PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

New Breakthroughs or Dead-Ends?

Margaret A. Boden

Phil. Trans. R. Soc. Lond. A 1994 **349**, 1-13
doi: 10.1098/rsta.1994.0109

New breakthroughs or dead-ends?

By Margaret A. Boden

School of Cognitive and Computing Sciences, University of Sussex, Brighton BN1 9QH, U.K.

Artificial intelligence (AI), at its inception, offered new concepts for formulating psychological theories and a new methodology for testing them. It also promised an 'existence proof' that intelligence could be implemented in a physical system. These promises are still controversial, both in AI and in philosophy.

Some researchers favour connectionism, a form of AI that has blossomed relatively recently. Others believe 'classical' AI insights are needed to model many types of human thinking. Some eschew classical AI (and the associated frame problem) in favour of robots 'embedded' in the real world. Similarly, some reject functionalist interpretations of AI, arguing that intentionality cannot be grounded in syntactic and/or simulated and/or non-evolved systems. Consciousness is highly problematic: many doubt that any computational (or even scientific) account could explain it.

The papers presented at this Royal Society/British Academy meeting explore these issues. Even without dead-ends, the routes taken in AI accounts of the mind may lead in unexpected directions.

1. The promises

Artificial intelligence (AI), at its inception, promised new concepts for formulating psychological theories and a new methodology for testing them. Both promises were scientifically exciting, for they offered new breakthroughs towards the long-awaited science of the mind. Both are still controversial also. (It is unclear, for example, how to decide which aspects of a program are theoretically significant and so, up for testing, and which are mere implementation details (Pylyshyn 1984).) However, the first is the more interesting philosophically.

Any psychologist – indeed, any scientist – can clarify a theory and test its coherence and implications, by implementing it on a computer. However, AI-influenced 'computational' psychologists do more than this. They use computational ideas as substantive concepts in theorizing about mental processes. The claim is that AI models instantiate (something like) what is really going on in intelligent organisms.

Different computational psychologists favour distinct types of AI model. These include symbolic AI, or good old-fashioned AI (GOFAI) (Haugeland 1985); various forms of connectionist AI; and some less familiar AI approaches, such as situated and evolutionary robotics. Clearly, the term 'AI' is not being used here (as sometimes) to refer exclusively to GOFAI. The restrictive use is misleading. Not only do GOFAI and connectionism share the same historical roots, but they, along with

Phil. Trans. R. Soc. Lond. A (1994) **349**, 1–13
Printed in Great Britain

more recent approaches, are part of the same general scientific enterprise. To restrict 'AI' to symbolic models is like restricting 'physics' to good old-fashioned physics. In its wider sense, AI covers any computational attempt to understand cognition. (This definition is imprecise, for there is no universally agreed definition of computation; for present purposes, however, this may be ignored.)

AI theorists sometimes assert that the execution of certain computational processes is necessary and sufficient for intelligence and for semantic processes such as designation and interpretation (Newell & Simon 1972; Newell 1980). John Searle (1980) calls this philosophical claim 'strong AI', and (using the notorious 'Chinese Room' example) argues that it is fundamentally mistaken. On his view, symbolic AI deals only with syntax, not semantics, so cannot account for meaning, intentionality or understanding. Formally specified causal processes cannot explain thought of any kind. Connectionism, too (which he likens to a 'Chinese Gym'), he sees as incapable of explaining meaning (Searle 1990).

A different, equally dismissive critique of the AI project has been mounted by Herbert Dreyfus (1979; Dreyfus & Dreyfus 1988). Drawing on philosophers such as Merleau-Ponty, Heidegger and the later Wittgenstein, he rejects the basic epistemological presuppositions of classical AI. He is less critical of connectionism, but even that fails. On his view, no empirical science can give a philosophically fundamental explanation of mental and bodily skills, because these skills are presupposed by science.

Other philosophers of mind are less dismissive. Some of these commend GOFAI (Fodor 1976, 1987); others reject it, favouring connectionism instead (Churchland 1984; Churchland 1986). Yet others recommend some combination of these AI approaches (Boden 1972, 1991; Clark 1989, 1993; Cussins 1990; Dennett 1992), but in general they argue that AI, in one form or another, can help us understand how meaning, concepts, purpose, creativity and even consciousness are possible.

If any of them is right, then AI (complemented by neuroscience) holds the key to the centuries-old puzzle of how mechanism can support meaning. If Searle or Dreyfus is right, this heady promise cannot be honoured: we do not have any new breakthroughs, only dead ends leading to a solid brick wall right in front of our noses.

2. The beginnings

Strong AI did not spring fully formed from the head of Zeus, or even Alan Turing. It took some time to develop. Most early AI workers, like Turing himself (Turing 1936), used introspection as a source of ideas about information processing. Often, these ideas were used merely as scaffolding, the aim being to construct a program with a satisfactory input–output profile, not to match its internal processing with real thinking. From the earliest days, however, some AI researchers were primarily interested in modelling human minds.

The suggestion that theoretical psychology should be computational was first made in the early 1940s. The seeds of both symbolic and connectionist AI were sown by Warren McCulloch & Walter Pitts (1943). Soon afterwards, they published another seminal paper, focused on the computational potential (and neurophysiological plausibility) of parallel processing, self-equilibrating systems (Pitts & McCulloch 1947). The text and titles of these two papers showed that they sought to solve many philosophical problems, too.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Just over a decade later, computer models of a wide range of psychological phenomena were reported in the first published collections of AI research (Blake & Uttley 1959; Feigenbaum & Feldman 1963; Tomkins & Messick 1962). The most influential of these early efforts was the pioneering research programme of Allen Newell & Herb Simon. Their 'Logic Theorist' (Newell *et al.* 1957) and 'General Problem Solver' (Newell & Simon 1961) were intended as computer models of human problem-solving and drew on studies by Gestalt psychologists. Over the next thirty years, Newell & Simon constructed increasingly sophisticated models of mental processing, and carried out many careful experiments on problem solving (Newell & Simon 1972; Newell 1990) and motor skills (Card *et al.* 1983).

Around 1960, several psychologists were inspired by AI research to couch their theories in computational terms. Some drew on the hierarchical planning introduced by Newell & Simon, whereas others (Rosenblatt 1958; Selfridge 1959; Uhr & Vossler 1963) developed parallel-processing models based on the connectionist ideas of McCulloch & Pitts. (They described their models as 'simulations' of the mind: the view that they were instantiations of mental processes was made explicit only later, by Newell & Simon.)

The first authors to apply AI concepts across the whole range of psychology, from instinct to personality, were George Miller, Eugene Galanter & Karl Pribram (1960). The concepts they used were unavoidably primitive, and their work highly speculative. (Wags commented: 'Miller thought of it, Galanter wrote it, and Pribram believed it'.) But they had vision. Other writers soon followed, using computational ideas to discuss motivation and affect as well as cognition (Abelson 1963; Boden 1965, 1972; Colby 1963, 1964; Neisser 1963, 1967; Reitman 1963, 1965; Reitman *et al.* 1964). Much of this AI-inspired psychology was programmatic rather than programmed. Even so, a significant number of working models was produced. (Since then, of course, many psychologists have expressed their theories in programmed form (Boden 1988); but most later models focus on cognition, not motivation or emotion.)

As well as formulating psychological theories, Newell & Simon made explicitly philosophical claims. (McCulloch & Pitts had done so too, but in a less systematic fashion.) Their 'physical symbol system' hypothesis (Newell & Simon 1972; Newell 1980) clearly articulated the position of strong AI. Passing from simulation to instantiation, they argued that certain types of symbolic process constitute intelligence wheresoever it may be found.

Meanwhile, a new philosophy of mind, functionalism, was being developed by Hilary Putnam (1960, 1967). Mental states, he suggested, are definable in terms of their abstract causal relations with sensory input, motor output, and other mental states. Moreover, these relations are in principle specifiable in computational terms. Putnam did not discuss the infant AI programs, but he referred at length to Turing machines and likened the software–hardware distinction to the relation between mind and body. He saw computers as an existence proof that abstract causal functions can be attributed to purely physical systems, and that a given function (compare: mental state) can be physically implemented in many different ways.

Functionalism was fundamentally similar to Newell & Simon's position, and it was to be no less influential in philosophical circles than theirs was in psychology. Nearly half a century later, functionalism (and its strong-AI equivalent, the physical symbol system hypothesis) is still influential, and still controversial.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

3. Current approaches in AI

Scientific AI is controversial too, not least to the scientists involved. That there is no generally accepted research paradigm is evident from the 50th volume of *Artificial Intelligence* (Bobrow 1993). This special volume (dedicated to Newell, who died shortly before it went to press) shows that AI practitioners disagree about what is the most promising approach.

GOFAI, the most well known type of AI, is still a leading contender. Symbolic modelling is widespread in current computational research, whether psychologically or technologically inspired. It is used to study (for example) problem solving, planning, learning, natural language understanding, analogy, human–computer interaction, the perception and performance of music, and creativity in art and science. A few people are using it to study motivation and emotion, too (Sloman 1987). Among the advantages of this methodology are its ability to represent hierarchical structure, to define 'strong' problem-constraints, and to provide models that are relatively transparent.

A very different, and widely used, approach is connectionism (of various types). Its visibility has burgeoned over the last decade, but its roots lie in the mid-century parallel-processing models mentioned in § 2. One form of connectionism is parallel distributed processing (PDP), wherein representations are implemented not by specific tokens (symbols) but by the overall pattern of activity of a network of computational units (Rumelhart & McClelland 1986). Processing is conceived not as classical computation – do this, then do that – but as transformations of activation vectors (sometimes described in terms of dynamic equilibrium-seeking (Smolensky 1988)). Most connectionist models are learning systems, in which the activation weights on the connections between units can change with experience. These use a family of statistical learning-rules, whose varied potential and limitations are still being explored.

PDP systems lack the three advantages of GOFAI just listed, but offer others in compensation. Their inherent properties provide 'natural' abilities for pattern matching, content-addressable memory, learning by example and graceful degradation, all of which are difficult to program into classical models. The advantage of biological plausibility is weaker than it is often claimed to be: though originally inspired by neuroscience, current connectionism bears only a very sketchy resemblance to processing in the brain. Nevertheless, some models show a strong resemblance to human deficits, such as specific types of dyslexia (Hinton *et al.* 1993).

Most AI researchers today use only classical, or only connectionist, models. Because these have complementary strengths and weaknesses, there is growing interest in 'hybrid' models, which try to get the best of both worlds (Hendeler 1989; Hinton 1991; Stark 1993; Thornton 1991).

Hybrids come in various forms. A connectionist system may mimic, to some extent, the properties of GOFAI machines. For example, recurrent networks, in which information is fed back from higher to lower levels, can capture (tacitly, not explicitly) some features of hierarchical structure (Elman 1991). Conversely, some classical work incorporates bottom-up, parallel processing of subcognitive micro-features (Marr 1982), and the widely used 'blackboard' architecture (Newell & Simon 1972) is a sequential–parallel hybrid. In addition, two different virtual machines (classical and connectionist) may be combined, at a specially designed

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

interface: each part performs the tasks to which it is best suited, control passing from one to the other as appropriate during problem solving (Chrisley 1991).

Both classical and connectionist AI share a commitment to internal representation as integral to intelligent thought and action. This commitment has been abandoned in recent AI work in situated robotics (Brooks 1991*a*, *b*, 1992; Maes 1991; Mataric 1991), seeking thereby to avoid the notorious 'frame problem'. Classical robotics relies on planning, done within an internal world-model. A classical robot must anticipate a host of intended consequences and unintended side-effects (such as the fact that moving a chair also moves its cushion, but does not affect a table six feet away). For many tasks, moreover, it must do this in real time. In the real world, this is impractical. Instead of manipulating complex internal representations of the world, situated robots deal directly with it.

This new approach favours a 'behaviour-based' architecture, as opposed to one based on classical functional decomposition. Situated robots engage in simple, hardwired behaviours, triggered by specific, ecologically relevant, environmental cues. A given type of behaviour can be inhibited by another one (so walking may be inhibited by turning), but the higher-level activity is itself automatically triggered by the current world-input (such as hitting an obstacle). The robot functions even if the higher levels are removed, but is better adapted to its environment if all levels are engaged.

The layered 'subsumption architecture' enables complex, but non-programmed, behavioural patterns to emerge. Obstacle avoidance or wall following, for instance, emerge as a result of the interaction of low-level, localized processes, without any centralized representation of them as such. Current results include insect-like robots that can walk on rough terrain by using their six legs in an apparently coordinated way: the coordination results from the interaction of various specific responses to specific environmental cues, not from general perceptual analysis of the objective properties of the terrain, nor from top-down planning. Apparently goal-directed behaviour, but without goal-representations, can also be engineered in this way (Maes 1990).

Whether high-level human thinking could ever be so modelled is quite another matter. Situated roboticists are now taking a less aggressive line on this than previously (Brooks 1991*a*, *b*). They are also prepared to allow that representations, of a sort, may be involved in fairly simple behaviour. For instance, subject-centred 'representations' can help in tracking a robot's movements within a given niche, but are of no use for general problem solving because they do not refer to an independent, objective world (Mataric 1991). Similarly, the philosopher Andy Clark has argued that some of the individual behavioural modules of situated robots may be described as representing the input classes to which they respond (Clark 1994; Clark & Toribio 1994). This is entirely compatible with there being no central representation of the type favoured in classical AI.

Situated roboticists see their work as more biologically realistic than the classical variety. They describe their robots as 'embedded' in the real world, in contrast to GOFAI robots that live in a world of make-believe: namely, planning. Moreover, they point out that improvements to situated robots are engineered not by wholesale redesign, but by adding some new behavioural module to a creature that already functions in an acceptable manner. This is analogous to biological evolution, wherein creatures must be viable at all times.

However, the added modules of situated robots are designed by the human

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

roboticists. Some AI research aims at a still closer biological parallel, programs (and even hardware) being not so much written as evolved. It uses 'genetic algorithms' (GAs), whereby random mutations in a program give rise to successive generations, at each of which the more useful rules are automatically identified and (probably) used for breeding (Holland 1975). After many generations, perhaps thousands, the system in its evolved form may be highly efficient, or even optimal. This AI technique is used for inductive problem-solving and classification (Holland *et al.* 1986), and for evolutionary art, where the selection at each generation is done by the human user (Sims 1991; Todd & Latham 1992).

Genetic algorithms are also used for evolutionary robotics, in which robots are not consciously designed but automatically evolved (Brooks 1992). For instance, the 'brains' of simple robots, and their sensorimotor morphology, can be evolved adaptively in this way (Cliff *et al.* 1993). The latter case is an example of work in artificial life, or 'A-Life' (Boden 1994*b*; Langton 1989). A-Life studies self-organizing, self-replicating, adaptive systems (individuals and groups), and the emergence of ordered complexity from simple rules. It is closely related to AI. Indeed, because intelligence is a property of living systems, AI can be seen as a sub-area of A-Life.

In A-Life (as in situated robotics), the emphasis is on 'autonomous' systems adapted to their environment, and on a bottom-up approach to complex behaviour. Because the environment is assumed to be noisy, dynamic and largely inconvenient, GOFAI's commitment to representation and top-down planning is rejected. However, as remarked above, this theoretical stance may be inadequate for certain types of behaviour. Whether some types of intelligence require internal representations remains an open question, not least because just what counts as a 'representation' is still unclear.

Evidently then, the AI of the 1990s is intriguingly diverse. Philosophical commentaries, too, have diversified, as outlined in § 4.

4. Associated philosophical debates

There are many philosophical debates associated with AI, of which only a few can be mentioned here. The issues outlined below are selected for their interest to both AI and philosophical researchers.

Searle's rejection of symbolic AI persists (Searle 1992). Criticizing GOFAI for positing internal processes conceptualized in semantic terms, he grants that connectionism can at least show 'how a system might convert a meaningful input into a meaningful output without any rules, principles, inferences or other sorts of meaningful phenomena in between' (Searle 1992, pp. 246–247). But he sees connectionism, too, as unable to explain intentionality, and still believes it to be intuitively obvious that neuroprotein can support intentionality, whereas metal and silicon cannot.

There have been many responses to Searle's argument, both rebuttals and endorsements. My own (Boden 1990) is that we have no specific reason to believe that only neuroprotein can support intentionality. Moreover, the fact that it does so is not obvious, but highly counter-intuitive. Insofar as we understand this at all, we appeal to the information-processing properties of the brain (the role of the sodium pump in enabling message passing, for example), not to its specific

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

material stuff (potassium would do just as well, if it fulfilled the same role). As for his claim that syntax alone cannot give us semantics, I agree. Meaning, purpose and understanding require certain sorts of causal relationship between a system's internal processing and its environment, plus a historical grounding in evolution: intentionality cannot properly be ascribed to an artefact, except in a secondary sense (Boden 1972).

However, causal-evolutionary views of intentionality are not universally accepted. Extended defences have been offered by Ruth Millikan (1984) and Fred Dretske (1988), but there are problems as to how an evolutionary account can explain the occurrence of error, or misrepresentation (Dretske 1986; Fodor 1987, 1990; Millikan 1993). The concept 'information' is contested also, so there is no agreement as to just what an 'information-processing system' is. Moreover, some philosophers oppose any naturalistic account of meaning, whether purely causal or causal-evolutionary (Morris 1992).

Other philosophical disputes concern the nature of our mental architecture. Putnam's functionalism was developed by Jerry Fodor (1976) in his 'Language of thought' theory. Fodor sees mental processes as involving computations over representations, which (like GOFAI) involve combinations of elementary units. The compositionality of language and the generality of thinking – such that someone who can think that John loves Mary must also be able to think that Mary loves John – are said to follow from the computations, and language, of thought. On this view, connectionism is philosophically irrelevant (Fodor & Pylyshyn 1988). Even arguments that connectionism can provide compositionality (Clark 1991) cut little ice. Fodorians may admit that the language of thought is, as a matter of fact, implemented in neural networks. But they insist that thinking is made possible by compositional syntax, describable only in classical, symbolic terms.

Other AI-inspired philosophers disagree, seeing connectionism as the more significant AI methodology. They argue that its emphasis on sub-symbolic processes, its lack of all-or-none rigidity, and its avoidance of explicit rules illuminate long-standing philosophical problems. These include how objective concepts can arise from pre-conceptual thought, how understanding rests on prototypes and family resemblances, how representations arise and how they can change, and how mature thinking – including scientific explanation – is possible (Churchland 1990; Clark 1989, 1993; Cussins 1990; Thagard 1988).

Others, besides, refuse to join the fray. Aaron Sloman (1991), for example, sees no fundamental philosophical difference between the two methodologies. The mental features of most interest to philosophers (and psychologists), he argues, depend on the global architecture of the mind/brain, and the causal relations between its functionally distinct sub-systems. In general, implementation details will be irrelevant to such questions. In practice, a certain function may need to be done (reliably, or in real time) by one type of process rather than another. Also, sometimes low-level biological mechanisms may be closely coupled to high-level psychological functions, as when a neurochemical changes our mood, or alcohol our reasoning powers. Even so, it is the abstractly defined functions which are philosophically important. In short, Sloman is closer to Putnam (and Fodor) than to those who champion a particular type of implementation.

What of consciousness? One of the earliest objections to functionalism was that it cannot account for phenomenal experience, or qualia. Suppose AI were to produce a robot outwardly just like us, passing the Turing Test (Turing 1950)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

in whatever form we cared to pose it. It is still conceivable (so the argument goes) that the robot would not be conscious. It might not be able to smell, taste, hear or see anything at all, despite having the capacity to make all the relevant behavioural discriminations and verbal remarks (to carry out all the relevant causal functions).

Searle claims that a robot made of inorganic materials simply could not be conscious. But at least he allows that neuroscience may one day explain consciousness. Some philosophers despair of any scientific explanation of it, whether computational or not. They argue that subjective consciousness (what it is like to be an experiencing subject) can never be captured by objective descriptions, even though certain structural aspects of experience might be so described (Nagel 1974). Or they argue that the mind–body problem is for ever beyond our capacities to solve, much as quantum physics is beyond the understanding of a dog (McGinn 1991). The latter claim is conceivably correct, but at this stage of the game unnecessarily defeatist. Compared with such a view, Searle is an optimist – about neuroscience, if not about AI.

Among philosophers sympathetic to AI, the greatest optimist about explaining consciousness is Dan Dennett (1991). He points out that there is no such thing as 'the' problem of consciousness. Rather, there are many interrelated problems. Multiple personality disorder (MPD), for example, seems utterly impossible on a cartesian view of the mind. From a computational viewpoint, however, it is intelligible (Boden 1972, ch. 7; Boden 1994*a*; Dennett 1991, ch. 13). (This point is independent of the 'reality' of the clinical syndrome: it holds even if MPD is merely role-playing, by suggestible patients egged on by publicity-hungry therapists.) Dennett offers computationally inspired accounts of various other questions relating to consciousness.

As for the most intractable question of all, Dennett's position can be gauged from the title of the relevant chapter: 'Disqualifying qualia'. He sees no reason to posit the existence of anything over and above behavioural discrimination and internal computation. The richness of our experience is explained in terms of a host of highly sensitive discriminations, coloured (as the saying goes) by all manner of idiosyncratic associations. Qualia 'as such' simply do not exist. Or rather, because he admits he cannot strictly prove this, there is no good reason to believe that they do. This suggestion strikes many as philosophically unacceptable – a polite way of saying 'absurd'. But that is not to say that anyone else has offered a convincing philosophy of consciousness.

As though all these philosophical disputes were not enough, neo-heideggerian murmurings are afoot (Haugeland 1994; Varela *et al.* 1991; Wheeler 1994*a*, *b*). They threaten the fundamental assumptions of AI, for they reject the subject–object distinction presupposed by materialists and idealists alike, and deny the epistemological primacy of science. Similar criticisms come from philosophers who see organisms as dynamic systems closely coupled with their environment (Port & van Gelder 1994).

Heideggerian critiques of AI are not new. They were first mooted by Dreyfus in the mid-1960s (Dreyfus 1965, 1967), and were later expanded (Collins 1990; Dreyfus 1979; Dreyfus & Dreyfus 1988). These past critiques had little effect on the AI community, partly because they often misdescribed contemporary AI research, and partly because they denied the very possibility of a scientific psychology.

However, similar critiques are now being put forward, not only by philoso-

phers inimical to computer modelling, but also by people who favour situated and evolutionary robotics and the simulation of animals and 'animats' (Meyer & Wilson 1991). Instead of positing internal representations of an objective external world, such people speak of whole systems embedded in, and adapted to, their own particular 'worlds'. The early ethologist Jacob von Uexkull (1957) argued persuasively that different animals inhabit different 'worlds', but he ignored the broader philosophical implications, and his pictures of the living-room as seen by the fly (or the dog) fell far short in detail, if not in charm, of current computer models. Seventy years later, computer simulations of actual and imaginary animals are receiving more careful philosophical attention.

It remains to be seen how these neo-heideggerian critiques will affect the aims and practices of scientific AI . It remains to be seen, too, whether the (broadly) empiricist epistemology and realist metaphysics of 'analytic' philosophy will give way to phenomenology.

5. Conclusion

With respect to final verdicts, both juries – scientific and philosophical – are still out. Nevertheless, there is solid evidence that the promises of the 1940s have borne fruit. AI ideas have given direction (and clarity) to many psychological projects. In some areas, they have been no more than suggestive. In others, they have been enormously productive (the best example, here, is vision).

Admittedly, there have been disappointments. For instance, we now know that there can be no truly general problem-solver. But a research area that leads to some dead ends can provide useful alleyways, too. Symbolic work on scene analysis (Boden 1987, ch. 5, 6) turned out to be less useful for low-level vision than bottom-up parallel processing is, but some of its theoretical insights were borrowed by the 'opposition' (Mackworth 1973; Marr 1982). Moreover, future changes of direction are inevitable, for there are many things that cannot be done by any current approach. We have only just begun to explore the space of computational possibilities. Changes of direction, and even the occasional dead end, should not be scorned as folly. Science grows not only by conjectures, but also by refutations (Popper 1963).

In philosophy, conjectures are many and refutations few. Nevertheless, progress can be made. The early hunches (of McCulloch & Pitts, for instance) that AI methods would redirect and revivify the philosophy of mind have been amply borne out. For those who share a broadly materialist standpoint, philosophy has been much enriched. Even anti-materialist philosophers may take on board, albeit at a less fundamental level, some of the new ideas about mental processes. Furthermore, the unthinkable is now being thought: reconciliation between the analytic and phenomenological traditions is being considered, partly because of recent advances in AI. Without doubt, then, the philosophy of mind has been advanced.

It is not surprising that such mutual influences, and advances, were being predicted in the 1940s. History had already shown that science and philosophy can illuminate each other. The papers presented at this joint meeting of the Royal Society and the British Academy illustrate some ways in which this is happening now.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

References

Abelson, R. P. 1963 Computer simulation of 'hot' cognition. In *Computer simulation of personality: frontier of psychological research* (ed. S. S. Tomkins & S. Messick), pp. 277–298. New York: Wiley.

Blake, D. V. & Uttley, A. M. 1959 *The mechanization of thought processes.* London: HMSO.

Bobrow, D. G. (ed.) 1993 Artificial intelligence in perspective. *Artificial Intelligence* **50**, 1–462.

Boden, M. A. 1965 McDougall revisited. *J. Personality* **33**, 1–19.

Boden, M. A. 1972 *Purposive explanation in psychology.* Cambridge, Mass.: Harvard University Press.

Boden, M. A. 1987 *Artificial intelligence and natural man*, 2nd edn. London: MIT Press.

Boden, M. A. 1988 *Computer models of mind: computational approaches in theoretical psychology.* Cambridge University Press.

Boden, M. A. 1990. Escaping from the Chinese room. In *Philosophy of artificial intelligence* (ed. M. A. Boden), pp. 89–104. Oxford University Press.

Boden, M. A. 1991 *The creative mind: myths and mechanisms*, 2nd edn. London: Abacus.

Boden, M. A. 1994*a* Multiple personality and computational models. In *Philosophy, psychology, psychiatry* (Royal Institute of Philosophy lectures 1993–94) (ed. W. Fulford & A. Phillips-Griffiths). Cambridge University Press. (In the press.)

Boden, M. A. (ed.) 1994*b* *The philosophy of artificial life.* Oxford University Press.

Brooks, R. 1991*a* Intelligence without representation. *Artificial Intelligence* **47**, 139–159.

Brooks, R. A. 1991*b* Intelligence without reason. In *Proc. 12fth Int. Conf. on Artificial Intelligence*, pp. 569–595. San Mateo, CA: Morgan Kauffman.

Brooks, R. A. 1992 Artificial life and real robots. In *Toward a practice of autonomous systems: Proceedings 1st European Conference on Artificial Life* (ed. F. J. Varela & P. Bourgine), pp. 3–10. Cambridge, Mass.: MIT Press.

Card, S. K., Moran, T. P. & Newell, A. 1983 *The psychology of human–computer interaction.* Hillsdale, N.J.: Erlbaum.

Chrisley, R. L. 1991 A hybrid architecture for cognitive map construction and use. *AISB Q.* **78**, 31–33.

Churchland, P. M. 1984 *Matter and consciousness.* Cambridge, Mass.: MIT Press.

Churchland, P. M. 1990 *A neurocomputational perspective: the nature of mind and the structure of science.* Cambridge, Mass.: MIT Press.

Churchland, P. S. 1986 *Neurophilosophy: toward a unified understanding of the mind–brain.* Cambridge, Mass.: MIT Press.

Clark, A. 1989 *Microcognition: philosophy, cognitive science and parallel distributed processing.* Cambridge, Mass.: MIT Press.

Clark, A. 1991 Systematicity, structured representations and cognitive architecture: a reply to Fodor and Pylyshyn. In *Connectionism and the philosophy of mind* (ed. T. Horgan & J. Tinson), pp. 198–218. London: Kluwer.

Clark, A. 1993 *Associative engines: connectionism, concepts and representational change.* Cambridge, Mass.: MIT Press.

Clark, A. 1994 Philosophical foundations. In *Handbook of perception and cognition* (2nd edn), vol. 14: *Computational psychology and artificial intelligence* (ed. M. A. Boden). New York: Academic Press. (In the press.)

Clark, A. & Toribio, P. 1994 Doing without representing? In *Synthese* special issue, 'Connectionism & the frontiers of AI'. (In the press.)

Cliff, D., Harvey, I. & Husbands, P. 1993 Explorations in evolutionary robotics. *Adaptive Behavior* **2**, 73–110.

Colby, K. M. 1963 Computer simulation of a neurotic process. In *Computer simulation of personality: frontier of psychological research* (ed. S. S. Tomkins & S. Messick), pp. 165–180. New York: Wiley.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Colby, K. M. 1964 Experimental treatment of neurotic computer programs. *Archs gen. Psychiat.* **10**, 220–227.

Collins, H. M. 1990 *Artificial experts: social knowledge and intelligent machines.* Cambridge, Mass.: MIT Press.

Cussins, A. 1990 The connectionist construction of concepts. In *The philosophy of artificial intelligence* (ed. M. A. Boden), pp. 368–440. Oxford University Press.

Dennett, D. C. 1991 *Consciousness explained.* New York: Little, Brown.

Dretske, F. 1986 Misrepresentation. In *Belief: form, content and function* (ed. R. Bogdan), pp. 17–36. Oxford University Press.

Dretske, F. 1988 *Explaining behavior.* Cambridge, Mass.: MIT Press.

Dreyfus, H. L. 1965 *Alchemy and artificial intelligence.* Rand Corporation P–3244. Santa Monica, California.

Dreyfus, H. L. 1967 Why computers must have bodies in order to be intelligent. *Rev. Metaphys.* **21**, 13–32.

Dreyfus, H. L. 1979 *What computers can't do: the limits of artificial intelligence*, 2nd edn. New York: Harper & Row.

Dreyfus, H. L. & Dreyfus, S. E. 1988 Making a mind versus modelling the brain: artificial intelligence back at a branchpoint. In *The artificial intelligence debate: false starts, real foundations* (ed. S. R. Graubard), pp. 15–43. Cambridge, Mass.: MIT Press.

Elman, J. L. 1991 Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning* **7**, 195–225.

Feigenbaum, E. A. & Feldman, J. (eds) 1963 *Computers and thought.* New York: McGraw-Hill.

Fodor, J. A. 1976 *The language of thought.* Hassocks, Sussex: Harvester Press.

Fodor, J. A. 1987 *Psychosemantics: the problem of meaning in the philosophy of mind.* Cambridge, Mass.: MIT Press.

Fodor, J. A. 1990 *A theory of content, and other essays.* Cambridge, Mass.: MIT Press.

Fodor, J. A. & Pylyshyn, Z. W. 1988 Connectionism and cognitive architecture. *Cognition* **28**, 3–71.

Haugeland, J. 1985 *Artificial intelligence: the very idea.* Cambridge, Mass.: MIT Press.

Haugeland, J. 1994 Mind embodied and embedded. In *Philosophy and artificial intelligence* (ed. L. Haaparanta & S. Heinamaa). *Acta Philosophica Fennica.* (In the press.)

Hendeler, J. A. (ed.) 1989 *Hybrid systems. Connection Sci.* **1**, 227–342.

Hinton, G. E. (ed.) 1991 *Connectionist symbol processing.* Cambridge, Mass.: MIT Press.

Hinton, G. E., Plaut, D. C. & Shallice, T. 1993 Simulating brain damage. *Scient. Am.* **265**, (10), 58–65.

Holland, J. H. 1975 *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence.* Ann Arbor: University of Michigan Press. (Reissued 1991, Cambridge, Mass.: MIT Press.)

Holland, J. H., Holyoak, K. J., Nisbet, R. E. & Thagard, P. R. 1986 *Induction: processes of inference, learning and discovery.* Cambridge, Mass.: MIT Press.

Langton, C. G. 1989 Artificial life. In *Artificial life, SFI Studies in the sciences of complexity, vol. VI (Proc. Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems, September 1987, Los Alamos)* (ed. C. G. Langton), pp. 1–48. Redwood City, California: Addison-Wesley.

McCulloch, W. S. & Pitts, W. H. 1943 A logical calculus of the ideas immanent in nervous activity. *Bull. math. Biophys.* **5**, 115–133.

McGinn, C. 1991 *The problem of consciousness.* Oxford: Basil Blackwell.

Mackworth, A. K. 1973 Interpreting pictures of polyhedral scenes. *Artificial Intelligence* **4**, 121–138.

Maes, P. 1990 Situated agents can have goals. *Robotics and Autonomous Systems* **6**. Also in *Designing autonomous agents* (ed. P. Maes). Cambridge, Mass.: MIT Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Maes, P. (ed.) 1991 *Designing autonomous agents.* Cambridge, Mass.: MIT Press.

Marr, D. C. 1982 *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.

Mataric, M. 1991 Navigating with a rat brain: a neurobiologically inspired model for robot spatial representation. In *From Animals to Animats: Proc. 1st Int. Conf. on Simulation of Adaptive Behavior* (ed. J.-A. Meyer & S. Wilson), pp. 169–175. Cambridge, Mass.: MIT Press.

Meyer, J.-A. & Wilson, S. W. (eds) 1991 *From Animals to Animats: Proc. 1st Int. Conf. on Simulation of Adaptive Behavior.* Cambridge, Mass.: MIT Press.

Miller, G. A., Galanter, E. & Pribram, K. H. 1960. *Plans and the structure of behavior.* New York: Holt.

Millikan, R. G. 1984 *Language, thought and other biological categories.* Cambridge, Mass.: MIT Press.

Millikan, R. G. 1993 *White Queen psychology, and other essays for Alice.* Cambridge, Mass.: MIT Press.

Morris, M. R. 1992 *The good and the true.* Oxford: Clarendon Press.

Nagel, T. 1974 What is it like to be a bat? *Phil. Rev.* **83**, 435–450.

Neisser, U. 1963 The imitation of man by machine. *Science, Wash.* **139**, 193–197.

Neisser, U. 1967 *Cognitive psychology.* New York: Appleton-Century-Crofts.

Newell, A. 1980 Physical symbol systems. *Cognitive Sci.* **4**, 135–183.

Newell, A. 1990 *Unified theories of cognition.* Cambridge, Mass.: Harvard University Press.

Newell, A. & Simon, H. A. 1961 GPS – A program that simulates human thought. In *Lernende Automaten* (ed. H. Billing), pp. 109–124. Munich: Oldenbourg.

Newell, A. & Simon, H. A. 1972 *Human problem solving.* Englewood-Cliffs, N.J.: Prentice-Hall.

Newell, A., Shaw, J. C. & Simon, H. A. 1957 Empirical explorations with the logic theory machine: a case study in heuristics. *Proc. Western Joint Computer Conf.* **15**, 218–239.

Pitts, W. H. & McCulloch, W. S. 1947 How we know universals: the perception of auditory and visual forms. *Bull. math. Biophys.* **9**, 127–147.

Popper, K. R. 1963 *Conjectures and refutations: the growth of scientific knowledge.* London: Routledge.

Port, R. & T. van Gelder (eds) 1994 *Mind as motion: dynamics, behavior and cognition.* Cambridge, Mass.: MIT Press. (In the press.)

Putnam, H. 1960 Minds and machines. In *Dimensions of mind: a symposium* (ed. S. Hook), pp. 148–179. New York: New York University Press.

Putnam, H. 1967 The nature of mental states (or: psychological predicates). In *Art, mind and religion* (ed. W. H. Capitan & D. D. Merrill), pp. 37–48. New York: McGraw-Hill.

Pylyshyn, Z. W. 1984 *Computation and cognition: Toward a foundation for cognitive science.* Cambridge, Mass.: MIT Press.

Reitman, W. R. 1963 Personality as a problem–solving coalition. In *Computer simulation of personality: frontier of psychological research* (ed. S. S. Tomkins & S. Messick), pp. 69–100. New York: Wiley.

Reitman, W. R. 1965 *Cognition and thought: an information–processing approach.* New York: Wiley.

Reitman, W. R., Grove, R. B. & Shoup, R. G. 1964 Argus: an information–processing model of thinking. *Behavioral Sci.* **9**, 270–281.

Rosenblatt, F. 1958 The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–407.

Rumelhart, D. E. & McClelland, J. L. (eds) 1986 *Parallel distributed processing: explorations in the microstructure of cognition.* Cambridge, Mass.: MIT Press.

Searle, J. R. 1980 Minds, brains, and programs. *Behavioral & Brain Sciences* **3**, 417–424.

Searle, J. R. 1990 Is the brain's mind a computer program? *Scient. Am.* **262** (1), 20–25.

Searle, J. R. 1992 *The rediscovery of the mind.* Cambridge, Mass.: MIT Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Selfridge, O. G. 1959 Pandemonium: a paradigm for learning. In *Proc. Symposium on Mechanization of Thought Processes* (ed. D. V. Blake & A. M. Uttley), pp. 511–529. London: HMSO.

Sims, K. 1991 Artificial evolution for computer graphics, *Computer Graphics* **25**, 319–328.

Sloman, A. 1987 Motives, mechanisms and emotions. *Cognition and Emotion* **1**, 217–233.

Sloman, A. 1991 AI, neural networks, neurobiology, architectures and design space. *AISB Q.* **78**, 10–13.

Smolensky, P. 1988 On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**, 1–74.

Stark, R. J. 1993 Connectionist variable binding architectures. D.Phil. thesis, University of Sussex.

Thagard, P. 1988 *Computational philosophy of science.* Cambridge, Mass.: MIT Press.

Thornton, C. (ed.) 1991 *Hybrid models. AISB Q.* **78**, 1–42.

Todd, S. & Latham, W. 1992 *Evolutionary art and computers.* London: Academic Press.

Tomkins, S. S. & Messick, S. (eds) 1962 *Computer simulation of personality: frontier of psychological research.* New York: Wiley.

Turing, A. M. 1936 On computable numbers, with an application to the *Entscheidungsproblem. Proc. Lond. math. Soc.* (2) **42**, 230–265.

Turing, A. M. 1950 Computing machinery and intelligence. *Mind* **59**, 433–460.

Uhr, L. & Vossler, C. 1963 A pattern recogniton program that generates, evaluates and adjusts its own operators. In *Computers and thought* (ed. E. A. Feigenbaum & J. Feldman), pp. 251–268. New York: McGraw-Hill.

Varela, F. J., Thompson, E. & Rosch, E. 1991 *The embodied mind: cognitive science and human experience.* Cambridge, Mass.: MIT Press.

von Uexkull, J. 1957 A stroll through the worlds of animals and men. In *Instinctive behavior: the development of a modern concept* (ed. C. H. Schiller), pp. 5–82. New York: International Universities Press.

Wheeler, M. W. 1994*a* From robots to Rothko: the bringing forth of worlds. In *Perception and understanding: cognition in science and art* (ed. C. Murath). (In preparation.)

Wheeler, M. W. 1994*b* Active perception in meaningful worlds. In *Proc. 4th Int. Conf. on Artificial Life.* Cambridge, Mass.: MIT Press. (In the press.)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES