), chess, or card games such as bridge, and 'problem-solvers', programs which would address intellectual puzzles such as the 'towers of Hanoi', the 'bridges of

[4] For such accounts, see McCorduck (1979), Gardner (1985), Pratt (1987), and Boden (forthcoming).

Konigsberg', the 'travelling salesman', 'missionary and cannibals' cases, etc. Early AI research focused on such problems because, as the editors of the first major anthology of such research papers put it,

> game situations provide problem environments which are relatively highly regular and well-defined, but which afford sufficient complexity in solution generation so that intelligence and symbolic reasoning skills play a crucial role. In short, game environments are very useful task environments for studying the nature and structure of complex problem-solving processes. (Feigenbaum and Feldman 1963: 37)

Playing games, proving theorems, and solving problems are clearly examples of intelligent human activities. They also have the advantage of being tractable and manageably small domains.

Since the mid-1950s, however, AI programs have come to range over a field vastly greater than this. Problem-solving, itself now covering a far greater area, takes its place alongside vision, natural-language understanding, planning, 'machine learning', expert systems, and several other areas at the centre of this kind of AI.[5] But although these 'classical', 'symbolic' AI programs, designed to perform well on such activities, are of interestingly different kinds, they all deploy *representations* according to *rules*. The representations can be mathematical or language-like. The rules can be either *algorithms* (procedures guaranteed to give the right answer to the question posed) or *heuristics* (rule-of-thumb procedures which can narrow down the search-space). What such programs importantly share (for our purpose) is that they manipulate data-structures composed of *symbols* according to *instructions*.

Alongside these first steps in classical, symbolic AI ran another research programme, now known as *connectionism*.[6] Connectionist research investigates the properties of *neural networks*, which consist of large numbers of artificial neurons, small and very simple processing units, related to one another by connections whose excitatory or inhibitory 'weight', and whose threshold for firing, can be altered by the programmer. (The networks studied by McCulloch and Pitts are an early example.) In some ('local') connectionist networks each unit or 'node' is associated with a distinct symbol, but in the more interesting recent ones, symbolic information is thought of as being encoded in a distributed way across collections of nodes. The ideas that form the background to this research were developed during the 1940s and 1950s by psychologists such as

[5] For an introduction, see Boden (1987).

[6] For introductions to the technical issues, see Wasserman (1989), Bishop (1997), and Haykin (1999). For the philosophical issues, see Bechtel and Abrahamsen (1999).

Donald Hebb, Karl Lashley, and Oliver Selfridge in a mathematical branch of their subject called *learning theory*.

The first computer simulation of a neural network was undertaken at the Massachusetts Institute of Technology in 1954. But the first important working networks were developed in the 1960s by Frank Rosenblatt, who showed how McCulloch–Pitts nets could be 'trained' to classify certain patterns as similar or distinct. The sorts of tasks networks excel at, such as pattern-recognition, pattern-completion, etc., contrast with those to which classical AI programs are usually put. Neural networks have been widely advertised as having some important similarities (of structure and operation) with brain components. But this isn't to say that the former are accurate models of the latter.

Boden sums up the influence of McCulloch and Pitts's 1943 paper thus:

> Their vision of implementing the 'logical calculus' influenced von Neumann in designing the digital computer, and inspired AI pioneers to attempt the formal modelling of thought. And their discussion of 'nervous activity' contributed to Hebb's psychophysiological theory of cell-assemblies, and engendered various models of neural networks. . . . If 'nets' are thought of as approximations to real neural connectivities, then we have a broadly connectionist research-programme. Interpreted as highly abstract idealizations of neural activity, the prime focus being on binary logic rather than real cell-connectivities and thresholds, we have the digital information-processing typical of traditional AI. Both types of AI research were initiated as a result of McCulloch and Pitts's paper. (Boden 1990: 2–3)

Turing, as befits his heroic status, anticipated and laid foundations for *both* these research programmes. But his work on neural networks, although undertaken in 1947, was not published until more than twenty years later, and his role in the history of this kind of computation has only recently begun to be emphasized.[7] The advice he issued at the very end of his 1950 paper, that one ought to try both approaches ('abstract activities' such as chess, as well as providing a machine with sense-organs and educating it, like a child), although prescient, appears not to have been importantly influential on the founders of AI.

Conceptions and Claims of AI

There have always been at least two major conceptions of AI. The first, and more radical, saw it as the attempt to design and build machines which display a range

[7] See e.g. Leiber (1991), Proudfoot and Copeland (1994), Copeland and Proudfoot (1996).

of genuine psychological attributes: problem-solving, thinking, understanding, and reasoning, and perhaps ultimately even consciousness, feeling, and emotion. This seems to have been the official self-image of the AI project. According to the second, more modest conception, the aim was to enable machines to do things which, when humans do them, are counted as examples of these same psychological phenomena, but *without* the implication that the machines should be credited with the psychological attributes in question. At the centre of the mainstream cognitive science Searle discerns a theory of mind based on the first view of AI, according to which minds are really (certain kinds of) programs.

These two different conceptions are related to (but don't map exactly onto) one of Searle's best-known contributions to the debate, the distinction between 'Strong' and 'Weak' AI. These are not primarily (as some suppose) approaches to AI, nor ideas about how AI engineers should spend their time, but rather conceptions of the activity, the field, and its aims. Weak AI says simply that electronic digital computers are powerful instruments for helping us to model, and thereby understand, the mind. This is contested neither by the Chinese Room Argument, nor by Searle's later work.[8] In fact, he expresses enthusiasm for it (e.g. Searle 1982*b*: 57).

Searle's arguments are *not* aimed at rubbishing AI research, or those who carry it out. They are directed instead at a two-part claim *about* such research, the *Strong AI* thesis, which Searle originally formulated as saying that

(*a*) an appropriately programmed computer really would *have* (or be) a mind in the same sense that you or I have,

and

(*b*) its following the program(s) in question would explain its ability to do the psychological things it does.[9]

Strong AI, he also said, is 'a precise, well-defined thesis: mental processes are computational processes over formally defined elements' (Searle 1980*a*: 422 (B81)).[10] It has been, in one form or another, an important part of the 'prose' of AI engineers, the accounts they generate in order to explain and justify what they do, both to those who provide its research funding, and to the wider

[8] However, as some commentators have noted, it's hard to see why 'Weak AI' deserves its name at all, since it ignores the possibility of genuine artificial intelligence (in electronic digital computers, at least). 'Weak AI' is better thought of as *cognitive simulation*.

[9] Later, Searle formulates (*a*) as 'all there is to having a mind is having an appropriately-programmed digital computer' (1987*b*: 295).

[10] In articles for the popular press (e.g. 1982*a*: 3, 1982*b*: 56), Searle sometimes characterizes Strong AI as a baggier group of theses.

public, via the media. (The extent to which this is true can partly be gauged from laypeople's reactions to AI projects.)

Strong AI, however, is something more than a thesis. It forms a sort of picture whose implications, if Searle is right, ramify within many different aspects of cognitive science. The Strong AI *thesis*, one might say, is by no means the only important matter to which Searle's arguments are supposed to pertain. Although he has gone to some lengths to clarify exactly what he takes the negative import of the CRA to be, he states it in rather different ways in different writings, and there is still unclarity about its scope. The first sections of Larry Hauser's chapter in this volume address this issue, taking Searle to task for the unclarity.

Examples of Strong AI

Ever since Searle's original paper, some have supposed that 'Strong AI' is a straw man, a position hardly anyone in the cognitive science community subscribes to, or has ever subscribed to. This is certainly not so. That it has indeed been held can be seen in careful theoretical statements, as well as in certain exuberant and unqualified predictions, made by the precursors and originators of AI. As far as Turing goes, his presentation of the imitation game implies that computers whose performance comes to be indistinguishable from that of humans (in this respect) should be counted as thinking things. Searle himself shows that the founders of AI all subscribe to Strong AI (Searle 1984: 29–30, 1987*a*: 210–11). Perhaps the centrepiece of Newell and Simon's theoretical work in AI is their 'Physical Symbol System Hypothesis', that 'the necessary and sufficient condition for a physical system to exhibit general intelligent action is that it be a physical symbol system' (Newell 1980: 170), that is, a machine that produces a changing collection of symbol-structures. In early 1956, Simon claimed that the two of them had just invented a thinking machine, and by 1957, they expressed their view that

> [T]here are now in the world machines that think, that learn, and that create. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the range to which the human mind has been applied. (Simon and Newell 1958: 8)

Simon, in his chapter with Stuart Eisenstadt for this volume, resolutely maintains that genuine computer comprehension of natural language is not just empirically possible, but has been with us since 1972. Their chapter bears witness

that some cognitive scientists still prefer their terms to have operational definitions, and think that the right such definitions are what's needed to definitively resolve the issue of which things can really have psychological properties.

Only a little contact with university departments of the core cognitive sciences is necessary to testify to the ongoing prominence of Strong AI (among both staff and students). Daniel Dennett, one of the most important contemporary philosophers of mind and cognitive science, explicitly identifies himself as one of its defenders. But *anyone* who *seriously* describes electronic digital computers in psychological terms, or who proposes to replicate psychological phenomena in such computers solely by programming them in appropriate ways, is committed to Strong AI. Plenty of cognitive science literature is so committed. Some important contemporary cognitive scientists subscribe to it even in their more scholarly and less marketable productions, although it is most manifest in recent popular books by authors such as Hans Moravec and Ray Kurzweil. If you think you don't know of anyone who endorses Strong AI, you haven't looked hard enough!

Searle, on the other hand, claims to have found an argument that undercuts the idea that electronic digital computers (whether they run current AI programs, or *any* programs) can be said to exhibit any of the contested psychological capacities purely in virtue of their programs. Philosophers certainly have no special insight into what technical tasks programmed machines might be able to perform, or when. But they can have a say about how it makes sense to characterize the abilities in question, especially when that characterization is in non-technical terms, such as those we all use for mental phenomena. Like anyone else they may, by using thought-experiments for example, establish or refute theses about what is logically possible.

The Yale Programs

The AI lab at Yale University, having secured funding for visiting speakers from the Sloan Foundation, invited Searle to speak to them about cognitive science. Not knowing much about AI at the time, Searle bought a book recently written by the head of the lab, Roger Schank, and his colleague from the Psychology Department, Robert Abelson, in which they described what they called their 'story-understanding' programs.[11]

[11] From an interview with Searle conducted on 24 April 1999.

Schank and Abelson postulated certain theoretical entities which, they said, 'must form the basis of human memory organization' (Schank and Abelson 1977: 17). Human memory, they argued, is organized around *episodes*, personal experiences, and therefore must include a procedure for recognizing repeated or similar sequences. As an economy measure for storing episodes, they therefore proposed that 'when enough of them are alike they are remembered in terms of a standardized generalized episode which we will call a *script*' (ibid. 19, emphasis added):

> A script is a structure that describes appropriate sequences of events in a particular context. A script is made up of slots and requirements about what can fill those slots. The structure is an interconnected whole, and what is in one slot affects what can be in another. Scripts handle stylized everyday situations. They are not subject to much change, nor do they provide the apparatus for handling totally novel situations. Thus, a script is a predetermined, stereotyped sequence of actions that defines a well-known situation. (Schank and Abelson 1977: 41)

The part of their work Searle focused on aimed to simulate and explain the human ability to understand stories. One of the notable facts about this is that people can answer questions about stories even when the correct answers to those questions aren't explicitly contained within the story itself. The Yale programs aimed to reproduce this ability (among others) using scripts which represent the sorts of information humans have about typical scenarios, to answer questions about the story they are given by referring to the appropriate representations.

Searle's Intervention: The Chinese Room

Before he got to Yale, Searle came up with a thought-experiment which he believed would show that no matter how good they were, programs like these could only ever simulate, but never duplicate, the psychological abilities in question.

Searle claims that nothing in his argument depends upon the details of the Yale programs, that it applies to *any* computer simulation of human mental phenomena. Although AI research isn't focused only on the simulation of *human* abilities (it isn't just 'Weak AI'—cognitive simulation), his idea is that the Yale programs are just the *kind* of things with which AI engineers hoped to construct machine intelligences. How fair is this?

Some commentators have complained that, even during the 1970s, many AI researchers were already unhappy with the assumption that language-understanding is self-standing, and had started paying more attention to the ways in which language is integrated with perception and action. Searle, as we shall see, has his own response to critics who insist that AI programs deal with the causal commerce between such domains. For the moment, he assumes that there is nothing untypical about the Yale programs. They may be relatively unsophisticated by today's standards, but they are not bad examples of AI programs. Much of the rest of AI is 'more of the same', and the Strong AI claim would be that machines running such programs (when perfected by reference to cognitive psychology, of course), really would *understand* the stories they manipulate, in the same sense in which you and I do.

Searle believed his thought-experiment shows such claims to be utterly false. Here it is, in his own words:

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch a 'script', they call the second batch a 'story', and they call the third batch 'questions'. Furthermore, they call the symbols I give them back in response to the third batch 'answers to the questions', and the set of rules in English that they gave me, they call the 'program'. Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view—from the point of view of someone reading my 'answers'—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program. (Searle 1980*a*: 417–18 (B68–9)).

The central claims of Searle's original paper are clear. Something is a digital computer in virtue of performing computations. But computations alone cannot, in principle, give rise to genuine cognition. Computation being nothing but the manipulation of symbols in accordance with purely *formal* or syntactical rules, something that is only computing cannot be said to have access to or know or understand the '*content*', the *semantic* properties (meaning, interpretation) of the symbols it happens to be manipulating. For the computer, as it were, what it manipulates are 'just formal counters' (Searle 1982*a*: 4), not symbols. But it's a conceptual or logical truth that syntax is not sufficient for semantics. Computers therefore cannot be credited with *understanding* the rules they apparently follow, or the programs those rules compose, or the symbols they manipulate. In fact, their states entirely lack what philosophers call *intentionality*, 'the feature of certain mental states by which they are directed at or *about* objects and states of affairs in the world' (Searle 1980*a*: 424 n. 3 (B72), emphasis added). Computation, therefore, can neither be nor explain cognition. While something can *simulate* intelligent performances purely in virtue of performing computations, it cannot thus *duplicate* them.

What the Argument is Supposed to Target

At the very least, the Chinese Room immediately implies that one's own conscious awareness of one's linguistic ability is not captured merely by running the right programs, since the Room runs those, but neither it nor its inhabitant is *conscious of* being able to understand Chinese. But, arguably, you can't understand a language without knowing that you do so (a related joke: 'Can you speak German?' 'I don't know, I've never tried'). Some think this

would be enough to show that AI cannot duplicate *all* important psychological phenomena. Others, like Hauser and Jack Copeland, detect here a pernicious Cartesian assumption about the incorrigibility of introspection, extended in an even more problematic way beyond the normal situations in which it is most plausible.

Regardless of this, Searle's aim is to crack a much bigger nut. The CRA focuses on language-understanding partly because to show that computers cannot *understand* would be to attack Strong AI 'on what its proponents take to be their strongest ground' (Searle 1980*b*: 453). Of all psychological concepts, the understanding of symbols (the symbols they themselves manipulate) is the one *most* plausibly attributed to electronic digital computers, the one partisans of Strong AI think most obviously can be ascribed to such devices.

Among AI engineers, especially, one quite often encounters the idea that the Chinese Room scenario is flawed because the program in question has not *learnt* to understand Chinese, and that this deficiency could be remedied by providing it with, or by substituting for it, a machine learning program (of the sort that Turing envisaged, or some more recent variant). This move will not work. Machine learning cannot rescue Strong AI from the CRA, since one could run an argument entirely parallel to the CRA focusing on learning, rather than understanding. Its conclusion, of course, would be that however good they may be at 'passing' Turing Tests relevant to learning, the machines in question don't *learn* anything. Any programs that can be run on Turing machines, 'any programs at any level at all' (Searle 1989*b*: 702), however embodied, and however they got there, are targeted by the Chinese Room Argument.

What is clear is that Searle never intended the argument to show anything as general as that machines cannot think (or understand, or have any genuine psychological capacities), that computers cannot think (etc.), or that there cannot be artificial thinking machines (artificial intelligence *proper*?). He is adamant that some machines *can* think (since people are machines, for example), that some computers can think (since we are computers), and allows that it may be possible to produce an artificial (man-made) thinking machine (Searle 1980*a*: 422 (B82)). Even if we understand the terms 'machine' and 'computer' in their contemporary and colloquial uses, in which people are contrasted with both machines and computers, Searle's argument doesn't seek to rule out the logical possibility that paradigm examples of computing machines (e.g. the PC on which I'm writing) might have psychological properties. The real focus of the CRA is *programs* (rather than whatever runs them), and it's best thought of as denying that anything could have any genuine psychological properties *solely in virtue of its running a program*. My PC could, as far as this argument is concerned, understand Chinese. But it couldn't do so in virtue of any program which AI engineers could make it run. Artificial Intelligence may be possible, but it cannot result solely from programming, however sophisticated.

This, however, already provides a large enough and important target. AI has changed since the 1950s, coming to conceive itself less as part of science and more as engineering. An increasing proportion of AI research is neither explicitly nor implicitly aimed at the goal which poorly informed philosophers sometimes suppose it *must* have, of producing an 'artificial mind' or brain. Indeed, concepts like that play a decreasingly important role in the prose and practice of AI. Nevertheless, plenty of AI research is still supposed to be about producing (rather than merely *simulating*) psychological phenomena on electronic digital computers just by programming them (even though AI engineers don't generally care whether their systems perform their tasks *in the same way* that intelligent creatures do). AI, even now, isn't *just* cognitive simulation.

The view which Hauser, Stevan Harnad, and Georges Rey here call *computationalism* or *the computer model of the mind*, which says that computation is both necessary and sufficient for cognition, because mental states are (solely) computational states, also falls within the target area (as does the related view Ned Block here identifies by the same term, that the brain is a digital computer). This forms the core of a wide research programme within cognitive science, the contemporary way of *doing* cognitive science, that Searle sometimes calls '*cognitivism*', according to which:

> Thinking is processing information, but information processing is just symbol manipulation. Computers do symbol manipulation. So the best way to study thinking (or as [cognitivists] prefer to call it, 'cognition') is to study computational symbol-manipulating programs, whether they are in computers or in brains. (Searle 1984: 43)[12]

Hauser and Rey are particularly keen to show that Searle's arguments are ineffective against computationalism; Harnad agrees with Searle that his arguments do refute some forms of that view.

Versions of *functionalism* and the *representational theory of mind* according to which the mind is to the brain as computer program is to computer hardware also fall within the argument's sights. Searle sometimes identifies this analogy with, and sometimes treats it as a summary of, the thesis of Strong AI (Searle 1984: 28, 1987*a*: 210, 1987*b*: 295, 1990*a*: 20). Elsewhere he identifies it as 'the basic idea of the

[12] Unfortunately, at other times Searle calls the view that the brain is a digital computer 'cognitivism' (1990*a*: 122, 1992: 202).

computer model of the mind', implicit in Turing's 1950 paper, as well as announced and defended in many important textbooks and articles since (Searle 1990*b*: 21, 1992: 200). He thinks of Turing machine functionalism (at least) as a variant on a view which Putnam himself had already done his bit to discredit: behaviourism. (Behaviourism, of course, as if it needed any further refutation, is also within the sights of the CRA). The CRA is supposed to refute functionalism because the Room, if the programmers have got things right, not only *behaves* as if it understood Chinese, it *functions* (externally *and* internally) as if it does. And yet it doesn't understand. Rey stringently denies this, arguing that functionalism is in no way committed to the Turing Test, that it has resources far deeper than the behaviourism with which Searle misguidedly associates it, and that once one takes account of these, Searle himself might even count as a functionalist. He does, after all, accept (a non-computationalist version of) the representational theory of mind.

Against the Turing Test

One of the points at issue in the debate is the adequacy of the Turing Test, which Strong AI is widely supposed to use as its criterion of the mental.[13] Searle expresses what he takes to be the Strong AI view thus:

> The conclusive proof of the presence of mental states and capacities is the ability of a system to pass the Turing test . . . If a system can convince a competent expert that it has mental states then it really has those mental states. If, for example, a machine could 'converse' with a native Chinese speaker in such a way as to convince the speaker that it understood Chinese then it would literally understand Chinese. (Searle 1982*a*: 3)

In this volume Roger Penrose, otherwise a fierce critic of Strong AI, argues in favour of accepting some form of the Turing Test on the grounds that only if we assume that the presence of mental phenomena is publicly detectable do we meet the appropriate standards of scientific objectivity. Whether as a core commitment of Strong AI, or just an optional extra, plenty of AI researchers, other cognitive scientists, and some of their philosophical cheerleaders do still seem to take something like the Turing Test as a sufficient condition of mentality.[14] By this I mean that whether or not they would explicitly agree that psychological phenomena are to be credited to anything which apparently

[13] Searle (1982*a*,*b*) explicitly treats the Turing Test as a component of Strong AI.

[14] For its most vigorous philosophical defence, see Dennett (1985) and Leiber (1991).

displays the appropriate linguistic skills, their practice betrays a readiness to credit them with such phenomena under just such conditions.

Searle thinks that this is totally wrong-headed. His argument can be presented as an attack on, and as embodying an alternative to, Turing's proposed sufficient condition for mentality.

He claims, first, that if the Turing Test is supposed to be not just a way of pragmatically avoiding philosophical discussion but a theoretically significant criterion of mental phenomena (such as intelligent thought), then the CRA refutes it (Searle 1982*a*: 5, 1987*b*: 295, 297, 1989*a*: 45, 1995*b*: 208). The case he makes reinforces the common complaint that the Test is behaviouristic (in the philosophical sense), casting doubt on its adequacy by insisting that 'there could be two "systems", both of which pass the Turing-test [for understanding], but only one of which understands' (Searle 1980*a*: 419 (B74)). His general diagnosis of the Test's failure is that it confuses epistemology with ontology: 'it makes a fundamental confusion between the way we would verify the presence of a mental phenomenon from the third person point of view [and] the actual first person existence of the phenomenon' (Searle 1989*a*: 45. See also Searle 1993*b*).

In this volume, Stevan Harnad agrees that the Room refutes the original Test under computationalist assumptions, but then goes on to investigate how much room for manoeuvre computationalists have. They might, he proposes, strengthen the Turing Test into something that would survive the CRA by requiring functional, or perhaps structural *and* functional, indistinguishability (instead of the mere linguistic indistinguishability of the original Test). But to require structural conditions would be to give up their doctrine that computational states are implementation-independent. Although Harnad thinks the CRA over-reaches itself in several ways, he also suggests that Searle's work has had a positive influence on cognitive science, helping to open up kinds of research which (unlike classical symbolic AI or connectionism) take the *embodiment* of minds seriously.

Terry Winograd, however, takes issue with Searle's entire conception of language, particularly of understanding. He complains that Searle is not entitled to claim that there are 'clear cases in which "understanding" literally applies and clear cases in which it does not apply' (Searle 1980*a*: 419 (B71)), and tries to show that the questions Searle raises simply don't permit of objective (right or wrong) answers. Winograd's central contention that Searle's question, 'Does the computer understand Chinese?' is meaningless when and because it has been removed from an appropriate context bears comparison with Turing's claim that the question 'Can a machine think?' is meaningless.

Searle's accusation that Strong AI and computationalism conflate ontology and epistemology also has implications for other views of the relation between the mind and behaviour. The Chinese Room scenario, after all, takes advantage of the fact that mere symbol-manipulation can in theory be used to simulate an enormous range of (perhaps even *all*) behavioural phenomena. It thereby suggests that our conception of the mental can be *entirely* divorced from our ways of telling what mental phenomena another being is undergoing, in Searle's words, that 'where the ontology—as opposed to the epistemology—of the mind is concerned behavior is, roughly speaking, irrelevant' (Searle 1993*b*). In this respect it parlays a traditional objection to behaviourism into an objection to anyone who thinks there is more than a purely contingent relationship between mental phenomena and behaviour.

Jeff Coulter and Wes Sharrock, by contrast, although agreeing with Searle over the bankruptcy of the Turing Test, diagnose the problem differently. For them the CRA shows, contrary to the Test, that the nature of a performance cannot be identified independently of the conditions of its production.

Second, Searle presents his argument not as a way of testing for any particular mental phenomenon, but as a way of testing theories of mind (1980*a*: 417 (B68)). Thought of thus, the experiment embodies the following general criterion of adequacy: given any theory of mind, ask yourself what it would be like if you yourself instantiated that theory, if your mind worked in the way the theory suggests. If the resulting mental world departs markedly from what you would experience, the theory in question has to be false.

It's difficult to see how such a criterion could be challenged at least when applied to *conscious* mental phenomena. If a theory fails to capture how things seem to a conscious subject, it has failed to capture the 'contents' of consciousness. In connection with this, Searle issues a general injunction always to 'insist on the first person point of view' (Searle 1980*b*: 451, 1982*c*: 346). Whether this is still thought of as problematic within cognitive science and its philosophy purely because of the residual behaviourism and operationalism Searle discerns there, or whether this injunction is itself methodologically inappropriate, or a philosophical liability, are some of the larger issues involved.

Initial Reservations

Certain misunderstandings of the Chinese Room scenario should be squashed from the outset.

That Searle cites only caricatures of rules of the very simplest kind ('"squiggle squiggle" is followed by "squoggle squoggle"' (1980*a*: 419 (B73)), 'Reach into basket 1 and take out a squiggle-squiggle sign, and go put that over next to the squoggle-squoggle sign that you take from basket 2' (Searle 1987*a*: 213)), for example, is no objection. Even the most complex classical AI programs are concatenations of instructions of roughly this *kind*: they specify sequences of symbols or relationships between symbols which hold purely in virtue of their 'formal' or syntactic properties alone.

Some have complained that ordinary-language mental concepts such as understanding, in which the CRA is framed, are somehow not suitable to pose important questions to AI. But just as cognitive simulation aims to *simulate* everyday psychological phenomena such as story-understanding, so Strong AI conceives AI programs as means by which to *duplicate* (and thereby explain) such phenomena. They comprise the level of psychological abilities with which cognitive science ultimately has to make contact, if it's to explain what it claims to. Unless it can explain things *like* this, cognitive science hasn't got a subject-matter.

It has often been objected that Searle's scenario is unrealistic in other, important ways, for example, in its supposition that a human could possibly handwork the suite of programs that would undoubtedly be needed accurately to reproduce the performance of a native Chinese speaker in real time. The right response to this objection, I believe, is not that speed is irrelevant to how we ascribe psychological abilities such as intelligence. Rather, it is that the fact that the person in the room couldn't handwork the programs fast or reliably enough doesn't matter. Neither would it matter to Searle's response to the 'Systems Reply' (see below) if memorizing the programs in question turned out to be beyond human capacities. The Chinese Room is a thought-experiment, an investigation into what would follow if something thoroughly counterfactual were to be the case. But in this respect it doesn't differ from Einstein's request for us to imagine what we would observe if, *per impossibile*, we were riding on the front of a beam of light. In such scenarios, one is allowed to imagine what would happen if some contingent and variable limitations (such as the speed of human activity, the capacity of memory, and the reliability of operation) were idealized.[15]

What secures the relevance of Searle's scenario is the idea that any digital computer program could, 'in principle', be handworked in the way that the CRA demands. This is guaranteed by the fact that the person in the Chinese

[15] For serious philosophical reservations about thought-experiments, however, see Wilkes (1988).

Room has almost exactly the same resources Turing granted the human computer whose performance is modelled by what we now call 'Turing machines' (some sheets of paper, some very simple rules and instructions, and the wherewithal and means to follow them). To object, as some have, that the idealized version of Searle in the Chinese Room doesn't really constitute a computer, is to cut the ground from under Turing's original analysis of computation, and to jeopardize the basic concept of computer science. Criticisms of the CRA (along with new analyses of computation) must avoid implying that a person performing a computation is not really computing. If a Turing machine isn't an appropriate model of a human computing, such machines have nothing like the scientific importance currently attached to them. The beauty, as well as the import, of the CRA, is its close proximity not just to the Turing Test scenario, but also to the original explanation of a Turing machine. And the Chinese Room is no stronger an idealization than such a machine, since it abstracts from the human computer in much the same way (by ignoring limitations of speed, memory, and reliability, for example).

What is the Chinese Room *Argument*?

So far we've looked in an informal way at the Chinese Room scenario—the thought-experiment—and some of the conclusions Searle draws from it. As we shall see, Searle sets out the underlying *argument* against Strong AI and computationalism in different ways on different occasions. However, he never explicitly presents it as a piece of reasoning about the thought-experimental scenario. If it were presented thus, its premises would presumably be:

1. The person in the room has access only to the formal, syntactic features of the symbols he or she is presented with.
2. To understand the Chinese input, the person in the room would need access to the semantic features of those input symbols.
3. No set of formal or syntactical principles is sufficient for understanding.

But what exactly is this argument's conclusion? *If* the conclusion pertains only to the person in the room (if, for example, it's simply that the person in the room doesn't understand the Chinese input), then it's relevant to Strong AI only if that view makes a claim about the analogous part of a suitably programmed computer. It's often argued that Strong AI makes no such claim.

Exactly how is the Room supposed to be analogous to a computer? Searle says that when ensconced in the Chinese Room he 'simply behave[s] like a

computer', is 'simply an instantiation of the computer program', that he *is* the computer and that he has 'everything that artificial intelligence can put into me by way of a program' (Searle 1980*a*: 418 (B69–70)).[16] The English-language rules constitute the computer program, and the first two batches of symbols the database to which the program has access.

However, some commentators, such as Ned Block and John Haugeland, urge that the person in the room is analogous not to the computer, as Searle usually claims, but *only* to its *central processing unit* (CPU), the executive part of the computer which controls and coordinates everything else happening in the machine. Haugeland, for example, argues here that Searle fails to apply his own proposed criterion for adequate theories of mind, asking himself only what it would be like to be *part of* the Chinese-understanding system, rather than the system itself. These commentators then remind us that Strong AI's claim is *not* about the CPU, but about the computer as a whole, the entire *system*. However if, as Copeland points out in what he calls the '*logical reply*', the Chinese Room Argument is supposed to pertain to the system as a whole, then although germane, it isn't watertight. It would then be of the form: 'No amount of symbol manipulation on the person's part will enable him to understand the Chinese input, therefore no amount of such manipulation will enable the wider system of which he is a part to understand that input' (Copeland 1993: 125).

Since the conclusion is about a system that isn't even referred to in the premises, *this* Chinese Room Argument (as it stands) must be logically invalid. It commits, as Haugeland puts it, a part–whole fallacy.

As I mentioned, Searle never presents the Chinese Room Argument in this way. In the abstract of his original 1980 article, he set it out as a derivation from axioms, thus:

(1) Intentionality in human beings (and animals) is a product of causal features of the brain.
(2) Instantiating a computer program is never by itself a sufficient condition of intentionality,

therefore

(3) The explanation of how the brain produces intentionality cannot be that it does so by instantiating a computer program.

[16] '*[N]o digital computer solely in virtue of its being a digital computer has anything that I don't have*' (Searle 1987*a*: 213; emphasis in original); '[N]o computer just by running the program has anything the man does not have' (Searle 1989: 45).

(1) is supposed to entail:

(4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain.

(In more recent work, Searle explicitly says '*threshold* causal powers', since the brain may have more than is necessary to produce mentality. See e.g. Searle 1997: 158–9, 191, 202–3.) And from (2) and (4) is supposed to follow:

(5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. (Searle 1980*a*: 417, not reproduced in Boden)

The central argument of the original article, Searle tells us, is directed towards establishing axiom (2).

Whenever he presents what he thinks of as the underlying CRA's abstract logical structure (Searle 1984: 39, 1987*b*: 296–7, 1990*a*: 21, 1989*b*: 703, 1997: 11–12, 109, 1999*a*: 39), however, or the *point* of the Chinese Room scenario (1991: 526), Searle always does so as follows:

1. Programs are purely formal (syntactical).
2. Minds (human ones, at least) have semantics, mental (i.e. semantic) contents.
3. Syntax by itself is neither the same as, nor sufficient for, semantic content.

Therefore,

4. Programs by themselves are not constitutive of nor sufficient for minds.

It's noteworthy that in such presentations of what we might call (following Searle, and Hauser 1997) 'the Brutally Simple argument', the Chinese Room scenario is said merely to *illustrate* or remind us of the truth of premise (3), rather than to constitute the argument against Strong AI. That so many commentators took the scenario to constitute (i.e. exhaust) the argument is why Searle felt able to say, in 1987:

> [I]n all of the vast amount of literature that has grown up around the Chinese room argument, I cannot see that any of my critics have ever faced up to the sheer logical structure of the argument. Which of its axioms do they wish to deny? Which steps in the derivation do they wish to challenge? (1987*b*: 301)

Whether treated as the argument underlying the Chinese Room scenario, as a streamlined reformulation of that argument, or as a separate though related piece of reasoning, the Brutally Simple argument does attract its own

commentators. Dennett, perhaps the most determined critic of the Chinese Room Argument, has explicitly denied all three of its premises (Dennett 1987). In this volume, Copeland, Haugeland, and Hauser concern themselves explicitly with the Brutally Simple argument. Haugeland, for example, seeks to show that serious AI, while not committed to denying Searle's logical truth (that syntax is not sufficient for semantics), can respond to the CRA by denying that computer programs are *purely* syntactical. To do so, he outlines the conceptual foundations of AI in a way that takes account of the causal powers of programs and data.

The 'Systems Reply'

Copeland's 'logical reply' challenges only the logical validity of the Chinese Room Argument, and doesn't purport to take a stand on its conclusion. But it seems a short step to what Searle calls the 'Systems Reply', that although the person in the Room doesn't understand Chinese, the entire system comprising the Room and its contents does so. This is one of the replies to his Chinese Room scenario that Searle identified before publishing it, and then spent much of his original article responding to (thereby recapitulating, ironically?, the structure of Turing's 1950 paper).

Searle finds the Systems Reply deeply implausible, 'totally unmotivated' except by behaviourism and the question-begging Turing Test (Searle 1980*a*: 419 (B74), 1980*b*: 453, 1982*c*: 346), and thinks it a 'desperate move' on the part of the defender of Strong AI (1999*a*: 39). Whether this is so depends on the integrity of the analogy between the person in the Room and the computer: if the person is analogous only to the CPU (as Searle himself allows at one point (Searle 1987*b*: 297)), the Systems Reply, far from being a *reply* to the Chinese Room Argument, is actually what the defender of Strong AI was saying all along, and the original argument was misdirected. The cogency of Searle's argument then depends entirely on the success of the Systems Reply, or some such similar claim about the kind of thing which his opponents think is capable of understanding.

This means that Strong AI and Searle only really lock horns within his critique of the Systems Reply, which consists of several points. First, that the System has no more means of attaching meaning to, or interpreting, the Chinese symbols than the person in the Room had. It 'has a syntax but no semantics' (Searle 1982*a*: 5). To escape the feeling that the other elements of the System (ledgers, pieces of paper, walls, etc.) might matter in this respect, Searle urges that the

person in the Room 'internalizes' these other elements (memorize the rules in the ledger, do all the calculations in his or her head, etc.) and works outdoors. Even so, Searle insists, he or she would still not understand Chinese. Secondly, Searle claims that it's wildly implausible to think that 'while a person doesn't understand Chinese, somehow the *conjunction* of that person and bits of paper might understand Chinese' (Searle 1980*a*: 419 (B73)), simply because the mere addition of the Room, and the pieces of paper with Chinese symbols and English instructions, can't possibly make the difference between understanding and not understanding. If the person alone doesn't understand Chinese, no amount of adding *these* kinds of things will turn the resulting conglomeration into something which does so. Thirdly, Searle urges that the Systems Reply has independently absurd consequences. By implying that 'all sorts of noncognitive subsystems [such as stomachs, livers, etc.] are going to turn out to be cognitive' (ibid. 420 (B74)) it stands convicted of being hopelessly over-liberal. (This is the familiar problem of *non-standard realizations*, often directed towards functionalism.)

Searle treats the first consideration as the decisive objection to the systems reply (e.g. Searle 1993*b*), but I find the second more persuasive. A person plus some pieces of paper and walls just *isn't the right kind of thing* to be a properly basic subject of mental phenomena. When we ascribe such phenomena to systems which include people as parts, which we sometimes do, either our ascriptions are not what Searle would call 'literal ascriptions of intrinsic intentionality' (1982*c*: 346), or the talk can always be 'cashed out' in terms of the mental phenomena exhibited by the people in question. Of course, systems can have important properties which none of their creature components have. But whether this is true of specifically *psychological* properties is another matter. Here, I think, defenders of the Systems Reply have been over-hasty. We could make the underlying principle explicit thus:

> If a system (not just a creature, but an arrangement in which one or more creatures is embedded) really exhibits some genuine psychological phenomenon ϕ, it does so only in virtue of one or more of its component creatures exhibiting ϕ.[17]

This secures the logical validity of the argument based directly on the Chinese Room scenario by adding a premise to the effect that if the person in the Room doesn't understand Chinese, the Room itself cannot do so either. The suggested

[17] A qualification must be added to allow for collective psychological phenomena, such as decisions, which result from compromises between the individual decisions attributable to the component creatures, but it isn't germane here.

principle fits with Searle's insistence that systems such as business corporations have no intrinsic mental phenomena other than those of their employees, officers, etc. (Searle 1982*c*: 346). I believe it also improves upon his well-known assertion that we're justified in making literal ascriptions of intentionality to things because they are (at least) 'made of similar stuff to ourselves' (Searle 1980*a*: 421 (B80)), i.e. because they have the same biochemistry. The relevant similarity is not so much in the stuff things are made of, but in the *kind* of things they are.

However, none of this will impress or help Searle, since he agrees with proponents of the Systems Reply in rejecting such principles. He explicitly denies, in an early reply to Dennett, that his objection to the Systems Reply is that a system can have no psychological properties not possessed by its subsystems (Searle 1982*b*: 57). And he (rashly, I think) flouts the principle in question when he insists that *brains* (as well as their owners) have psychological states (1980*b*: 451); brains, after all, aren't creatures.

The 'Robot Reply'

More natural than the Systems Reply, perhaps, as a reaction to the Chinese Room scenario, is the idea that there's no understanding there because there aren't the right kind of causal relations between the symbols in the program and the things they refer to. Within a computationalist framework this becomes the 'Robot Reply', according to which an appropriately programmed robot, in rich causal contact with its environment (capable of reacting to stimuli, negotiating terrain, and operating upon things) would indeed have genuine understanding (and other mental phenomena).

Searle responds, first, that this move concedes the falsity of Strong AI (and computationalism) because to insist that syntax plus external causation would produce semantics is to *admit* that syntax is insufficient for semantics (Searle 1987*b*: 297). Causal theories of 'content' that might be taken to relate the robot's states and activities to the meanings of the terms it uses have recently become popular in the philosophy of mind (Rey mentions these). But they run strongly counter to the sort of 'internalist' semantic theory that Searle advocates, since they all attempt to reduce semantic relations to non-intentional natural phenomena. Such attempts, he urges, will fall prey to a combination of common-sense and technical objections. Among the former are that they leave out the essence of mental content, which is *subjectivity*, that they leave out intentionality

itself, and (most crucially, I suspect) that they leave out the *normative* dimension of intentional concepts (Searle 1992: 50–1). Intentionality, in short, is *irreducible*.

Secondly, Searle again presses the Chinese Room into service, appropriately modified, to show that adding robotic 'perceptual' and 'motor' capacities adds no understanding whatsoever to the original program. A person inside a computer inside a robot, after all, is still presented *only* with symbols to manipulate, and still has no way of coming to know what they mean. That the resulting arrangement would be far more impressive, in its behaviour, than a 'bed-ridden' computer, merely testifies to its being able to pass a more sophisticated kind of Turing Test. But Searle's point is that Turing Testing is *always* inadequate, since such tests always fall short of establishing the existence of genuine psychological phenomena. The Robot Reply, he concludes, 'had the wrong level of causation. The presence of input-output causation that would enable a robot to function in the world implies nothing whatever about the presence of bottom-up causation that would produce mental states' (Searle 1987*b*: 300).

In this volume, Selmer Bringsjord and Ron Noel pursue this issue in depth. They begin by responding, on Searle's behalf, to the modified Robot Reply favoured by Stevan Harnad, which sets the bar for simulation higher than the Turing Test. Harnad, they argue, follows Dennett in attempting to combine the Systems Reply with the Robot Reply. So they identify what they call the 'missing thought-experiment', a scenario which would combine Searle's responses to both. They approach this thought-experiment, which involves surgically blurring the distinction between people and robots, via computational simulations. Just as it is possible to produce behaviourally accurate simulations of simple animals which have been 'zombified' (rendered incapable of having sensory experience), Bringsjord and Noel argue, it's possible to imagine a scenario in which a person, implementing the parts and processes of a robot, behaves normally and yet has no experience at all.

The Brain Simulator Reply

Searle sprung his thought-experiment on the world of AI at a time when the 'classical' symbolic approach dominated the field. In the late 1970s and early 1980s, the other main kind of AI research, concerned with neural networks, was at a low ebb, at least partly because people had been convinced (by Minsky and Papert 1968) that the kinds of neural nets studied up until that time simply could not compute important functions. Nevertheless, Searle's original article discusses an objection this sort of research suggests.

Searle initially found it ironic that AI should start appealing to the workings of the brain in order to evade the Chinese Room Argument. As he put it, 'I thought the whole idea of strong AI is that we don't need to know how the brain works to know how the mind works' (Searle 1980*a*: 421 (B77)). But neural network research doesn't represent an alternative to 'Strong AI'. Rather, it has its own version of the Strong AI claim (connectionist Strong AI) according to which an appropriately configured and trained connectionist network would have (the relevant) genuine psychological properties, and would do so purely in virtue of its having the configuration and training in question.[18]

Searle isn't alone in thinking that even weak AI of the 'classical', symbolic kind can now, at the beginning of the twenty-first century, be judged to have failed (Searle 1995*b*, 205ff.). Some of his most determined computationalist opponents would agree. However, although more sympathetic to weak AI of the connectionist kind, Searle consistently deems the above sort of appeal to neural networks irrelevant, on the grounds that the computational power of neural networks is no stronger than that of Turing machines. We know from the Church–Turing thesis, he asserts, that 'any computation you can do on a parallel machine, you can also do on a classical machine' (Searle 1999*a*: 39, see also 1990*a*: 22, 1993*b*, and 1999*b*: 37), and therefore the original CRA applies. Recent developments in neural network research, following the important revival of connectionism in the mid-1980s, aren't supposed to affect this claim. The contemporary connectionist *style* of computation (massively parallel, featuring distributed 'representations'), he also urges, cuts no ice in this respect.

Copeland's chapter in this volume vigorously disputes this understanding of the Church–Turing thesis, even though Searle shares it with most cognitive scientists. Copeland also convicts Searle's attempt to extend the CRA to connectionism of what he calls the *simulation fallacy*, the fallacy of supposing that if x is a simulation of y and y has property Φ, x has property Φ.

However, in response to connectionist Strong AI Searle also, just in case, formulates the 'Chinese Gym' thought-experiment:

> Imagine that instead of a Chinese room, I have a Chinese gym: a hall containing many monolingual, English-speaking men. These men would carry out the same operations as the nodes and synapses in a connectionist architecture . . . and the outcome would be the same as having one man manipulate symbols according to a rule book. No one in the gym speaks a word of Chinese, and there is no way for the system as a whole to learn

[18] The last clause here is needed to rule out appeals to non-computational (e.g. electrochemical) properties of networks, which Searle allows might make the difference between duplicating and merely simulating mental phenomena (Searle 1993*b*).

the meanings of any Chinese words. Yet with appropriate adjustments, the system could give the correct answers to Chinese questions. (Searle 1990*a*: 22)

Copeland suggests not only that the logical reply is once again sufficient to refute this version of the argument, but that it also either commits the simulation fallacy or begs the question.

The broader underlying issues are also addressed here by neuroscientists. Igor Aleksander and John Taylor, although sympathetic to the original thrust of the CRA, both agree that recent neural modelling shows how to escape its clutches. According to them, a certain kind of inner experiential perspective which is a necessary precondition for genuine computer understanding is now within reach. Aleksander examines the way in which neural modelling bears on whether there could be a computational form of intentionality. He argues that special, emergent representations ('neural depictions'), developed in the context of a recent research programme on artificial consciousness, not only capture ego-centred experience of the world, but may also have an emergent intentionality. Taylor, meanwhile, seeks to expound the kind of semantics which might be used by the brain. Invoking evidence from functional brain imaging and deficit disorders, he suggests a neural model of language-processing within the frontal lobes in which semantic relations figure as 'virtual actions', residues of previously taken bodily movements. A virtual action-based semantics, in Taylor's view, leads to the possibility of meaning being attached to the computer's symbolic representations of external objects, and thus shows, he argues, that the CRA's challenge can now be answered.

Syntax and Semantics

A large part of the power of the Chinese Room Argument derives from its being premised on a distinction from linguistics which lies near the heart of contemporary cognitive science, between *syntax* and *semantics*, syntactic and semantic features or properties. Thinkers from outside computational cognitive science, such as some followers of Wittgenstein, have occasionally tried to challenge the integrity of this distinction. But that move isn't available to computationalists.

It's impossible to overestimate how much Searle wants to wield, and rely on, this distinction.[19] Many of his replies to critics of the CRA involve him repeating

[19] In the original article, though, the terms 'syntax' and 'semantics' appear infrequently, and only in the final sections, since the point is put there largely in terms of the distinction between form and content.

his fundamental challenge: neither the person in the Room, nor the programmed electronic digital computer, nor the system comprising the entire Chinese Room arrangement, nor a robot incorporating the Chinese understanding program and interacting causally with appropriate things in its environment have any way of attaching meanings *to* the symbols they are manipulating, or any way of finding (rather than just hallucinating) the meaning *in* those symbols. 'Formal symbols by themselves can never be enough for mental contents, because the symbols, by definition, have no meaning (or interpretation, or semantics) except insofar as someone outside the system gives it to them' (Searle 1989*a*: 45). The only meanings that the symbols in the program have is given to them by something (namely, someone) *outside* the system. Someone *inside* simply has access only to the syntax, and no way of 'getting from' the syntax to the semantics.[20]

What Searle does with the distinction, as we've seen, is to use the Chinese Room scenario to remind us of what he calls a 'very simple logical truth, namely, syntax is not sufficient for semantics' (Searle 1984: 34, 39, 1987*a*: 231, 1987*b*: 296, 1989*a*: 45, 1989*b*: 701, 1990*b*: 21, 1992: 200, 1993*a*: 15, 1997: 12).

Nothing in Searle's arguments is supposed to depend on the present state of computer technology. This derives from the fact that they pertain to what it is to be a digital computer, its original *definition*. And this, of course, is the Universal Turing machine. Turing machines are precisely what Searle has in mind when he writes: 'Digital computer programs by definition consist of sets of purely formal operations on formally specified symbols. The ideal computer does such things as print a 0 on the tape, move one square to the left, erase a 1, move back to the right, etc.' (Searle 1982*a*: 4). Digital computers, on this conception, are by definition symbol-manipulating devices (1990*a*: 20). Computers can be said to manipulate symbols, but the symbols 'have no meaning; they have no semantic content; they are not about anything' (1984: 31). They have to be specified purely in terms of their formal or syntactical structure. The operation of computers can, by definition, be specified purely formally, in terms of abstract symbols (1984: 30). A digital computer is 'a device which manipulates symbols, without

[20] Sometimes Searle writes as if the symbols do have meanings, but they can't be accessed by the person in the room. At other times he says that they have no meanings at all (e.g. when he says that while in the Room he doesn't understand Chinese because 'the entire system, me, symbols, rule books, room, etc., contain only Chinese symbolic devices but no meanings' (Searle 1989*a*: 45)). I take it that this isn't a vacillation: the underlying view is that the symbols have meaning, but only in so far as someone outside the system gives them meaning, and that without going outside the system, it's impossible to work out their meanings. (What on earth would a meaningless symbol be, anyway?)

any reference to their meaning or interpretation' (1989*a*: 45). Likewise, computer *programs*, by definition, are 'purely formal and abstract sets of rules for manipulating symbols' (1989*a*: 45), entirely defined by their formal or syntactical structure. The distinction between formal symbols and semantics, manipulation of syntactical elements and meanings 'applies at every level of program implementation' (1989*b*: 703).

Hauser, however, challenges Searle on the grounds that although programs are purely syntactic, the *processes* in which they run are not. Could Searle have failed to take into account the fact that program runs have (causal and dynamic) properties that inert programs don't? This raises a problem about the transition from the Chinese Room *scenario* to the Chinese Room *argument* that's supposed to represent it. The scenario, of course, involves *running* a program (by hand). But does the first premise of the 'Brutally Simple' argument survive if we focus on running programs, rather than inert ones?

Intentionality and the Brain: Searle's 'Biological Naturalism'

According to computationalists, Turing solved the problem of rationality (or intelligence) by showing how, given states which have both syntactic and semantic features, minds could process them in virtue of the former in a way that doesn't mangle the latter. Turing did not, however, ask (let alone answer) the question 'What is it for a state (physical or mental) of an organism or device to *have* semantic features?'. Turing's idea doesn't address this, the problem of *intentionality*, but it isn't supposed to.

The problem of intentionality, however, is precisely where Searle's main interest lies. In contrast to digital computers as Searle presents them, you and I do have ways of attaching meanings to, of discovering the meaning in, and of understanding, the symbols we use (in thought or language). We have access to both syntax *and* semantics. '[W]e know from our own experience that the mind has something more going on in it than the manipulation of formal symbols; minds have contents' (Searle 1997: 10). 'By virtue of their content', says Searle, our thoughts, perceptions, understandings, etc. 'can be about objects and states of affairs in the world' (1990*a*: 21). We know that the symbols we use *are* symbols, that their purpose is to represent, and we know their meanings. The mental, semantic 'contents' the mind has are 'biologically produced by the brain' (1990*a*: 24).

This brings us to a large topic, Searle's own positive theory of mental phenomena, which he calls '*biological naturalism*', and which has been no less controversial than his negative arguments. Its major theme is that mental phenomena are *biological* phenomena (1987*a*: 217). Searle applies biological naturalism principally to consciousness, which has been the focus of his more recent work. In this volume Penrose, and Coulter and Sharrock dispute his idea that consciousness is a biological phenomenon.

Biological naturalism is founded on the idea that the relation between mental phenomena and brains is best understood on analogy with that between microstructural features and the macrostructures they are features of. Water, for example, consists of micro-particles, molecules of H_2O. These have a certain limited range of properties, like weight, dimension, chemical valence, etc., and also stand to one another in a limited range of relationships. But the objects these micro-particles compose, physical 'systems' such as particular bodies of water, also have 'surface' or 'global' properties such as liquidity, solidity, and transparency (sometimes called 'emergent properties'). Such surface or global properties, according to Searle, can be causally explained by the behaviour of elements at the micro-level. So the liquidity of a particular glass of water is supposed to be explained by the structure of and interactions between the H_2O molecules of which it is composed. In short: micro-properties can (sometimes) cause macro-properties, and when they do, they also causally explain those macro-properties.

Searle's idea is that the relationship between micro-properties and macro-properties serves as a good model for the mind–brain relation, and thus dispels some of the mystery in the mind–body problem. (Not all of it, because he also thinks purely natural physical phenomena are themselves pretty mysterious.) His doctrine on the mind–body problem is thus that mental phenomena are both *caused by* brain processes and '*realized in*' the physical system we call the brain. This is what allows mental phenomena to enter into causal relations with other physical phenomena. Properties like consciousness are (emergent) properties of brains (macro-level objects), not of their micro-level constituents (particular neurons, or anything smaller, like molecules). (This is one reason why it's important to Searle to be able to attribute psychological phenomena to brains, rather than just to people.) Individual neurons aren't themselves conscious. Nevertheless, consciousness is caused by the operations of neurons, the component parts which brains are (partly) made of.

Searle's conception of intentionality as a *biological* property concerns several contributors to this volume. Alison Adam's paper directly addresses his use of

the concept of intentionality to distinguish genuine from counterfeit minds. Invoking the work of anthropologist Mary Douglas, Adam interprets Searle's use of the distinction as a vestige of a problematic picture of the status of humans, our relation to nature, and the deeply valued boundary between us and machines. She seeks, instead, to guide us beyond the antagonism between philosophy and AI which arguments such as the CRA seem to have generated, as well as beyond the existing spectrum of options on the relationship between humans and machines. In order to do so, she appeals to recent research in social science (actor-network theory) and feminist thought (Donna Haraway's 'cyborg feminism'), which blur the boundary in question.

Adam's constructivist approach to social reality contrasts vividly with Searle's own writings on these topics (Searle 1995*a*), but has something in common with Kevin Warwick's claim, in this volume, that arguments like Searle's must involve species-bias. Warwick, at the forefront of research on cyborgs, has first-person experience of an interface between human tissues and silicon-based electronic computational devices, and is keen to erase any suggestion of a boundary between humans and machines. Taking Searle to deny the possibility of computer-controlled conscious robots, he argues that the CRA's attack on computer understanding can be rebutted if the machines in question can be given consciousness. Using Searle's own premise that consciousness is a feature of the brain, he suggests that there is a continuum of consciousnesses, that different kinds of brains have different kinds of consciousness, and that neither simulated silicon brains nor artificial neural networks can be denied consciousness of some kind. He then seeks to show that computers can indeed be conscious, deploying something like an operational definition of that term, claiming that humans are programmed too, in their genes, and appealing to 'machine learning'. For Warwick, Searle's suggestions that creatures with brains are special because of their *biology* is not to the point, since certain kinds of existing robots are already 'non-biological' living things.

Coulter and Sharrock, while finding merit in the CRA, also take serious issue with Searle's theory of intentionality (in part following work by other commentators of a Wittgensteinian orientation (e.g. Malcolm 1991, Hacker 1992)), particularly with his view of how language is dependent on the mind. Searle's theory issues in problems such as how minds impose intentionality on linguistic entities that aren't intrinsically intentional, and how one 'attaches' meanings to the words going through one's mind. These, Coulter and Sharrock argue, are pseudo-problems, deriving entirely from Searle's overly mentalistic renditions of concepts such as meaning and understanding. They return to

Searle's theory of speech acts (Searle 1969) in order to critically evaluate the foundations of his approach to language.

Diane Proudfoot, meanwhile, investigates resemblances between the CRA and Wittgenstein's own remarks on cognition (some of which were spurred by dialogue with Turing). Despite agreement between Wittgenstein and Searle that symbol-manipulation is insufficient for, and hence cannot constitute, understanding, and that computing machines don't genuinely follow rules, Proudfoot argues that Wittgenstein's perspective provides persuasive objections to the CRA, and to the model of mind underlying it. Like Hauser, she finds Searle making an illicit appeal to 'first-person authority' to decide whether he understands Chinese. But Proudfoot also argues that Wittgenstein's later work has something more positive to offer: a perspective which has more in common with contemporary philosophy of mind, and with new developments in cognitive science.

Searle's New Foundation for Cognitive Science

Searle suggests that cognitive science should begin not from the morass of computationalism, even devotees of which can't agree about their foundational concepts, but from things we already *know* about how the world works. Among these are things that few but the most determined philosophical sceptics deny, such that 'we all really do have mental states and processes', that they are intrinsic to us, and not merely a matter of how others choose to treat us, and that many of these states and processes are intentional (Searle 1993*b*). But as well as these unproblematic theses, Searle supposes that we also now know that brains cause minds, and that brain operations of the right kind are *sufficient for* mental phenomena (that is, for '*pure*' mental phenomena, the ones picked out by non-factive verbs):

> To put it crudely, and counting the rest of the nervous system as part of the brain for the purposes of this discussion, everything that matters for our mental life, all our thoughts and feelings are caused by processes inside the brain. As far as the causation of mental states is concerned, the crucial step is the one that goes on inside the head, and not the external stimulus. And the argument for this is simply that if the events outside the brain occurred but caused nothing in the brain, there would be no mental events, *whereas if the events in the brain occurred the mental events would occur even if there were not an outside stimulus.* (Searle 1987*a*: 222, emphasis added)

This is what Searle calls the *principle of neurophysiological sufficiency* (ibid. 229).

Although widely misrepresented as believing otherwise, Searle allows that having a working organic brain may not be necessary for having a mind (1997: 131, 203). All we can say is that it's a causal precondition for having a mind *as far as we now know*. But does a working brain constitute a mind, and is having certain things go on in a brain sufficient for having mental phenomena? I think the answers to these questions may not be as obvious as Searle suggests. Plenty of contemporary philosophers are convinced that brain events are necessary but not sufficient for important intentional mental phenomena. Most of them (the 'externalists') insist that what's also needed is an appropriate relationship between the brain and the world external to it. Searle's determined 'internalism', epitomized by his insistence that a brain in a vat could have the very same pure mental phenomena as an embodied and socially situated person (Searle 1983: 230), constitutes, to my mind, one of the most problematic aspects of his views.[21]

Computers Don't Follow Rules At All?

The CRA grants the computationalist premise that computer programs can be characterized in terms of syntactic properties (or, equivalently, that computers follow syntactic rules). A more radical option is to deny that computers follow any rules at all. This path, taken by certain thinkers influenced by the later work of Wittgenstein, elicits agreement or sympathy in this volume from Coulter and Sharrock, Proudfoot, and Mark Bishop.

One of its basic points is that even the simplest rule-following operations require agents capable of exhibiting the capacities characteristic of normativity: they must *understand* the rules being followed, be capable of *explaining*, *justifying*, and *correcting* what they (and others) do by reference to the rules in question. As a result, Turing was wrong in thinking that the primitive operations of a Turing machine are truly *mechanical* in the sense that they presuppose no intelligent agent for their execution. Perhaps this is what Wittgenstein meant by insisting that Turing machines are '*humans* who calculate' (Wittgenstein 1980, § 1096).

One way of developing this viewpoint, pursued by Peter Hacker and Stuart Shanker,[22] issues in the claim that Turing (and his followers) profoundly mis-characterized his own achievement. Contrary to the computationalist view, what Turing really did was to design a way of producing the output to

[21] See e.g. Glock and Preston (1995).

[22] See e.g. Hacker (1990, ch. IV) and Shanker (1998).

processes which previously required intelligence *without* their using any intelligence at all. If this is right, we have to take seriously the possibility that what AI engineers are really doing, and the conclusions they are really establishing, are not very much like what they *say* they are doing and establishing. They should be taken as showing how it's possible to design artefacts and get people to treat them as really performing tasks which they merely *appear* to be performing. As a result, when we credit computers with certain achievements, the concepts in question are not used in the same sense as that in which they apply to humans.

Searle doesn't consistently endorse this line of reasoning. For him, computers can indeed be literally credited with performing *some* of the tasks we naturally think of them as performing, even when those tasks are psychological: 'Just by manipulating meaningless symbols the computer can prove theorems, win chess games, and form new hypotheses' (Searle 1989*a*: 45).[23] In fact, he dismisses Wittgenstein's later work as 'part of a larger tradition of seeking behaviouralistic or quasi-behaviouralistic analyses of mental concepts', whose efforts are 'doomed to failure' (Searle 1987*a*: 231). His own views do make contact with this tradition, though, in his claim that the kind of intentionality that computers (and certain other systems) apparently display is not the kind which humans exhibit. Coulter and Sharrock's claim that intentional terms are used in different (but not metaphorical) senses when applied to humans and to machines bears comparison both with the closely related view of Searle, and with Hauser's critique of it. Hauser appeals to linguistics to show that such uses of psychological concepts aren't ambiguous, but are being given a strict and literal application. Searle's distinction between intrinsic intentionality and as-if intentionality (which isn't intentionality at all), Hauser suggests, is inferior to H. P. Grice's important distinction between natural and non-natural meaning, which supports attributions of ambiguity in the right cases, but not in the cases Searle and the Wittgensteinians claim to detect.

Searle's Trivialization and Observer-Dependence Arguments

More recently, Searle has moved closer to this Wittgensteinian view. Since the inception of the debate, his only major change of mind on the issues consists in his withdrawal, a decade after his original paper, of the assumption that

[23] In fact, in more recent writings (e.g. 1999*b*: 36) Searle questions whether, for example, IBM's 'Deep Blue' really plays chess.

computers follow syntactic rules. In later work he presents other arguments which, while still resting on the syntax/semantics distinction, concede less to Strong AI and computationalism than the Chinese Room.

He argues, first, that it's clear from the original definition of the kind of computation that machines are supposed to perform that

(1) For any object there is some description of that object such that under that description the object is a digital computer,

and that

(2) For any program there is some sufficiently complex object such that there is some description of the object under which it is implementing the program. (Searle 1990*b*: 26–7, 1992: 208)

The phenomenon of 'multiple realizability' beloved of computationalists and functionalists, turns into a *universal* realizability which trivializes their doctrines.

Ironically, Putnam, the founder of Turing machine functionalism, had already argued that computational descriptions are too cheap ('*everything has every functional organization*' (Putnam 1988, p. xv; emphasis in original), taking this to mean that functionalism would imply behaviourism (ibid. 121–5)). Searle's trivialization argument has been addressed most vigorously by Copeland and by David Chalmers (Copeland 1996, Chalmers 1996*a*, 1996*b*). In this volume, the issue is pursued by Block, Haugeland, and Rey, who criticize Searle's argument, and by Bishop, who seeks to reinstate Putnam's triviality proof in the face of Chalmers's critique, arguing that Strong AI implies a fanciful and unacceptable panpsychism.

What Searle means by 'the original definition of computation' is simply Turing's specification, in his 1936 paper, of what we now call Turing machines. Searle proposes to return to this because he finds very little agreement among contemporary cognitive scientists on fundamental questions about computation. That specification talks about the machine's elementary operations, which include the ability to print a '0' or a '1' in each square of its indefinitely long tape. (That any Turing machine program can be stated in terms of this set of symbols is perhaps what tempts some to think that computers use binary arithmetic.) But, as Searle points out, these 0's and 1's are not to be thought of as physical inhabitants of the computer: one wouldn't find them if one opened the machine up: 'To find out if an object is really a digital computer, it turns out that we do not actually have to look for 0's and 1's, etc.; rather we just have to look for something that we could *treat as* or *count as* or *could be used to* function as 0's and 1's.' (Searle 1990*b*: 25, 1992: 206).

If Searle is right about this, almost *anything* counts as a digital computer, since anything can be treated as 'running a program' (no matter how simple) consisting of 0's and 1's. Computationalists can accept this. For them, what matters is whether anything with the necessary structure can be treated as running *any* program. If almost any process can count as almost any computation, then the computationalist view of cognition, instead of being the interesting (and empirical) hypothesis its advocates intend, is vacuous. Rather than being a substantial, informative answer to the question 'How does the brain (or the mind) work?', an answer which would pick out some *observer-independent* fact about brains which specifies what processes take place within, the computationalist answer would become trivial. Likewise, to say of a complicated enough system that it's running a particular program would be empty, since such a system will be treatable as if it's running *any* program (compatible with its level of complication). Cognition will be computation simply because almost any process is computational.

According to Searle, all this is a consequence of the deeper point that, as he puts it, *syntax is not intrinsic to physics*. In other words, 'syntax' is not the name of a physical feature or property: 'the ascription of syntactical properties is always *relative to* an agent or observer who treats certain physical phenomena as syntactical' (Searle 1990*b*: 26, 1992: 208; emphasis added); '[s]omething is a symbol only relative to some observer, user, or agent who assigns a symbolic interpretation to it' (Searle 1993*a*: 16, see also 1995*b*: 209–10). We might be able to tighten up the original definition of computation, probably by imposing some *causal* conditions, in order to block the inference to universal realizability. However, Searle says, 'these further restrictions on the definition of computation are no help in the present discussion because the really deep problem is that syntax is essentially an *observer-relative* notion' (Searle 1990*b*: 27, 1992: 209, emphasis added).

Perhaps this is best treated as a second new argument. The idea is that computation is defined in terms of syntax, but syntax is observer-dependent, so computation is observer-dependent too. It can therefore never suffice for semantics, since whether or not a state has a given semantic 'content' is one of its observer-independent or 'intrinsic' features. Therefore, there's no prospect of our ever 'discovering' that something (be it a brain, a mind, or anything else) is a machine carrying out computations *independently of someone's having assigned it such a role*. This new argument concedes less to computationalism than the CRA, since it implies that computationalism doesn't even succeed in being false, but rather is incoherent, having no clear sense (Searle 1993*a*: 15, 1995*b*: 205, 1997: 14). Whereas the CRA, if successful, shows that computation isn't sufficient for cognition, the new argument is supposed to show that it cannot be necessary, either.

This argument from observer-dependence here exercises Block, who responds that the observer-dependence in question is far more limited than Searle thinks, Rey, who urges that the argument is a *non sequitur*, Haugeland, Penrose, and Coulter and Sharrock. Haugeland and Penrose argue forcefully that syntax and computability (respectively) are not observer-dependent. According to Rey, Searle misunderstands the computationalist project, conflating it with extraneous claims that any serious defender of it should reject. Coulter and Sharrock, though, seek to impugn Searle's distinction between the intrinsic and the observer-dependent, discerning there a failure to recognize that functional properties are indeed intrinsic.

Looking Forward: Hypercomputation, Non-computability, and Dynamical Systems

Just how far into the existing and future technology are Searle's arguments supposed to reach? They purport to cover any formal symbol-manipulation system, that is, anything whose operations can be run (rather than just simulated) on a digital computer. As we have seen, Searle already explicitly argued that a modified version of the CRA, the Chinese Gym, covers neural networks and parallel distributed processing.

What about even later developments? In recent years, new kinds of computers (some of which, nevertheless, Turing may have had inklings of) and new kinds of computational environments have come to be thought of as logically or even physically possible. It has been argued, for example, that certain kinds of hypercomputers or 'super-Turing' computers could compute functions which Turing machines can't, and that even Turing machines can compute such functions when embedded within certain kinds of relativistic space-times (Hogarth 1994).

In recent work (e.g. Copeland 1996), Jack Copeland has suggested and explained ways in which the Church–Turing thesis has been mis-stated or misunderstood. In his chapter for this volume, he makes further contributions to this project, germane to Searle's own complaint that cognitive scientists have failed to explain why and how they take such technical results and theses within computability theory to be relevant to questions of how brains work, as well as to questions of how electronic digital computers do so, and whether and how they might constitute or give rise to minds (Searle 1990*b*: 22, 24; 1992: 205). Copeland's work, however, suggests that there may be kinds of computers

which exceed the computational power of human beings, in virtue of using operations that no human being, unaided by machinery, can perform. If there could be such devices, whose programs could not possibly be handworked within the Chinese Room, the CRA, even if sound, would be powerless to refute the claim that the human brain is such a device.

Copeland's views make a vivid contrast with those of Roger Penrose. Although he thinks the original Chinese Room Argument has considerable force against computationalism, Penrose finds Searle too ready to accept the reigning orthodoxy that brains are digital computers (on the grounds that '*everything* is a digital computer'). He famously prefers, as more rigorous than the CRA, an argument from Gödel's theorem, according to which in any formal system there are statements which mathematicians can see are true and yet which cannot be proved true by any computational reasoning within that system. He regards the argument as establishing that there must be *non-computable* neural processes, processes that cannot even be simulated by a Turing machine, underlying human mathematical understanding. This means he opposes Searle over *Weak* AI. Penrose speculates that such processes may take place at the problematic and badly understood borderline between classical and quantum physics and, in particular, that non-computable processes underlying consciousness may take place at the sub-cellular level. The final section of his chapter continues a debate about consciousness that he and Searle have been engaged in for several years.

Another important way of dissenting from the orthodoxy in AI and cognitive science, the *dynamical systems approach* to cognition, is the subject of Michael Wheeler's chapter. The questions which exercise him concern the relation between this approach and Searle's argument, and whether the former provides new insights into the latter. Wheeler's paper includes a vivid introduction to the new approach, characterizing dynamical systems, setting out basic concepts of dynamical systems analysis, and suggesting how dynamical systems research might be conducted. It then tackles the question of the relationship between dynamical and computational systems (defined in terms of the Turing machine paradigm), showing that the latter set is contained within, but by no means exhausts, the former. However, since the difference between these two sorts of systems pertains to the role of *time*, the Chinese Room Argument still applies to dynamical systems, because they are specified in purely *formal* (non-semantic) terms, and the CRA's conclusion is that no such process is sufficient for mentality. Searle could, as Wheeler points out, construct a new version of the CRA in which he manipulates formal elements in accordance with the equations which govern dynamical systems. Invocation of dynamical neural networks would

not block this version of the CRA. However, Searle's own account of the specific non-formal causal powers of the brain which suffice for mentality, Wheeler concludes, is deeply problematic, and the dynamic systems approach, if and when attacked by Searle, bites back. The confrontation, in my opinion, constitutes one of the most important new features of this volume's debate.

As we saw, it's no coincidence that the Chinese Room scenario gives expression to the fundamental concept in traditional computation theory, the Turing machine, via its explicit invocation of what can be achieved by a human computer, idealized in certain respects, but unaided by machinery. Among the challenging questions that Searle's argument poses to these new approaches are: why should we suppose that what the machines in question are doing is really *computation*? Could their operations be repeatable, reliable, and normative enough to count as such? Even if there are logically possible machines which can compute some functions that Turing machines cannot, why should we suppose that their computations could not be simulated by the Chinese Room arrangement? Could there be primitive computational operations that could be performed by a device (whether brain or machine) but not by an unaided human being?

Conclusion

This volume was conceived as a forum in which to provide opportunities to restate the original argument and envisaged responses, to develop those responses, and to indicate lines of argument which Searle did not anticipate. Several contributors have honed and developed their arguments by reference to each other's contributions. Searle was not asked, or given the opportunity here, to respond to the chapters in this volume. But although it's not supposed to contain the last word on the debate, the volume does not simply take stock. Rather, it attempts to latch onto a new phase of the debate, in which detailed analysis and unpicking of the arguments pro and con predominates over the original 'replies' which Searle himself enumerated.

In preparing the volume, the editors became more aware than ever of a sort of consensus among cognitive scientists to the effect that the CRA is, and has been shown to be, bankrupt. Despite the fact that several notable 'names' within the philosophy of mind agree with Searle,[24] it's true that the negative

[24] Baumgartner and Payr (1995) include some of the confessions in a fascinating volume.

consensus among computationalists has become, if anything, even more solid. Some prominent philosophers of mind declined to contribute on the grounds that the project would give further exposure to a woefully flawed piece of philosophizing. Even some who have contributed to the volume think of the CRA not just as flawed, but as pernicious and wholly undeserving of its fame.

Despite this consensus it is notable, however, that there is (still) little agreement about exactly how the argument goes wrong, or about what should be the exact response on behalf of computational cognitive science and Strong AI. We should probably find it extraordinary how much opinions can differ, and how wide the variety of topics which can be raised by, a scenario as apparently simple as the Chinese Room. But Searle's thought-experiment is a microcosm of much contemporary philosophy and cognitive science. Its importance, both philosophically and practically (in its impact on the self-image of current and proposed cognitive science research programmes) has ensured that it has been widely attacked (and defended), with almost religious fervour. It raises a host of issues about mind and mentality, language, meaning and understanding, intentionality, computers, cyborgs, and our self-conception. It can also be used to raise large methodological questions about how cognitive science should be done (computationalism versus 'cognitive neuroscience', versus some more person-centred alternative?), as well as about what philosophy should be ('scientific'? or 'analytic'? or perhaps 'phenomenological'?). At the very least, it forces those involved in contemporary cognitive science into clarifying exactly what general theoretical theses they want to defend.

References

Baumgartner, P., and Payr, S. (eds.) (1995) *Speaking Minds: Interviews with Twenty Eminent Cognitive Scientists* (Princeton: Princeton University Press).

Bechtel, W., and Abrahamsen, A. A. (1999) *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*, 2nd edn. (Oxford: Blackwell).

Bishop, C. M. (1997) *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press).

Boden, M. A. (1987) *Artificial Intelligence and Natural Man*, 2nd edn. (London: MIT Press).

—— (ed.) (1990) *The Philosophy of Artificial Intelligence* (Oxford: Oxford University Press).

—— (forthcoming) *A History of Cognitive Science*.

Chalmers, D. J. (1996*a*) 'Does a Rock Implement every Finite-State Automaton?' *Synthese*, 108: 309–33.

CHALMERS, D. J. (1996*b*) *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press).

CHURCHLAND, P. M., and CHURCHLAND, P. S. (1990) 'Could a Machine Think?' *Scientific American*, 262 (Jan.), 26–31.

COPELAND, B. J. (1993) *Artificial Intelligence: A Philosophical Introduction* (Oxford: Blackwell).

—— (1996) 'The Church-Turing Thesis', *Stanford Online Encyclopedia of Philosophy*.

—— (1998) 'Turing's O-Machines, Penrose, Searle, and the Brain', *Analysis*, 58: 128–38.

—— (2000) 'The Turing Test', *Minds and Machines*, 10: 519–39.

—— and PROUDFOOT, D. (1996) 'On Alan Turing's Anticipation of Connectionism', *Synthese*, 108: 361–77.

DENNETT, D. C. (1985) 'Can Machines Think?' in M. Shafto (ed.), *How We Know* (San Francisco: Harper & Row), 121–45.

—— (1987) 'Fast Thinking', in his *The Intentional Stance* (Cambridge, Mass.: MIT Press), 323–37.

FEIGENBAUM, E. A., and FELDMAN, J. (eds.) (1963) *Computers and Thought*, (New York: McGraw-Hill). Reprinted Cambridge, Mass.: AAAI Press and MIT Press, 1995.

FODOR, J. A. (1995) 'The Folly of Simulation', in Baumgartner and Payr 1995: 85–100.

GARDNER, H. (1985) *The Mind's New Science: A History of the Cognitive Revolution* (New York: Basic Books).

GLOCK, H.-J., and PRESTON, J. M. (1995) 'Externalism and First-Person Authority', *The Monist*, 78: 515–33.

HACKER, P. M. S. (1990) *Wittgenstein: Meaning and Mind, Part 1: Essays* (Oxford: Blackwell).

—— (1992) 'Malcolm and Searle on "Intentional Mental States" ', *Philosophical Investigations*, 15: 245–75.

HAUSER, L. (1997) 'Searle's Chinese Box: Debunking the Chinese Room Argument', *Minds and Machines*, 7: 199–226.

HAYKIN, S. (1999) *Neural Networks: A Comprehensive Foundation*, 2nd edn. (Englewood Cliffs, NJ: Prentice-Hall).

HODGES, A. (1983) *Alan Turing: The Enigma of Intelligence* (London: Unwin).

HOGARTH, M. L. (1994) 'Non-Turing Computers and Non-Turing Computability', in D. Hull, M. Forbes, and R. M. Burian (eds.), *PSA 1994*, i (East Lansing, Mich.: Philosophy of Science Association), 126–38.

KURZWEIL, R. (1998) *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (New York: Viking).

LEIBER, J. (1991) *An Invitation to Cognitive Science* (Oxford: Blackwell).

MCCORDUCK, P. (1979) *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (San Francisco, Calif.: Freeman).

MCCULLOCH, W. S., and PITTS, W. H. (1943) 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of Mathematical Biophysics*, 5: 115–33. Reprinted in Boden 1990.

MALCOLM, N. (1991) ' "I Believe that p" ', in E. Lepore and R. van Gulick (eds.), *John Searle and His Critics* (Oxford: Blackwell), 159–67.

MINSKY, M. L., and PAPERT, S. (1968) *Perceptrons: An Introduction to Computational Geometry* (Cambridge, Mass.: MIT Press).

NEWELL, A. (1980) 'Physical Symbol Systems', *Cognitive Science*, 4: 135–83.

PRATT, V. (1987) *Thinking Machines: The Evolution of Artificial Intelligence* (Oxford: Blackwell).

PROUDFOOT, D., and COPELAND, B. J. (1994) 'Turing, Wittgenstein, and the Science of the Mind', *Australasian Journal of Philosophy*, 72: 497–519.

PUTNAM, H. (1988) *Representation and Reality* (Cambridge, Mass.: MIT Press).

SCHANK, R. C., and ABELSON, R. P. (1977) *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures* (Hillsdale, NJ: Lawrence Erlbaum).

SEARLE, J. R. (1969) *Speech Acts: An Essay in the Philosophy of Language* (Cambridge: Cambridge University Press).

—— (1980*a*) 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3: 417–24.

—— (1980*b*) 'Intrinsic Intentionality', *Behavioral and Brain Sciences*, 3: 450–6.

—— (1982*a*) 'The Myth of the Computer', *New York Review of Books*, 29/7: 3–6.

—— (1982*b*) 'The Myth of the Computer: An Exchange', *New York Review of Books*, 29/11: 56–7.

—— (1982*c*) 'The Chinese Room Revisited', *Behavioral and Brain Sciences*, 5: 345–8.

—— (1983) *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press).

—— (1984) *Minds, Brains and Science* (London: BBC Publications and Cambridge, Mass.: Harvard University Press).

—— (1987*a*) 'Minds and Brains without Programs', in C. Blakemore and S. Greenfield (eds.), *Mindwaves* (Oxford: Blackwell), 208–33.

—— (1987*b*) 'Turing the Chinese Room', in T. D. Singh and R. Gomatam (eds.), *Synthesis of Science and Religion: Critical Essays and Dialogues* (San Francisco and Bombay: The Bhaktivedanta Institute), 295–301.

—— (1989*a*) 'Artificial Intelligence and the Chinese Room: An Exchange [with Elhanan Motzkin]', *New York Review of Books*, 36 (16 Feb.), 45.

—— (1989*b*) 'Reply to Jacquette', *Philosophy and Phenomenological Research*, 69: 701–7.

—— (1990*a*) 'Is the Brain's Mind a Computer Program?' *Scientific American*, 262 (Jan.), 20–5.

—— (1990*b*) 'Is the Brain a Digital Computer?' *Proceedings and Addresses of the American Philosophical Association*, 64: 21–37.

—— (1991) 'Yin and Yang Strike Out', in D. M. Rosenthal (ed.), *The Nature of Mind* (Oxford: Oxford University Press), 525–6.

—— (1992) *The Rediscovery of the Mind* (Cambridge, Mass.: MIT Press).

—— (1993*a*) 'The Problem of Consciousness', *Social Research*, 60: 3–16.

—— (1993*b*) 'The Failures of Computationalism', *Think*, 2: 68–71.

—— (1995*a*) *The Construction of Social Reality* (London: Allen Lane).

—— (1995*b*) 'Ontology is the Question', in Baumgartner and Payr 1995: 202–13.

—— (1997) *The Mystery of Consciousness* (London: Granta).

SEARLE, J. R. (1999*a*) Interview (with Julian Moore), *Philosophy Now*, winter, 37–41.
—— (1999*b*) 'I Married a Computer' (review of Kurzweil 1998), *New York Review of Books*, 47 (8 Apr.), 34–8.
SHANKER, S. G. (1998) *Wittgenstein's Remarks on the Foundations of A.I.* (London: Routledge).
SIEGELMANN, H. T. (1999) *Neural Networks and Analog Computation: Beyond the Turing Limit* (Boston: Birkhäuser).
—— and SONTAG, E. D. (1994) 'Analog Computation via Neural Networks', *Theoretical Computer Science*, 131: 331–60.
SIMON, H. A., and NEWELL, A. (1958) 'Heuristic Problem-Solving: The Next Advance in Operations Research', *Operations Research*, 6: 1–10.
TURING, A. M. (1936) 'On Computable Numbers, with an Application to the *Entscheidungsproblem*', *Proceedings of the London Mathematical Society*, series 2, vol. 42: 230–65 (with corrections in vol. 43: 544–6). Reprinted in M. Davis (ed.), *The Undecidable* (New York: Raven Press, 1965), to which page references are given.
—— (1948) 'Intelligent Machinery', reprinted in B. Meltzer and D. Michie (eds.), *Machine Intelligence 5* (Edinburgh: Edinburgh University Press, 1969), 3–23.
—— (1950) 'Computing Machinery and Intelligence', *Mind*, 59: 433–60.
WASSERMAN, P. D. (1989) *Neural Computing: Theory and Practice* (New York: Van Nostrand Reinhold).
WILKES, K. V. (1988) *Real People: Personal Identity without Thought Experiments* (Oxford: Oxford University Press).
WITTGENSTEIN, L. (1980) *Remarks on the Philosophy of Psychology, Volume 1* (Oxford: Blackwell).

2

Twenty-One Years in the Chinese Room

John R. Searle

I

I want to use the occasion of this volume dedicated to the twenty-first anniversary of the Chinese Room Argument to reflect on some of the implications of this debate for cognitive science in general, and indeed, for the current state of our larger intellectual culture. I will not spend much time responding to the many detailed arguments that have been presented. I have already responded to more criticisms of the Chinese Room Argument than to all of the criticisms of all of the other controversial philosophical theses that I have advanced in my life. My reason for having so much confidence that the basic argument is sound is that in the past twenty-one years I have not seen anything to shake its fundamental thesis. The fundamental claim is that the purely formal or abstract or syntactical processes of the implemented computer program could not by themselves be sufficient to *guarantee* the presence of mental content or semantic content of the sort that is essential to human cognition. Of course a system might have semantic content for some other reason. It may be that implementing this program in this particular hardware is sufficient to cause consciousness and intentionality, but such a claim is no longer Strong Artificial Intelligence. It is at the very heart of the Strong AI thesis that the system that implements the program does not matter. Any hardware implementation will do, provided only that it is rich enough and stable enough to carry the program. This is why I can, at least in principle, carry out the steps in the program in the Chinese