THE DAWNING OF (MACHINE) INTELLIGENCE

S. G. Shanker

1. *Turing's 'Computational Revolution'*

It was recently announced that 'The Modern World ... began on November 10, 1619' when Descartes foresaw 'the unification of all science' and thence the 'mathematization of the world' [15, p. 3]. The sober historian of mathematics will be forgiven for wishing to push this date back by some two millennia, but perhaps there is some truth to Davis and Hersh's bold thesis insofar as, prior to Descartes, the mathematician was allowed complete freedom to penetrate the mysteries of the universe, but the secrets of his own body and mind were deemed to lie irrevocably beyond the compass of his deductive tool. Like all revolutionary episodes in the history of ideas, Descartes' epiphany was really an act of defiance bordering on hubris. Far from just musing on the promise of analytic geometry, Descartes was challenging this orthodox confinement of the 'science of quantity and space' [see 40]. Nor were the wider implications of Descartes' program (as spelled out in *Treatise of Man* and *Passions of the Soul*) lost on his contemporaries, as the immediate outbreak of the mechanist/vitalist debate makes clear. Humanist prejudices die hard, however, and it was not until late in the nineteenth-century that vitalism became a spent force; now, late in the twentieth, there are many who anticipate the imminent vindication of the latter part of what they regard as Descartes' prophetic vision, as mathematics extends its way into the inner recesses of the mind. The problem in all this, however, as this special volume attests, is that we have yet to reach a clear understanding on the nature of mathematics itself, let alone that branch of mathematics on which this burgeoning 'science of mind' is to be based.

Indeed, we are only just beginning to appreciate the full significance of the fact that mathematics does not operate in a conceptual vacuum. For the manner in which a mathematician interprets his results is as heavily influenced by the scientific trends around him as the latter draw on current mathematical

thought. When there is a shared transformation in both mathematical and scientific outlooks the result can be said to be a paradigm-revolution; but as philosophers of science have all too clearly demonstrated, a convergence in theoretical aims and assumptions, no matter how consistent, is by no means a guarantor of truth. This is especially so when the foundational issues on which the fledgling science rests have yet to be resolved, while those who, eager to get on with the business of theory-construction, hope that future empirical discoveries will remove those fundamental epistemological obstacles that have troubled philosophers of mathematics. Certainly this has been the case with Artificial Intelligence (AI); for aspiring cognitive scientists have made free use of Church's Thesis without appreciating the subtlety of the problems that it embodies [see 61]. The very fact that AI does represent such a melding of interests entails, however, that the philosopher of mathematics cannot begin to address its foundations without broadening his scope immeasurably. For computationalism quite clearly cannot be seen as an isolated mathematical affair; basic issues in the theory of algorithms have been inextricably linked to cybernetic theories of action and psychological theories of learning: largely as the result of a figure who many would see as Descartes' spiritual heir, both in terms of his accomplishments and his breadth of vision.

Certainly Turing stands out, like Descartes, as a revolutionary figure in the field of mathematics in which he was most involved (not to mention the zest for iconoclasm which he demonstrated in such papers as 'Computing Machinery and Intelligence'). The developments which Turing initiated on the basis of this comparison of 'a man in the process of computing a real number to a machine which is only capable of a finite number of conditions' were in many ways foreign to the intentions of his immediate predecessors in recursion theory. Where they were preoccupied with the search for a criterion whereby the class of effectively calculable functions could be demarcated, Turing's version of Church's Thesis served as a transitional impossibility proof whose real significance was immediately regarded as philosophical rather than mathematical. With hindsight one can see how Turing's results were the product of the formalist framework in which they were embedded, which perhaps accounts for the striking promptness with which the epistemological import of 'On Computable Numbers' was grasped. Apart from the conception of mathematical propositions and truth which Turing inherited from Hilbert, however, our concern in this paper will not be with the formalist foundations of Turing's interpretation of his cryp-

tological method for automating a class of mathematical procedures (as a means of demonstrating that machines can follow a species of primitive 'mechanical' rule) [see 61]. Rather, we shall concentrate on the formalist assumptions which underpinned the mechanist implications that Turing sought to graft on to Church's Thesis, and most interestingly, the platonist consequences which this had on Turing's successors. But before we can approach the foundational problems which have thus been absorbed into the fabric of AI, we must first consider briefly the state of the mechanist environment in which Turing found himself, and the manner in which this influenced his thought.

Prior to Turing, mechanists had struggled in vain to clarify the relationship between the neuro-chemical operations of the brain and so-called 'psychic processes'. As late as 1926 Clark Hull was recording in one of his 'Idea Books':

> It has struck me many times of late that the human organism is one of the most extraordinary machines - and yet a machine. And it has struck me more than once that so far as the thinking processes go, a machine could be built which would do every essential thing that the body does (except growth) so far as concerns thinking, etc. And ... to think through the essentials of such a mechanism would probably be the best way of analyzing out the essential requirements of thinking [31, p. 820].

But such a program was more than a century old, and for all the mathematical sophistication of his theories Hull conspicuously failed to advance matters significantly beyond that which had been achieved by the reductionists. The cardinal doctrine of AI is that Turing changed all this: to the point where Marvin Minsky goes so far as to dismiss this kind of 'pre-computational mechanism' as nothing of the kind. According to Minsky, the people who

> considered themselves to be mechanists tended to be something else. I don't know if there's a word for them. There should be - let's say simplists. Striking examples are people like Pavlov and Watson and the whole family of people who believed in conditioning as a basis for learning, the mechanical associationists. Although on the surface they could be considered mechanists because they seem to talk more openly about the mind being a machine, their real trouble is that their image of the machine is precomputational [quoted in 41, p. 71].

It is noteworthy that, from the countless examples of 'precomputational mechanists' available, Minsky should have seized on the one group that, if not formally allied to the Logical Positivists, certainly shared their anti-metaphysical ardour. The implication here is that AI stands diametrically opposed to behaviourism: a point which, as far as scientific attitudes are concerned, is undoubtedly the case. Indeed, what those who would see Hull as one of the fathers of AI overlook is precisely the fact that the 'logic of mind' as this has developed since Turing is above all else 'a metaphysical doctrine' [51, p. 4]. That is, the changes wrought by 'On Computable Numbers' signify more than just a further refinement in the notion of 'machine'; herein lies the impetus for yet another swing in the empiricist/rationalist pendulum, and yet another rehabilitation of transcendental deduction (this time under the guise of calculating the computational properties of mental processes).

Just as introspectivism succeeded the mechanistic psychology that flourished in the middle of the nineteenth-century, so behaviourism marked a sharp reaction to the excesses that had developed in the former's approach to animal psychology. It was equally inevitable that a new school would then repudiate the constrictions imposed by the exclusion of mentalistic notions from scientific explanations of behaviour. Bruner recalls how in the late 1940s there was a growing 'cultural movement to change the image of man from a passive receiver and responder to an active selector and constructor of experience' [9, p. 103]. Nowhere was this more clear than in Bruner's own work in perception: the so-called 'New Look' which postulated the existence of a 'hypothesis generator' which formulates the 'prerecognitional assumptions' that govern perception. But the problem with this 'Judas Eye' was the lack of a theoretical framework in which to probe its operations. Thus Bruner recounts how 'it was not until we were able to look at perception as a genre of "information processing" in the metaphor of a computer that we were able to see the *necessity* its being the result of a prolonged process - for all its phenomenal immediacy' [9, p. 82]. The significance of this account as far as the evolution of AI is concerned is clear: the pressures for a computational model of 'mental processes' had begun to emerge prior to the widespread awareness of Turing's results. As several commentators have pointed out, the *Zeitgeist* was exerting its unseen influence on several fields long before they became aware of their convergent interests. Indeed, the realization that this was the case did not occur until the mid 1950s: at roughly the time when AI was baptized as such.

If there was an organizing principle underlying these isolated

movements, therefore, it lay in the anti-reductionist animus which first found expression in the Gestalt theory of perception (but can also be discerned in the growing signs that behaviourists themselves were becoming increasingly restless during the 1930s with the sweeping proscription of mentalistic concepts). But having said that, the question remains whether AI is as conceptually divorced from behaviourism as the shift in *Weltanschauungen* might suggest. Perhaps the greatest irony in these developments was to be that 'the mind came in on the back of the machine' [48, p. 26]. For the paradigm which was to unite the thitherto disparate forces now banded together under the banner of Cognitive Science was itself the off-spring of the union formed between recursion theory and one of the most fundamental of behaviourist notions. As Minsky indicates in the above passage, the 'computational revolution' was the direct result of Turing's conception of 'mechanical calculability'. In strictly mathematical terms what Turing had proved in 'On Computable Numbers' is that every mechanically calculable function is 'Turning-machine computable'. That is, an 'effective procedure' is an algorithm that can be so encoded (e.g. in binary terms) as to be machine-executable [see 61]. But for AI - as opposed to computer - scientists, Turing proved far more than this: what he really accomplished was to transform machines into a species of rule-following system. And following in Turing's footsteps, McCulloch and Pitts were to do exactly the same thing for the brain [see §2].

The manner in which Turing achieved this feat was by postulating a category of meaningless (sub-) rules which could guide the operations of a machine and/or the brain, thereby providing the rudiments for a new understanding of 'machine' and thence the creation of artificial intelligence. The concept of *machine* had already undergone radical changes during the nineteenth-century; whereas at the beginning of the period it had been confined to the static motions dictated by Newtonian mechanics, it had begun to evolve by the 1870s into the teleological homeostatic system envisaged by Claude Bernard (not to mention the 'logical' or 'reasoning' machines conceived by Babbage and Jevons). There was widespread dissension at the time, however, as to whether machines could ever approximate self-regulating *adaptive* behaviour and thus, whether the body *qua* homeostatic system could indeed be described as a machine. The key word here is *ever*, which of course indicates that the issue was regarded as empirical. The obvious solution would be to 'think through the essentials of such a mechanism' but, in G.H. Lewes' words, 'An automaton that will learn by experience, and adapt

itself to conditions not calculated for in its construction, has yet to be made; till it is made, we must deny that organisms are machines' [38, p. 436]. This is precisely the problem which, as we saw above, was continuing to frustrate mechanists fifty years on; and indeed would have remained beyond the compass of their ambitions had Turing not completed the mathematical transformation of machines. What is all too often overlooked by AI theorists, however, is that by providing the computational means for overcoming the impasse in which mechanism found itself, Turing was committed to the very framework - as defined by its network of assumptions - which had created it!

As important as the Turing-inspired 'computational shift' was for the evolution of AI, no less significant were the classical associationist assumptions which provided the means for the transformation of Turing's 'slave machines' into 'intelligent automatons'. If Turing's major accomplishment in 'On Computable Numbers' was to expose the epistemological premises built into formalism, so his main achievement in the 1940s was to recognize the extent to which this outlook both harmonised with and extended contemporary behaviourist thought. Thus Turing sought to synthesize these disparate theories so as to forge an internal relation between mechanical rules and learning programs. Through their joint service in the Mechanist Thesis each thereby served to validate the other: and the framework from whence each derived. It is to the latter that we must look, therefore, in order to understand not simply the genesis, but more importantly, the presuppositions of AI. For it suggests, not just that the significance of the computational revolution might not be quite so pronounced as the cognitivist assumes, but at an even more fundamental level, that the gulf between pre- and post-computational mechanism is not nearly so great as Minsky contends. The reason why he thinks otherwise would appear to be because his attention is firmly fixed on the contrast between behaviourist and cognitivist conceptions of intentional behaviour. But before accepting the radical disparity postulated here, it will be salutary to confirm the extent to which Turing saw himself as working within a broadly behaviourist framework: as taking the theory a step further by incorporating such 'higher level' activities as chess-playing and theorem-proving into the picture. This will enable us to see that the route leading from Huxley's 'sentient automatons' through Hull's 'learning machines' to Turing's 'learning systems' is far more direct and continuous than is commonly acknowledged.

In order to understand the significance of Turing's contribution to this conceptual evolution it is important to be aware,

first, that the thought-experiment machines portrayed in 'On Computable Numbers' are not credited with cognitive abilities as such; on the contrary, they are specifically referred to as devoid of intelligence. The routines which they execute are described as 'brute force': a reminder not just of the repetitious strategy they employ to solve computational problems but also, that they belong to the intellectual level of the brutes (with all the Cartesian overtones which this carries). In Turing's words, these machines 'should be treated as entirely without intelligence'; but, he continued, 'There are indications ... that it is possible to make the machine display intelligence at the risk of its making occasional serious mistakes' [73, p. 41]. Just as a student has been exposed to 'teachers [who] have been intentionally trying to modify' his behaviour so that 'at the end of the period a large number of standard routines will have been superimposed on the original pattern of his brain', so too 'by applying appropriate interference, mimicking education, we should hope to modify the machine until it could be relied on to produce definite reactions to certain commands' [72, p. 14]. The key to accomplishing this feat lay in the introduction of 'learning programs': self-modifying algorithms that revise their rules in order to improve the range and sophistication of the tasks they can execute, thereby satisfying Lewes' demand by enabling such a system to adapt to conditions not calculated for in its construction.

The ultimate philosophical issue which this argument raises is whether or in what sense such programs can be described as 'learning', and if not how they can best be understood [see 58]. But before this problem can be explored there lies the pressing question of why learning should have assumed such importance in mechanist thought. In mythopaeic terms the automaton only springs to life once it displays the ability to recognize and master its environment (at which point humanist anxieties invariably surface in the form of the creator's loss of control over this now autonomous being). In both physiological and psychological terms the nature of learning was to dominate the mechanist/vitalist debates during the nineteenth-century. And in terms of the history of AI the first and in some ways most potent objection raised against the Mechanist Thesis was voiced nearly a century before the invention of computers. In her Notes on Menabrea's 'Sketch', Ada Lovelace cautioned that Babbage's 'Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we *know how to order it* to perform. It can *follow* analysis; but it has no power of *anticipating* any analytical relations or truths. Its province is to

assist us in mmaking *available* what we are already acquainted with'[1] [5, p. 284]. As he made clear in 'Computing Machinery and Intelligence', the crux of Turing's version of the Mechanist Thesis turns on the very premise which Ada conceded in the above passage. For 'Who can be certain that "original work" that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles' [71, p. 21]. The important point is that, granted that the operations of a machine can be guided by rules (however simple these might be), it is possible to develop programs of sufficient complexity to warrant the attribution of intelligence. It was this argument which was to have so dramatic an effect on mechanist thought. For Turing was to insist that the essence of a learning program is its ability to simulate the creative aspect of human learning [see 73, pp. 122-3]. To serve as a defence of machine intelligence this argument must assume that *learning* 'denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time' [65, p. 28]. On first reading this statement reads as little more than a strained attempt to tailor the concept of learning so as to mesh with the concept of 'mechanical rules'. For if all learning amounted to were the adaptation of something to its environment we should be forced to conclude not just that machines but indeed, all matter was capable of learning. Hence it would seem *prima facie* that Turing only succeeded in subverting the concept of learning in his zeal to reduce it to a level commensurate with the minimal 'cognitive abilities' of his machines. But such an argument fails to do justice to the manner in which AI evolved from the union of mathematical and pre-existing mechanist thought, and the extent to which the latter had come to dominate learning theory. Moreover, it completely ignores the evolution of 'machines' which underpins this outcome and the bearing which this had not just on Turing's thought, but as a result of his influence, on automata theory and thence AI. That is, it obscures the extent to which behaviourist presuppositions were absorbed into the foundations of AI.

This behaviourist orientation is particularly evident in 'Intelligent Machinery', the report which Turing completed for the National Physical Laboratory in the summer of 1948. The purpose of this paper was to defend the claim that self-modifying algorithms can legitimately be described as 'learning' programs. The opening premise recalls Lashley's theory that learning is the result of reflex pathways brought about by conditioning. According to Turing, 'the cortex of the infant is an unorganized

machine, which can be organized by suitable interfering training' [72, p. 16]. By enabling the system to modify its own rules Turing thought he had demonstrated that his machines could in principle simulate the formation of neural reflex arcs that take place during conditioning. The ensuing argument then expands on this notion of 'modification' in terms of the 'Spread of Effect' experiments inspired by Thorndike. From Turing's point-of-view, the most important element in this classic associationist theory is that learning does not involve conscious reflection but rather, reduces to (quantifiable) stimulus-response units.[2] Turning explained that, in so far as 'The training of the human child depends largely on a system of rewards and punishments', 'It is intended that pain stimuli occur when the machine's behaviour is wrong, pleasure stimuli when it is particularly right. With appropriate stimuli on these lines, judiciously operated by the "teacher", one may hope that the "character" will converge towards the one desired, i.e., that wrong behaviour will tend to become rare'[3] [72, p. 16]. By enabling the system to modify its own rules, Turing thought he had demonstrated that his machines could in principle simulate the formation of neural reflex arcs that take place during conditioning. The ensuing argument then expands on this notion of 'modification' in terms of the 'Spread of Effect' experiments inspired by Thorndike. From Turing's point-of-view, the most important element in this classic associationist theory is that learning does not involve conscious reflection but rather, reduces to (quantifiable) stimulus-response units.[2] Turing explained that, in so far as 'The training of the human child depends largely on a system of rewards and punishments', 'It is intended that pain stimuli occur when the machine's behaviour is wrong, pleasure stimuli when it is particularly right. With appropriate stimuli on these lines, judiciously operated by the "teacher", one may hope that the "character" will converge towards
the one desired, i.e., that wrong behaviour will tend to become rare'[3] [72, p. 17].

The concept of *modification* on which Turing placed so much emphasis in his 'learning'-based version of the Mechanist Thesis was thus directly culled from behaviourist writings: it was by employing 'analogues' of pleasure and pain stimuli that he hoped 'to give the desired modification' to a machine's 'character' [72, p. 20]. In 'Intelligent Machinery, A Heretical Theory' he explained that

> Without some ... idea, corresponding to the 'pleasure principle' of the psychologists, it is very difficult to see how to

> proceed. Certainly it would be most natural to introduce some such thing into the machine. I suggest that there should be two keys which can be manipulated by the schoolmaster, and which can represent the ideas of pleasure and pain. At later stages in education the machine would recognize certain other conditions as desirable owing to their having been constantly associated in the past with pleasure, and likewise certain others as undesirable [70, p. 132].

The metaphor would now appear to be twice removed from the meaning of 'learning'; where behaviourists had taken the liberty of depicting habituation as a lower form of learning, Turing went a step further and added the premise that machines display 'behaviour' which can be 'conditioned' by 'analogues of pleasure and pain stimuli'. Whether or not he was aware of Hull's writings, Turing clearly shared his conviction that 'an automaton might be constructed on the analogy of the nervous system which could learn and through experience acquire a considerable degree of intelligence by just coming in contact with an environment'. [31, p. 820]. But that is precisely a consequence of the fact that Turing shared the framework in which Hull approached this issue.

This framework is exemplified (although not inspired by) Thorndike's experiments on the 'learning curve'. Thorndike designed a 'puzzle box' to measure the number of times a cat placed inside would randomly pull on chains and levers to escape. He found that when practice days were plotted against the amount of time required to free itself, a learning curve emerged which fell rapidly at first and then gradually until it approached a horizontal line which signified the point at which the cat had 'mastered the task'. According to Thorndike his results showed how animal learning at its most basic level breaks down into a series of brute repetitions which gradually 'stamp' the correct response into the animal's behaviour pattern by creating 'neuro-causal connections'. Repetition alone, however, does not suffice for the reinforcement of these connections; without the concomitant effects produced by punishment and reward new connections would not be stamped in. But why should such conditioning be greeted as a confirmation of the thesis that learning is nothing more than adaptation (cf. the quotation from Lewes, *supra*)? Foregoing discussion of the associationist premise on which this is based, we can see that the answer rests on a picture of the *continuum of learning*,

> with simple negative adaptation (habituation, or accommodation, and tropisms, which are orientating responses and are known to be mediated by fairly simple physico-chemical means) at one end, and maze-learning, puzzle-box learning ... and ape-learning ... in stages of increasing complexity, leading to human learning at the other end. The conditioned response ... falls somewhere towards the middle of the continuum [19, p. 180].

The crucial point is the idea that learning results from the formation of stimulus-response 'connections' that require a modicum of intelligence.[4] The 'higher' forms of learning are thus distinguished by the complexity of the behaviour acquired through this process, but the cognitive abilities rendered by the atomic associations which form the basis for all levels of learning remain identical and thus provide the rationale for describing what had hitherto been regarded as disparate phenomena as a continuum of learning.

Needless to say, this argument imparted a vital impetus to Turing's version of the Mechanist Thesis; for provided the lowest level of the continuum can be artificially simulated, there is no *a priori* reason why machines should not be capable of ascending this cognitive hierarchy. Moreover, it should be clear from the brief account presented above that this was exactly the theme which Turing exploited in the 1940s in his defence of machine intelligence. In so doing he instituted yet another and what he regarded as the crucial modification to this behaviourist theory. Prior to Turing the mechanist conception of the learning continuum had faced a major obstacle: on this picture the mastery of a concept which forms the mainstay of learning is rendered categorially identical to phylogenetic adaptation. The problem was that while Pavlov and Thorndike's conditioning experiments on dogs and cats marked an advance over Loeb and Jennings' habituation studies on caterpillars and mammalian eggs, it was not at all clear how to ascend higher than this to the upper reaches of the continuum so as to bring the normative practices which characterize learning proper under this conditioning umbrella. It was one thing to stipulate that by virtue of the continuum these neural structures must exist, but quite another to know exactly how they would be configured. Before we consider the post-computational response to this latter issue, however, it is important that we bear in mind some of the fundamental objections to this (associationist) behaviourist theory that have been raised. For the major question that will concern us in the sequel is whether Turing surmounted or

subsumed these problems, and to what extent this can be said to have impinged on AI.

The first thing to notice is that this picture of a *learning continuum* was putatively one in which the higher forms of learning are built up out of simpler components. But in actual fact the framework evolved in the opposite direction; it was only by first postulating that the network of learning concepts can be applied in a diminished state to the descending levels of the continuum that the converse compositional theory could be installed. Hence the theory assumed that the family of concepts tied to learning are only externally related, and can be hived off from the declining orders until the foundational level is reached where learning is a function of reflexive habituation. To stipulate with Jacques Loeb that his caterpillars 'learned where the light was coming from' was thus to say nothing more than that they had acquired knowledge about the light. On this classical empiricist conception 'we suppose that the organism had some specific experience which caused or was in some way related to the change in its knowledge state' [8, p. 13]. With this premise in place the theory could then reintroduce the various cognitive concepts, now hierarchically arranged, at each successive stage on the evolutionary scale. To understand what 'learning' involves at any given stage, therefore, is to know in advance what cognitive abilities are possessed by the organisms in question (e.g. be they 'negative adaptation', 'maze learning', 'ape-learning', or 'human learning'). Otherwise the scientist runs the risk of misinterpreting an experiment by assigning inappropriate knowledge-claims to the organism in question; e.g. of supposing that, 'In the make-believe world of talking animals, Pavlov's dog might say to itself, "The bell was followed by food" and the giant axon of a squid might say, "Irritation of my nerve ending is followed by a hell of a shock"' [8, p. 14].

The dangers manifested by such conclusions is a consequence of the tendency to impute the scientist's knowledge to the organism under study. The problem here is not to clarify how the 'knowledge' acquired by the lower orders should be described; it is to see how very misleading is the assumption that adaptation is coextensive with cognition. Only if one were already committed to the premise that 'experience causes a change in the state of knowledge ... of the organism' [8, p. 14] could it be supposed that the fact that Loeb's caterpillars were attracted to the light signifies a change in their 'knowledge state'. But knowledge is not a state, much less one that is caused by external stimuli, or whose inception and duration can be accurately measured [cf. 78, §82]. Nor is the concept of sensation

internally related to that of cognition in the manner assumed by the associationists. Finally, knowledge can neither be identified with a change in behaviour - for one can behave in a certain way without knowing what one is doing (as conditioning experiments have so abundantly demonstrated) - nor denied on the grounds that no behavioural change was discernible (as those seasoned in dissimulation can attest). On this approach 'we infer someone's knowledge from inputs to him and outputs from him, and we *infer* learning caused by an experience because of before-to-after changes in his inferred knowledge' [8, p. 14]. To be sure, a change in someone's behaviour may be evidence for his having learnt how to φ: but the nature of this evidence is logico-grammatical, not inductive. What someone has learnt does not *cause* them to φ; rather it is their ability to φ that licenses our judgement about what they have learnt. But such judgements are always defeasible: inescapable proof that the relationship between behaviour and learning is not one of equivalence or entailment. Hence the possibility that a subject may behave in an 'appropriate' manner without having learnt how to φ and conversely, have learnt how to φ and yet conceal this fact in their behaviour.

Like the notion of understanding to which it is so intimately connected, learning is a family-resemblance concept which embraces a wide spectrum of activities that are loosely based on the attainment of an ability, not 'cortical connections'. Regardless of any changes that might occur in its neural map (or even, none at all), Thorndike's cats would still be said to have learnt how to escape from the box if they consistently demonstrated such an ability; for that is what the term means. What renders the so-called 'higher forms of learning' more complex is not the compilation of such simple skills but rather, the ability to govern one's actions according to a *rule*. Far from explaining, the continuum picture only serves to undermine the normative foundation on which the latter concept rests. Indeed, the theory could only proceed by assuming the very phenomenon which it had undertaken to explain. For a subject's cognitive abilities are displayed by what they can learn: not the reverse. The source of the confusion operating here stems from the initial move to sever the internal relation which binds concept-learning to understanding. Because of their preoccupation with vitalism the physiological mechanists had misguidedly treated the relation between purposive behaviour and choice as external, and following in their footsteps, the fathers of behaviourism were to do exactly the same thing with the concepts of behaviour and learning. Moreover, they were to commit this conceptual trans-

gression in the very same manner: e.g. by confusing the ability to speak a language with *physiological adaptation*. But the former refers to the mastery of a concept: not the (acquired or unconditioned) response to a stimulus. Hence it demands the abilities to explain, teach, justify, correct, etc. the use of that concept. Without these abilities one is left, not with a 'lower form of learning' but rather, an activity which is categorially divorced from the family of learning concepts.

To be sure, humans do display a spectrum of learning abilities, but these are not 'hierarchically' arranged and *a fortiori*, in no way a reflection of the associationist continuum picture. Rather, this is determined by the complexity of the skills and concepts that the infant, child, adult is capable of mastering. To extrapolate from this the premise that since apes can perform linguistic feats similar to the young human learner there is no reason why simple organisms should not be compared to the human foetus is to misunderstand the sense in which the language-learning ability of apes can be compared to that of children. For it is only in so far as Washoe or Suzy could satisfy the normative criteria which license the description of a child's behaviour as *learning* that such a hypothesis applies. Without this basis of comparison one is left with a repetition of the confusions inspired by Hans the wonder horse, and recent behaviourist experiments in animal learning would only amount to examples of the sophisticated results that the dedicated animal trainer can obtain. But then, conditioning experiments are becoming increasingly out of fashion, largely *because* of the current work in animal learning. Indeed, the AI-scientist can happily point to all of the above criticisms as proof of the insuperable barriers to a behaviourist theory of learning while maintaining that the type of automaton established by Turing is categorially different from these earlier mechanist perspectives. The emphasis as far as Turing's learning machines are concerned is on the system's ability to be guided by and modify its mechanical rules. Thus for the post-computationalist the mind emerges as the set of internal representations *embodied* in the brain and its profile is inferred from the actions brought about by such rules. Admittedly, the success of the theory is still to be deduced from observed behaviour; but the 'connections' now being tested are quasi-normative, not electro-chemical.

For the radical behaviourists, learning was seen as the physico-chemical states that cause a subject to φ; on Turing's neo-behaviourist interpretation we return to mental states - now characterized as stages in a program - that cause a subject to φ. The basic premise that learning and behaviour are *causally*

determined and hence a fit subject for scientific explanation was thus retained by Turing. Unlike Turing, cognitivists are aware of and anxious to overcome the philosophical dangers inherent in this form of strict reducibility. Their goal is not to disregard or eliminate the cluster of normative concepts outlined above characterizing learning at the level of ordinary language but rather, to explain the manner in which the computational configurations divined by Turing guide an agent (or system's) actions at the sub-behavioural level. But in Turing-like fashion, they could only implement this strategy by relying on one of the most fundamental of Hilbert's assumptions. It is certainly no coincidence that, despite the key role which learning theory played, this early phase in the development of AI should have been so completely dominated by mathematical logicians rather than psychologists. Nor should it come as a surprise that the foundations of AI should be so closely bound up with one of the central issues in the foundations of mathematics. For at the heart of the cognitive enterprise lies the premise that, in so far as models can be mapped onto physical structures, it follows that the former can be said to be *realized* in the latter. Few present-day cognitivists will be aware, however, of the formalist origins of this idea, or the philosophical problems which it presents. Our task in the next two sections will be to correct this situation.

2. *'A Logical Calculus of the Ideas Immanent in Nervous Activity'*

To understand both why Turing saw his 'learning programs' as a behaviourist addendum and also why his version of Church's Thesis had such an immediate impact on psychology, it is necessary to grasp the impasse in which behaviourism found itself. Purposive behaviour had been defined as a consequence of neuro-physiological adaptation: the 'stamping in' of 'cortical reflexes'. This meant that 'learning' could be quantified in terms of the repetition of stimuli required to reinforce a given response. But Thorndike had shown that frequency of repetition alone was insufficient to bring about the formation of neural connections: it must be supplemented by punishment and reward, pleasant or uncomfortable sensations that the organism would associate with the stimulus in question. The problem was that such notions as pleasure or discomfort were worryingly subjective, and it was not at all clear how to define them other than as those physical states which an organism tended to seek or avoid. But this only served to raise a problem of a different order, for now it was manifest that the internal state of the

organism was crucial to the forging of stimulus-response connections but not at all clear how behaviourism could assimilate this factor, having eschewed the explanatory force of 'unobservables' as its guiding principle. Hence a yawning gulf appeared between the physiological needs of an organism and the reinforcement of its neural accommodation.[5] To be sure, some behaviourists sought to circumvent this problem by postulating 'intervening variables' to mediate between causes and effects (e.g. Hull's 'thirst drive'). But these were seen as heuristic devices, and thus, eliminable. In no way were they intended to sanction the reintroduction of 'mental states' as a means of bridging the gap between the effect that an organism's internal structure had on its responses to its environment.

What was needed was some means of reconciling the behaviourist emphasis on observational techniques with a method of explaining *how* an organism interacted with its environment. And it was precisely this lacuna which automata theory sought to fill. Following the recursive route mapped out in 'On Computable Numbers', 'A Logical Calculus of the Ideas Immanent in Nervous Activity' by Warren McCulloch and Walter Pitts must be regarded as one of the *Gedankenbausteinen* of AI, even though its 'bottom-up' approach was almost immediately repudiated by the subject's founding fathers. Following Mountcastle's discovery that neurons are grouped vertically in cortical columns and, more importantly, the publication of Newell and Simon's 'Logic Theorist', the influence of McCulloch and Pitts' theory of formal neural nets began to wane in AI circles. For these reasons its significance is now deemed to be largely historical: an example of early insights and miscues, and a catalyst for subsequent development in neurophysiology as well as AI. From a philosophical point of view, however, its bearing on the subsequent evolution of AI is far more significant. What matters to us are the assumptions which survived the demise of the theory of axiomatized neural nets.[6] In this respect McCulloch and Pitts were framework builders, ranking on an equal footing with Turing and Shannon. Where they were so important was the manner in which they seized on the mechanist premises implicit in Turing's version of Church's Thesis. From a passing remark following a lecture by von Neumann it is clear that 'A Logical Calculus of the Ideas Immanent in Nervous Activity' was directly inspired by 'On Computable Numbers'.[7] That is not to say that the mechanist import of Turing's thesis would not have been grasped without their work; on the contrary, the epistemological pressures built into Turing's conceptual framework would have seen to that [see 61]. Where McCulloch and Pitts were so important was in their

generalization of Turing's results: in taking them outside the narrow parameters of recursion theory and applying them to the neurophysiological study of purposive behaviour and the psychological investigation of concept-acquisition, thereby bestowing a tremendous impetus to the burgeoning field of automata theory.

To do this they literally took over Turing's picture of mechanical comuting machines and applied it to the brain. For McCulloch and Pitts' 'neural nets' are not analogues, they are literally a species of Turing machines: viz. 'bivalent neural systems' that compute recursive functions. Hence they are rule-guided mechanisms of a minimal cognitive level. As in the case of Turing machines, it is the complexity of the neural net, with the number and configuration of the neurons in a net directly proportionate to the mental computational task involved, which delivers intelligence.[8] For the purposes of the theory the variety of neurons is ignored; all that need be known about 'idealized' neurons is that they consist of input synapses, one output axon, and an unanalyzed cell body. Each neuron has a firing threshold: the critical level of received impulses which triggers the firing of an ion impulse. Inputs and axons can be of two kinds - positive (excitatory) and negative (inhibitory) - and the firing threshold is the sum of these inputs. The theory makes the further assumption that the neurons in a net all share a uniform response time. A neuron fires an impulse along its axons at time $t+1$ if and only if the sum of the inputs reaches the firing threshold. The operations of each neuron can thus be represented as a function whose arguments are inputs $x_1, \ldots x_n$ at t and whose value is the output y at $t+1$; i.e.

$$\varphi(x_1(t)\ldots x_n(t))=y(t+1).$$

The 'black box' of the cell body corresponds to the function φ. How it might actually be constituted is of no concern to the theory: only the nature of the rules correlating arguments with values that regulates its activities (*infra*).

A neural net is defined as a collection of these neurons, each operating on the identical time scale, and connected to each other by (possible branching) axon outputs. Hence the total output of the network ($z(t+1)$) can be defined as a function at cell c_i of inputs $x_1 \ldots x_n$ at time $t+1$:

$$\varphi_i(x_1(t)\ldots x_n(t))=yi(t+1)=z(t+1)$$

The neural net is thus seen as a function whose arguments are the initial values of input fibres and output states, and whose value is the output state at $t+1$. Since (by the second axiom of the theory) the activity of a neuron is an 'all-or-none' affair (it either does or does not fire) the inputs and states of a neural

network can be represented in binary terms (0,1). In the case of a basic single-cell net consisting of two input fibres and one output, where the cell fires if and only if it receives an impulse from both inputs (simultaneously, by the first axiom), then (where 0 = does not fire and 1 = fires) we can formulate the following table:

x_1	x_2	y_1
0	0	0
0	1	0
1	0	0
1	1	1

But what we have here is simply the truth table for '&'. Likewise, the truth-table for 'v' (000, 011, 101, 111) could be used to represent a single-cell net which fires if either of two input fibres fire. The role of inhibitory impulses must also be borne in mind. For example, the table

x_1	x_2	y_1
0	0	0
0	1	0
1	0	1
1	1	0

presents the 'idealized neural analogue' for 'p&-q'. And since it is possible to build up any compound truth-function from these elementary logical constants it follows that there is no bound to the higher neural circuits that can be constructed.

The function-theoretic operations of such complex nets can be described by tables which map the set S of possible input fibres ($=\{s_1 \ldots s_n\}$) to the set Q of possible axon states ($=\{q_1 \ldots q_n\}$, where both $=2_n$). Since at any time t each of these can be either firing or not firing (0 or 1), a simple net consisting of three input fibres (x_1,x_2,x_3) and two output states (y_1,y_2) can be represented as:

	x_1	x_2	x_3		y_1	y_2
s_1	0	0	0	q_1	0	0
s_2	0	0	1	q_2	0	1
s_3	0	1	1	q_3	1	0
.	0	1	0	q_4	1	1
.	1	0	0			
.	1	0	1			
.	1	1	0			
s_8	1	1	1			

But a neural net does not simply map inputs onto outputs; for the existing axon states of the net play a key role in the transformations which the net undergoes. A suitable table, therefore, will be one which maps inputs and axon states onto new axon states. In the following table we adopt the convention that the system receives its inputs ($s_1 \ldots s_8$) serially. Note that a change in input does not entail a change in state; e.g. when the system begins from it remains in state q_1 through inputs s_5–s_7.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
q_1	q_2	q_4	q_1	q_3	q_1	q_1	q_2	q_3
q_2	q_3	q_1	q_3	q_4	q_2	q_2	q_2	q_1
q_3	q_4	q_3	q_1	q_3	q_2	q_1	q_4	q_2
q_4	q_1	q_2	q_2	q_1	q_3	q_4	q_1	q_3

But this is exactly the same kind of table as can be used to map out Turing machines. Hence McCulloch and Pitts concluded that it must be possible to realize a Turing machine in a neural net. For a neural net is no less an automaton than a Turing machine, in so far as each embodies a set of recursive rules. By axiomatizing an 'idealized' neuronal structure they had sought to show, in McCulloch's words, 'that brains were Turing machines, and that any Turing machine could be made out of neurons' [43, p. 155].

A Turing machine is essentially a finite automaton which scans a potentially infinite tape. Obviously, the absence of the latter has a significant bearing on the scope of the functions that can be computed in a neural net. But McCulloch and Pitts explained that

> every net, if furnished with a tape, scanners connected to afferents, and suitable efferents to perform the necessary motor-operations, can compute only such numbers as can a Turing machine [and] that each of the latter numbers can be computed by such a net. ... This is of interest as affording a psychological justification of the Turing definition of computability and its equivalents, Church's lambda-definability and Kleene's primitive recursiveness: If any number can be computed by an organism, it is computable by these definitions, and conversely [44, p. 129].

The important point here is that, disregarding the significance of the tape and scanning/printing device in Turing's argument, both neural nets and Turing machines are instantiations of finite

automatons. The latter are conceived as 'black boxes' which receive inputs, have a finite number of internal states, and emit a finite number of outputs which depend on the automaton's internal state. In mathematical terms, the automaton A which ranges over the set of inputs S can be represented by the quadruple $(Q,\Phi, q_i, 0)$, where Q = the set of internal sets,Φ= the function Φ: $Q \times S \rightarrow Q$ (which determines the next state of A on the basis of the system's present state and the input), the initial state q_i (where $q_i \in Q$), and the set of final states 0 (where $0 \in Q$). The nature of these 'internal states' will depend on the type of automaton under consideration. For example, the internal states of an artificial automaton might be the position of its cogs or levers; and for a natural automaton, it might be the state of excitation of a group of neurons. To refer to internal states as 'black boxes' is simply to credit them with the same attributes as the classical metaphysicians assigned to 'properties' [see 68, p. 63]. All we know about them is that they are the states an automaton is in when it executes its instructions, and that it is the combination of A's present state and the inputs it receives that determine both what act A will perform and what state it will next be in. In the present case the internal states will be *n*-tuples of the firings and non-firings of the neurons in a net. But the internal state must not be identified with that - idealized - physical structure. For given the preceding definition of automatons *qua* formal systems, it follows that multiple automatons can be in the same internal state. Hence the above represents a 'neural realization' of an internal state, which can only be understood as such in virtue of the function-theoretic role which it performs *vis-à-vis* inputs, outputs, and other states in a formal automaton.

A Turing machine *qua* finite automaton can thus be represented as a pair of functions which map the cartesian product of internal states and inputs onto altered states and operations. I.e.

$$M: Q \times S \rightarrow Q$$
$$N: Q \times S \rightarrow 0 \cup Q$$

(where 0 = {L,R,N}. If the output is 'stop' Z halts; otherwise the output of A causes 0 to print i and move one square left, right, or no squares at all on the tape.)

These operations can be written as a set of quintuples of the form

$$q^i s^j s^k R q_l,$$

where this quintuple is said to embody the rules:

> When A is in state q_i and D scans s_j, D prints s_k, moves one square to the right, and A goes to state q_l.

That is, the quintuple embodies two distinct kinds of rule: first, to do something (e.g. print i) and then, to change the internal state. But, as we have already seen, McCulloch and Pitts' neural nets can also be written as

M: $Q \times S \rightarrow Q$

N: $Q' \rightarrow 0$ (where Q' is the subset of ultimate axonal outputs of Q).

Hence it too can be written as a quintuple; e.g.

$q_i s_j q_l q'_m o_z$,

which embodies the rules that when the neural net is in state q_i receiving input s_j it changes to state q_l and when the ultimate axon outputs are in state q'_m the net emits the final output o_z.

It is natural to assume that the term 'neural net' was intended to refer to a unique physical structure in the brain (comparable e.g. to the neuron aggregates or 'nuclei' mapped out in topographical atlases of the brain). But neural nets, McCulloch repeatedly explained, *are* Turing machines. This is an important point, for there is a corresponding tendency to identify Turing machines with the *Gedankenmaschin* described by Turing (e.g. scanner, printer etc.). But 'Turing machine' is an abstraction: it refers, not to any of the physical models of Turing's argument, but to the 'network of rules' which could just as well be executed by a neural net as by the 'automatic calculating engines' envisaged by Turing. What then can we conclude from the fact that one and the same function-table can be realized by both a neural net and a Turing machine? Clearly nothing about the internal constitution of either. After all, both have been ruled 'black boxes' *ab initio*. All that this tells us is that the neural net and Turing machine (or any of the other possible realizations of that state table) all exhibit the same 'behaviour' in response to the same stimuli. Moreover, the theory makes no claims to providing a 'complete' description of the system (e.g. McCulloch and Pitts deliberately ignored the influence of glial cells); hence the neural net only aspires to furnish a partial description of a neural system. Which means that, not only can myriad systems satisfy the same table, but different aspects of the same system can be realized by different state tables. Most important of all, the argument is not confined to neural structures; the reactions of a simple organism, the operations of Bruner's 'Judas eye', or of a digital computer, could all be possible realizations, and the possible outputs might be ion impulses, genetic mutations, perceptions, or mechanical computations.

It would be pointless to attack this argument on the grounds of its neurosphysiological naivety. As Arbib points out, the

theory can only be obtained 'at the cost of drastic simplifications'[9] [1, p. 7]. It assumes complete synchronization of all the neurons, fixes the threshold of each neuron sempiternally, and it ignores the effects of other factors (e.g. hormones and chemicals) or the possible role of the glial cells in neural activity. Moreover, 'A Logical Calculus of the Ideas Immanent in Nervous Activity' is far from being a model of perspicuity or even internal consistency. But the early champions of the theory were fully aware of and unperturbed by these shortcomings [see 74]. For their attention was fixed elsewhere; overlooking the neurophysiological distortions which ensue on such an axiomatization, the point was to see how one could apply Turing's argument to generate a recursive model of 'purposive behaviour' as mechanistically conceived, thereby supplying the foundation for the 'non-causal non-contingent' nature of purposive behaviour which eluded the cyberneticians and thence the computationalist version of the behaviourist search for a theory of learning/adaptation. McCulloch and Pitts explained at the outset of 'A Logical Calculus of the Ideas Immanent in Nervous Activity' that learning is to be seen in terms of activities which 'have altered the net permanently'; it is an 'enduring change' to the net 'which can survive sleep, anaesthesia, convulsions and coma' [44, p. 117]. (E.g. a 'reverberating loop' remains inactive until one of the inputs fires; thereafter it fires whenever any of the inputs fire.) Hence the argument offered a solution to the behaviourist problem of mind which managed to provide a causal account of the mechanics of learning while retaining the normative element that characterizes learning proper. And it accomplished this feat by focussing on the computational structure of the atoms furnished by the reductionist account without committing the fallacy of seeking to collapse learning into these physical elements.

The argument presented the obverse side to the picture of human computation which Turing had used to explicate the notion of mechanical calculability in 'On Computable Numbers', thereby cementing in McCulloch and Pitts' eyes the epistemological resolution of Church's Thesis presented by Turing. Turing had argued for the artificial simulation of the mental states that occur in computing; McCulloch and Pitts offered in turn a mechanical theory of those mental states in virtue of the very fact that they could indeed, as Turing had postulated, be artificially simulated. Moreover, by interpreting the mind in these computational terms, McCulloch and Pitts hoped to base learning on the activities of the brain without succumbing to the reductionist fallacy that fostered behaviourist attempts to treat internal states as 'logical fictions'. Purposive behaviour could be

explained not as the sum total of the causal operations that take place in the brain but rather, as the consequence of being guided by the recursive rules embodied in the neural nets governing those activities. Hence learning could be seen as the result of a neurophysiological adaptability whose configuration and transformations could be inferred from overt behaviour. Following the tradition of nineteenth-century mechanism the picture had remained thoroughly Darwinian: the 'internal states' regulating the adaptation of an organism were the product of an evolutionary process designed to sustain homeostasis. In Ashby's words, 'learning usually changes in behaviour from a less to a more beneficial, i.e. self-promoting form'. Thus, 'when the nervous system learns, its behaviour changes for the better', and 'no use of any "vital" property or tendency will be made, and no *Deus ex machina* will be invoked' in computational explanations of 'learning'; the 'sole reason admitted for the behaviour of any part will be of the form that its own state and the condition of its immediate surroundings led, in accordance with the usual laws of matter, to the observed behaviour'[10] [4, pp. 3–4, 7–8].

In *Design for a Brain* Ashby presented a series of 'idealized forms' of neural mechanism which, while they may not correspond to any actual cell assemblies, should nonetheless enable us to grasp the nature of the governing principle which causes the 'organism as machine' to undergo changes which will ensure 'better adaptation'. Not surprisingly, the (cybernetic) principle in question turns out to be that the purposiveness displayed by an organism must be seen as the actions of a 'stable system [which] has the property that if displaced from a state of equilibrium and released, the subsequent movement is so matched to the initial displacement that the system is brought back to the state of equilibrium' [4, p. 54]. But then, 'Once it is appreciated that feedback can be used to correct any deviation we like, it is easy to understand that there is no limit to the complexity of goal-seeking behaviour which may occur in machines quite devoid of any "vital" factor [4, p. 55]. That is, the argument applies to any automaton, be it natural or artificial, in as much as the principles governing 'learning' are:

(1) Each mechanism is 'adapted' to its end.

(2) Its end is the maintenance of the values of some essential variables within physiological limits.

(3) Almost all the behaviour of an animal's vegetative system is due to such mechanisms [4, p. 58].

Hence it must be possible, as Turing had anticipated, to construct an 'artificial brain' which is capable of exhibiting the

same adaptability within pre-established parameters as characterizes human learning systems. For '*machines with feedback are not subject to the oft-repeated dictum that machines must act blindly and cannot correct their errors*. Such a statement is true of machines without feedback, but not of machines in general' [4; p. 55].

In the late 1940s and early 1950s this argument dominated mechanist thought in biology, psychology, neurophysiology, and mathematical logic. Several symposia were held to explore the cybernetic interrelatedness of these (and other) fields [see 62]. Predominant throughout this formative pre-AI period was the abstract notion of an automaton which, by its very nature *qua* formal system, could be realized in myriad natural and/or artificial systems (*infra*). Even those who were shortly to reject McCulloch and Pitts' bottom-up approach were to remain faithful to this picture; for the immaterial, non-spatial mind is the archetypal 'black box'. Mechanism may have been no closer to analyzing its internal constitution, but they hoped at least to explain its operations in the above function-theoretic terms, where the mind's inputs are neurophysiological (viz. sensations), its internal states are (as yet) a mystery, and its outputs are those overt behavioural events which the mind causes (e.g. perception). The theory of formal neural nets laid no claims to understanding the 'stuff' of which the mind is composed, therefore, or exactly how it causes bodily reactions; it sought only to disclose the laws governing its operations. But such a theory marked a significant advance in the evolution of mechanism. For it was no longer committed to a naive materialist reduction of mind to brain. The explanation of 'mental processes' on this approach are neither neurophysiological nor bio-chemical; rather, they are computational. The argument thus marked, not just a methodological revolution, but correspondingly, the transformation of organisms from passive into active agents. For on this theory the same stimulus can elicit different responses, depending on the system's internal state.

The table outlined above also shows how a different state would have been produced had the system in a state q_i been presented with a different stimulus. Since the relations between stimuli, internal states and responses can be functionally computed, and the same state table can be realized by any number of automatons, we can now see why it follows on this theory that the operations of the mind can be mechanically simulated and *a fortiori*, that artificial finite automatons can display cognitive abilities. For what is the latter but another way of depicting input-output behaviour (the functions which an automaton com-

putes)? To take but one example, the ability to speak a language can be measured by the number of sentences to which a speaker can respond appropriately, can be mechanically simulated, and thus, artificially realized. The effect which this argument has on Turing's learning thesis is immediately evident. On first reading, Turing's presentation of 'learning programs' has the appearance of a doubly strained metaphor. But McCulloch and Pitts' theory in effect removed the inverted commas from Turing's argument. For it reduced the explanation of behaviour to a level where the need for such commas are no longer relevant in so far as 'pleasure' and 'pain' stimuli are themselves the target for information-theoretic analysis. That is, they are merely those signals which upset the stability of a system, thereby activating the feedback mechanisms which restore an organism's equilibrium. Thus it is that repeated exposure to such conditioning results in the reinforcement of new cortical reflexes designed to adapt the organism to its environment, and the real comparison between human and machine learning comes down to the modification of neural nets versus that of 'Turing machine configurations'.

Far from signifying the overthrowal, this may indeed begin to look like the refinement of behaviourism which Turing envisaged. For all the key notions seem *prima facie* to remain intact in the explanation of purposive behaviour in terms of stimulus-response relations. Certainly there is no evidence in the works of this period that the scientists involved were intent on or even aware of the fact that they were undermining the behaviourist *Weltanschauung* which had recently dominated mechanist thought. It would, however, be misleading to construe this as nothing more than a species of or as marking the transition from behaviourism to neo-behaviourism. It has been argued that, since a convincing reduction of psychological to physiological events has yet to be offered, 'theses about behaviorism remain important. Psychological concepts, complex skills, and, in a still more traditional terminology, mental events as occurring at least in other persons and other animals can be known only from behavioristic evidence [67, p. 284]. On this reading, the transcendental deductions whereby the operations of the mind are inferred are transformed into hypothetico-deductions, and the cognitivist theory of mind *qua* system of recursive rules is not so much a departure from as a supplementation to behaviourism. But apart from the fact that this ignores the psychophysiological origins of behaviourist efforts to explain the mechanics of learning - and the consequent attempt to expand the notion of 'behaviour' so as to encompass neurophysiological events[11] - it also distorts the changes which the concept of 'stimulus' was to

undergo. For on the classic stimulus-response account stimuli are those external events which impinge on the passive organism/agent and to which it responds. With the shift to an information-processing format, however, stimuli themselves become those 'messages' which an organism actively seeks out [see 83, p. 88]. Finally, such an argument overlooks the shift from strict to empirical reductionism which inspires computationalism, and runs the risk of treating the cognitivist debate between bottom-up versus top-down approaches as categorial rather than methodological.

The crux of the computationalist theory is that both strategies should in principle meet in the middle; all that is at stake is the question of how one best proceeds in the initial stages.[12] Nor is it a minor point that computationalism should have been committed to this premise; not only did the evolution of AI - from Turing, Shannon, McCulloch and Pitts onwards - predispose the theory to move in this direction: it was compelled to do so in order to exploit Turing's computability results for a theory of the 'learning continuum' reformulated so as to base purposive behaviour on neural nets, rendering the complexity of the former a function of the latter. One of the most important problems - from a computationalist point of view at any rate - with labelling the cognitive revolution 'neo-behaviourist' is that this represents an attempt to reap the explanatory rewards of the former without embracing its mentalistic underpinnings. It is not enough to argue that behaviourism suffered from this glaring lack of a 'theory of internal states' and leave the matter at that: what the behaviourist seeks to coopt here lies *au fond* on a different level from the overt acts to which he longs to confine psychology. To be sure, the fact that they are function-theoretic - as opposed to functionalist - states demands the presence of a suitably endowed mechanism to execute this new species of embodied rules. But this is precisely where the neo-behaviourist interpretation comes unstuck, in so far as the impetus for the theory of mental processes is neither reductionist nor materialist. The thrust of Turing's thesis may well be that from the most elementary of effectively calculable algorithms complex recursive functions can be built up, but that does not entail that psychological predicates can be reduced to the neurological mechanisms underlying those embodied rules and representations.

The essence of the bottom-down/top-up debate is whether it is more profitable to begin with the complex functions and work our way down from them to the sub-tasks in the algorithms, or whether one should begin with the axiomatized neural nets and seek to compose the higher functions from these computational

elements. The former approach may seem the more immediately inviting, but the latter boasts the more direct confirmability (as e.g. in Lettvin, Maturana, McCulloch and Pitts' 'What the Frog's Eye Tells the Frog's Brain'). There is a tendency, however, to suppose that this conflict is really that between neuroscience and cognitive psychology; that is, between pre-and post-computational mechanism. The one studies the activities of the neuron, the other the system of rules executed by the mind. Hence one is asked

> to imagine the task of trying to understand what [a computer] program is doing (or attempting) in terms of a moment-by-moment listing of the electrical charges on all the thousands of transistors. ... Imagine that we had similar information about the physiological states of the *twelve billion neurons* in the human brain, each with up to *five thousand synapses*. ... This vast amount of information and its fantastic complexity would utterly dumbfound us; we could not hope to begin creating much order out of such vast quantities of particulate information. Rather, we would need some very powerful theories or ideas about how the particulate information was to be organized into a hierarchy of higher-level concepts referring to structure and function. ... Many psychologists feel that their task is to describe the functional program of the brain at the level of flow-charting information-processing mechanisms. What is important is the logical system of interacting parts - the model - and not the specific details of the machinery that might actually embody it in the nervous system [8, p. 476].

But the bottom-up approach alluded to here was certainly not the conception advanced by McCulloch and Pitts; they were not suggesting that any knowledge of the hardware would help one to grasp the recursive structure of the software. On the contrary, the whole point of the 'idealized' nature of their theory was to mitigate against such a reading. Rather, their point was that one seeks to understand the computational mechanics of mental processes by disclosing the simplest recursive components embodied in the nets. For example, in 'How We Know Universals' McCulloch and Pitts undertook to explain how 'Numerous nets, embodied in special nervous structures, serve to classify information according to useful common characters' [43, p. 46]. The recognition of universals is a result of a process of template matching. On this model a pattern is stored in a neural net and incoming signals undergo a series of transformations which are inverse to the stored pattern. In order to recognize a universal (i.e. independent of particulars) the incoming signals

are so transformed until a match occurs, which triggers that action which indicates what, on the molar level, we describe as the recognition of the universal in question.

Unless one is careful to distinguish between pre- and post-computational mechanism, therefore, there is a pronounced danger of confusing the bottom-up approach to the cognitive science of mind with central-state materialism. Nevertheless, the 'neo-behaviourist' interpretation does draw attention to the point emphasized in the opening section that, barring those isolated psychologists consciously in search of the 'New Look' in theories of perception, the key protagonists of automata studies seemed fully at ease in a behaviourist framework. This confirms the significance of the conceptual environment in which computationalism was spawned and as a result, the behaviourist presuppositions that were absorbed into this new theoretical setting: in particular, the continuum picture of learning in terms of neurophysiological adaptation which seemed to harmonize perfectly with the mechanically calculable functions discovered by Turing. Hence computationalism was more than just a symptom of an emerging *Zeitgeist*: it was the natural outcome of two seemingly disparate movements initiated a century before in the philosophies of nature and mathematics. But whether this development resulted form the realization of a pre-established harmony between natural and artificial formal systems or marked the synthesis of fundamental misapprehensions about the normativity of mathematics which, by its very union, only served to entrench one another, is the ultimate philosophical issue which we must now address. There are substantial grounds for questioning two of the key premises operating here: viz. that rules can be physically embodied [see 61], and that internal representations can be said to guide purposive behaviour; our present task, however, is to consider the concept of 'models' whereby these two themes were joined.

3. *The Interpretation of Formal Systems*

One of the main intents of the preceding sections has been to clarify how the birth of AI was the consequence of what in retrospect can be seen to have been more than just the fortuitous union of heterogeneous developments in the philosophies of nature and mathematics. For the former had embarked on a search for the mechanics of adaptation; the latter for the nature of mechanical procedures. It is thus little wonder that they should have proved to be tailor-made for one another. Without

the wider scope of effective procedures there would have been no reason to query the conventionalist terms of Church's Thesis; and without the effectively calculable functions furnished by recursion theory mechanists would likely have remained tethered to a crude empiricist attempt to exclude teleological considerations from the explanation of behaviour (or at least disguise their presence - e.g. by redesignating them as 'teleonomical'). As a result of Turing's 'computational revolution', however, McCulloch and Pitts could openly set out to explore the purposive character of neurophysiological adaptation without fear of subsiding - or being so construed - into vitalist metaphysics. To accomplish this mechanist reorientation they distinguished between those rules which the scientist employs (i.e. the hypotheses formulated in descriptive explanations) and those by which the system under study can be said to be guided. This demarcation rested on an abstract model-theoretic distinction between the *satisfaction* and the *embodiment* of a formal system/automaton. Significantly, their conceptual framework was methodologically neutral as far as the bottom-up/top-down dispute was concerned. Hence, when Newell and Simon repudiated McCulloch and Pitts' approach they nonetheless coopted the formalist premise which underpins the theory of formal neural networks, thereby enabling them to base their putative simulations of 'general problem-solving techniques' on the distinction between those rules which the AI scientist employs in his cognitive hypotheses and those which the mind/computer can be said to follow or by which it is guided.

There are two principle reasons why it is so important for the philosopher of AI to scrutinize these foundations of the subject before embarking on either a critique or contribution to its present theoretical state. The first is to confirm that, in so far as AI is the beneficiary of theories that are themselves plagued by philosophical controversy, there are solid grounds for proceeding with caution whatever one's scientistic predilections. One liability to which critics and advocates alike are prone is the pursuit of tangential and at times inconsequential issues. Popular questions such as whether computers can think or thinkers compute, or even more subtle inquiries into e.g. possible violations of the Homunculus fallacy, seem to offer both a more tractable and *prima facie* more relevant matter than obscure deliberations on the epistemological significance of Turing's mechanical version of Church's Thesis, or the platonist ramifications of McCulloch and Pitts' subsequent application of his results to the general field of automata studies. But without a grasp of the issues involved here one will be led to overlook

the implications and complexity of the problems contained in the foundations of AI; and most important, the nature of the philosophical challenge posed by AI (which is nothing less than the question whether empirical results can be used to resolve *philosophical* problems [see 62]). As is invariably the case with foundational studies, our concern here is with an assumption which demands close attention if one is to grasp the origins of the ensuing sceptical dilemma whose very presence establishes its gravity. In the case of AI, long before one can debate the merits of the 'Turning Test' it is necessary first to clarify what it means to say that the function-theoretic rules governing the operations (i.e. constituting) an automaton can be physically *embodied*.

Frequent suggestions to the contrary, it is not that neural nets are surrogate agents following those rules, which would indeed invite the necessary rejoinder that neurons cannot be credited with the cognitive abilities that apply to rule-following subjects [cf. 83, p. 14]. The matter is not quite so straightforward, however, when one shifts to the cognitive thesis that if anything is to be credited with the ability to follow rules, it is the mind *qua* automaton. To confuse the latter with neural assemblies would be to fall victim to central-state materialism, whose demise is one of the prime objectives of cognitive theories. Yet neither does the cognitivist wish to reduce the mind to its input-output patterns; for, *contra* behaviourism, the mind is a faculty whose inner states bear heavily on behaviour. The purport of the theory, therefore, is that the mind is somehow guided by recursive rules (programs) which are neurophysiologically embodied. But even this way of presenting the theory is misleading, for it immediately invites the request for further clarification of *what* exactly is guided. Admittedly, some cognitivists are tempted to fall back at this point on the last line of defence afforded by the 'black box', but it is crucial to understand that, from the orthodox computationalist point of view, any metaphysical conundrums on this score are the product of an ontological presupposition which is *au fond* misconceived. For the notion of an immaterial, non-spatial substance causing the body's actions is a throwback to the futile attempts of both empiricists and rationalists to explain the interaction between sensations and behaviour without the requisite mathematical tools.

With the advent of the computational revolution all need for a reified mind disappears, and in its place we can speak of the activities of an automaton being 'minded':

> the logic of mind is similar to the system of rules that

> governs the operations of a digital computer and hence it is correct to say that a mind is 'machine-like'. ... It is tempting to identify mind with these rules, although I prefer not to encourage the tendency to hypostatize which is often aroused by use of names such as 'the mind'. Instead, I will say that a necessary condition for a being *to have a mind* or *to be minded* is that its behaviour be guided by rules of a certain sort [51, pp. 2,9].

The notion of an immaterial mind thus turns out to be a logical fiction whose role was rendered otiose by the discovery of the systems of recursive rules that regulate the behaviour of natural and artificial automata. It might, however, appear that this theory still does not escape the charge of violating the Homunculus fallacy if it must still be possible for 'agents' (in a now extended sense of the term) to harbour faculties that follow rules of which only the cognitive scientist is aware. Furthermore, there is - from the mechanist standpoint - the danger that a preoccupation with language acquisition might promote an anthropomorphic bias which can only lead away from the continuum picture of learning/adaptation. Certainly this is one direction in which the theory has proceeded; particularly in the neo-rationalist investigations into transformational grammars. To be sure, much effort has gone into the attempt to extend the psycholinguistic model to isolated cases of animal learning, but even if these experiments should (in the eyes of their practitioners) prove successful, it still leaves little room for any further movement down the scale of purposive behaviour as mechanistically conceived.

This theory has already been the subject of extensive philosophical critique. Following the concerted empiricist attack on his notion of 'tacit knowledge' [see 25] Chomsky sought refuge in the dubious notion of 'cognizing', which combined the agreeable property of retaining the 'structure and character of knowledge' with the theoretical advantage of mastering the complex 'depth rules' devised by the transformational grammarian [see 13, pp. 69ff]. As Baker and Hacker explained:

> The use of 'cognize' (or 'tacitly know') is only 'explained' to the extent that it is said to be just like 'know', except that one who only cognizes cannot tell one what he cognizes, cannot display the object of his cognizing, does not recognize what he cognizes when told, never (apparently) forgets what he cognizes (but never remembers it either), has never learnt it and could not teach it, and so on. In short, cognizing is just like knowing, except that it is totally different in all respects. This is a travesty of the term

'know', of the introduction of technical terms in science, and of respectable reasoning [6, pp. 344-5].

There might, however, be some who (misguidedly) suppose that *Language, Sense & Knowledge* advances matters little over the original empiricist assault on 'tacit knowledge'. For according to Quine, one can only say that a subject's actions are guided by a rule when 'the behaver knows the rule and can state it' [55, p. 442]. In order to appreciate the full force of Baker and Hacker's investigation into the 'platonic mythology of rules' it is imperative that we see how, while the heirs of the Logical Positivists repudiated the neo-rationalist notion of 'tacit knowledge', they did not discard the underlying notion of a system/organism's following 'embodied rules'. Even Quine does 'not question the notion of implicit and unconscious conformity to rule, when this a matter of fitting.' [54, p. 444] For 'the native speaker must have acquired some recursive habit of mind, however unconscious, for building sentences in an essentially tree-like way; this is evident from the infinitude of his repertoire' [55, p. 443]. But to say that 'the mind follows a system of rules which operate below the level of consciousness for the most part' [51, p. 2] does not, as Turing demonstrated, demand a semantic sleight of hand in which all the attributes of conscious rule-following are retained in everything but name.

The alternative computationalist solution is to divorce the notion of normative behaviour from that of 'tacit knowledge' as far as possible by interpreting *changes in knowledge states* in the cybernetic terms of feedback mechanisms outlined above [see 23, and 20, pp. 31-2]. On this approach, the mechanist need not be coerced into the nebulous realm of 'cognizing' minds; for the whole point of the 'epistemological justification' which Turing presented in §9 of 'On Computable Numbers' was to reduce conscious rule-following to its causal components. McCulloch and Pitts carried his argument a step further by applying the latter framework to any recursive system, regardless of the possibility of consciousness. For it follows that, since the only constraint imposed on such recursive systems is that their input/internal-state/output 'behaviour' conforms to the functions devised by the computationalist, there is no reason why the schema should be limited to man (and, perhaps, the higher animals). Lower creatures on the evolutionary scale, and indeed, artificial systems, can all be 'minded' in the above sense, provided they can satisfy the minimal cognitive demands imposed by Turing's 'mechanical rules'.[13] But then, this still leaves open the central question whether, given that a scientist can construct rules (algorithms) to explain (predict) natural phenomena, there is any

sense in which the fact that one can map these rules onto the activities of a system/organism entails that the behaviour of that system/organism is causally determined by those rules? That is, does it make sense to describe an activity as rule-governed when the *possibility* of consciously following those rules has been excluded *ab initio*, and the relation between rule and behaviour has in effect been rendered causal [cf. 66]?

It must be emphasized that the problem here does not concern the manner in which the activities of these putatively rule-based systems is described (e.g. 'guided', 'governed', 'fit', and 'caused' have all been suggested). It is the assumption that, because algorithms can be mapped onto causal sequences, the latter mechanisms must embody the former. But what does that mean? The formalist response to this last question was, as McCulloch and von Neumann both made clear, supplied by Turing's proof that all effectively calculable functions are mechanically calculable and hence (by Turing's version of Church's Thesis) Turing-machine computable. By following in Turing's footsteps, McCulloch and Pitts were pursuing a conception of algorithms that provided the function-theoretic means for delivering the unified account of learning systems demanded by the continuum picture. For Turing had shown that algorithms decompose into sets of 'meaningless sub-rules', each of which can as such be followed by a machine, be it natural or artificial. In other words, McCulloch and Pitts based the theory of formal neural nets on the premise that Turing had exposed the causal mechanics of rule-following. In Turing's case, this argument was guided by the primary intention of defending the possibility of artificially simulating a cognitive ability. But it was the preceding century of physiological investigations into the mechanics of learning - i.e. purposive/adaptative behaviour - that had established the framework which both validated and exploited Turing's epistemological interpretation of his computability results. The next step was not so much to defend the further thesis that the mind is a Turing machine, therefore, as to reformulate what one should understand by the latter term.

What one is confronted with are automata: a species of formal system whose characteristic axioms are:

1) the number of units in the system is fixed in advance
2) each unit can only be in a finite number of states
3) the units all operate on a uniform time scale
4) the system alters its internal state as a function of its present state and environment (inputs)
5) the system emits its outputs as a function of its inputs and internal states.

In the original meaning of the term an automaton was an artefact which possessed the power of spontaneous or self-controlled movement. From the sixteenth-century onwards it was associated (in Gothic Romances) with the idea of a soulless man-created man, and by the beginning of the nineteenth-century it had come to refer to a human being behaving in a mechanical (i.e. soulless) fashion [see 14]. In 1936 Turing had sought to synthesize these various themes, and in their extension of Turing's Thesis McCulloch and Pitts had revealed that the kind of 'machine' envisaged by Turing is satisfied by any autonomous system which conforms to the above axioms: i.e. which is capable of governing its internal state transitions and outputs by conforming to the above function-theoretic rules. Thus, to paraphrase Gödel's famous endorsement of Turing's Thesis, McCulloch and Pitts had shown that 'an automaton *qua* formal system is nothing but a mechanical procedure for producing theorems. The concept of formal system requires that reasoning/thinking/perceiving/intending ... be completely replaced by "mechanical operations" on formulas in just the sense made clear by Turing machines' [76, p. 84].

The computationalist outlook as it now existed was still far removed, however, from what behavioural scientists were beginning to regard as the physical reality of 'information-processing systems'. To meet this need the next stage in the evolution of Turing's Thesis was supplied by von Neumann's theory of self-reproducing automata. This was to complete the computationalist response to the vitalist objection that a machine cannot reproduce itself, thereby bringing mechanisms full circle by returning it to its original focus. The problem, in von Neumann's words, was that

> When an automaton performs certain operations, they must be expected to be of a lower degree of complication than the automaton itself. In particular, if an automaton has the ability to construct another one, there must be a decrease in complication as we go from the parent to the construct [74, p. 2092].

Bearing in mind the dictates of the continuum picture of learning/adaptation, however, one and the same theory must apply to natural as well as artificial automata, and it is clear in the case of the former that

> Organisms reproduce themselves, that is, they produce new organisms with no decrease in complexity. In addition, there are long periods of evolution during which the complexity is even increasing. Organisms are indirectly derived from others which had lower complexity [Ibid.].

Von Neumann's solution was to adapt Turing's, purely computing machines so as to become automata 'whose output is other automata'. To accomplish this he endowed McCulloch and Pitts' formal neural networks (now conceived as idealized cellular arrays that are equipped with the logical connectives '&', 'v', and '–', are surrounded by inert cells, and whose internal states are determined by the present state of each cell together with that of its contiguous cells) with construction cells (which enable the automaton to change the state of its surrounding cells - e.g. by transforming inert cells into automaton parts and vice versa) and transmission cells (which, when thresholds are triggered, transmit messages from the control to the construction cells).

The essence of his proof was: i) to equip an automaton (A) with a universal constructor which 'when furnished the description of any other automaton in terms of appropriate functions will construct that entity'; ii) introduce a 'reproducer' (B) that can make a copy of any such automaton description; iii) add a 'control mechanism' which would activate each of these automata and then install the instruction description made by (B) into the automaton constructed by (A); and iv) include this control mechanism in the description supplied to (A) and copied by (B). The key to the resulting self-reproducing automaton is that it can, in fact, produce any automaton from the description supplied to it, including itself. For the automaton operates in two distinct modes: it oversees the construction of a new automaton-plus-control mechanism, and when this has been completed it makes a copy of the description which it attaches to the new automaton. The description thus performs two roles: in the first it issues the set of instructions which (A) follows in the construction of the new automaton, and in the second all semantic content is ignored. Hence when supplied with a description of itself the automaton will first build an identical automaton, and when this is finished duplicate its own blueprint and attach this to the new automaton, thereby completing the process of self-reproduction.

It was inevitable that this theory should have immediately found favour in the fields which had prefigured so largely in the early stages of the modern evolution of mechanism. For von Neumann himself was quick to emphasize that:

> it is quite clear that the instruction ID is roughly effecting the functions of a gene. It is also clear that the copying mechanism B performs the fundamental act of reproduction, the duplication of the genetic material, which is clearly the fundamental operation in the multiplication of living cells [74, p. 2097].

That is, the control and transmission cells can be seen to perform a role analogous to the higher centres and the neurotransmitters in the central nervous system, and the constructor cells offer a schematic model of the body's various regenerative organs. Indeed, the theory offered an ideal model for interpreting Watson and Crick's subsequent discovery of the structure of DNA; for von Neumann had shown, not just that an automaton which contained a complete description of itself could reproduce, but indeed, could simulate genetic mutations if random changes where programmed into the cycle. Thus DNA itself could be said to perform the role of von Neumann's description, ribosomes that of the universal constructor, and protein molecules the analogues - or instantiations - of the new automata constructed by the gene/automaton.[14]

It is not difficult to explain this link to genetics: given the gradual transition from physiology through behaviourism to automata theory it was only natural that the original concern with the mechanics of homeostasis should be preserved throughout and that the preoccupation with adaptation should shift from an individual to a species-based approach [see 83 chapter 1]. Furthermore, the biological application of the computationalist theory was self-corroborating in that it served to complete the continuum picture. For scientists could now maintain that 'Living entails perceptual creative activity; repeated making of what can be called *choices* and *decisions*'. And this applies to every automaton on the continuum, ranging from man down to the

> simplest cell [in which] the chemical processes must go on continually, but also must continually change. The creature has to adapt itself to the surrounding conditions, which are inevitably altering all the time ... In this endeavour we find a series of activities parallel to human actions. In the pursuit of its aim of living, every organism must *search* and *decide* what to do, which way to go to get what it needs. From moment to moment there are several possibilities open to it and the choice between them is made by the information it already contains [84, p. 47].

This 'information'

> is embodied (or 'written') in the nucleotide groups that have survived during millions of years of natural selection. In the same way we can say that by organization of the neurons the simpler sort of 'knowledge' had been written in the brain long before man appeared. The 'knowledge' how to breathe, for instance, how to eat and to walk and to mate. ... All the knowledge must somehow be recorded in the brain during the process of learning[15] [84, p. 49].

Only an argument which had proceeded from the assumptions outlined in the preceding sections could have arrived at a conclusion that treated autonomic reflexes as 'knowledge states', and effaced the logico-grammatical distinction between instinctive and normative behaviour.

Not only had the continuum picture of learning/adaptation been completed: it had been further expanded in the process. For mathematicians were simultaneously engaged in plotting the unpredictable periodic and aperiodic patterns that can be generated from elementary recursive rules.[16] Together these developments seemed to presage the immanent penetration of the secrets of a 'recursive universe' in which complex life forms are built up from basic 'cellular machines', and which can thus be seen as operating in a realm that is governed by recursive as opposed to physical laws. Almost immediately this inspired a widespread biological search for the embodied algorithms that control the dynamics of evolution.[17] But the theory reaches even further, for 'the concept of computation is so universal that it can be used to explain all kinds of physical phenomena' as well, ranging from the recursive structure of a snowflake to that of nucleation.[18] This signified not so much a turning as a returning-point in the evolution of mechanism; for automata theory, as von Neumann conceived it, is

> the study of the fundamental principles common to artificial automata (e.g. digital computers, analog computers, control systems) and natural automata (e.g. the human nervous system, self-reproducing cells, the evolutionary aspects of organisms). Von Neumann envisaged a systematic theory which would be mathematical and logical in form, and which would be a coherent body of concepts and principles concerning the structure and organization of both natural and artificial systems, the role of language and information in such systems, and the programming and control of such systems. The theory of automata is an interdisciplinary subject which combines viewpoints of logic, communication theory, and physiology [10, p. xxv].

Post-computational mechanism thus serves as the vindication - or consummation - of the *Weltanshauung* inspiring its nineteenth-century physiological antecedents; for according to automata theory, all of nature, whether this be animate, inanimate, or artificial, is based on the execution of simple mechanical rules in the sense made clear by Turing.

It was but a short step from here to the formal inauguration of 'artificial intelligence' at the 1956 Dartmouth Conference, instigated by McCarthy in order to distinguish his growing

interest in the simulation of cognitive processes from his earlier work in automata theory.[19] Rather than scrutinizing the consequences of this transition, however, the philosopher of AI's real concern here must be with the 'Platonic mythology of rules' that sustained it. But how could a theory whose empiricist origins were faithfully nurtured possibly be identified with the very standpoint that it most staunchly opposed? The answer lies once again in the mathematical framework which provided the foundation for the post-computational mechanist conception of the 'recursive universe'. Foremost here is the significance of Hilbert's epistemological outlook for the interpretation and reception of Turing's thesis [see 61]. But that by no means exhausts the significance of Hilbert's framework for the evolution of AI; for from Hilbert onwards there had been a tendency to confuse the *application* of a model with the notion of an *interpretation*: an assumption that not only encouraged but, in fact, demanded a picture of the recursive structures which inhere in nature awaiting discovery. It is a premise which dates back to Hilbert's axiomatization of Euclidean geometry. In a letter to Frege Hilbert explains that 'every theory is only a scaffolding (schema) of concepts together with their necessary connections, and that the basic elements can be thought of in any way one likes' [17, p. 42]. That is, Hilbert believed that a pure geometry *defines* primitive concepts by establishing their logical form, and applied geometry *determines their meaning* by mapping them on to systems of objects. This left an obvious problem, however, for it is 'certainly confusing to say that the primitive terms can be defined twice' [75, p. 134].

The logical positivist solution was to construe Hilbert's 'implicit definitions' as a species of 'pre-definition' which delimit the 'class of interpretations'. The concept of *straight line*, for example, is given different logical forms by Euclidean, Bolyai-Lobatchevskian, and Riemannian geometry, while its meaning is given by models: 'If we interpret it ... in such a way that the term "straight line" is coordinated to the term "path of a light ray," all the other terms then acquire a quite definite significance' [75, p. 133]. But then, that means that the Euclidean term 'straight line' will mean something different depending on what we are measuring. Moreover, this still does not clarify the logical status of the proposition 'The path of a light ray is (or is not) a Euclidean straight line'. Waismann's response was to contend that

> Geometry, conceived as a body of conventions, is not as *a priori* as it seems to be; its choice is governed by empirical facts. On the other hand, geometry conceived as a body of

> factual statements, based upon inductive evidence is not as *empirical* as it seems to be: it is not a naturalistic description of a large heap of facts obtained by measuring experiments, but an 'idealised' representation like any physical law is: i.e. it contains a conventional element [75, p. 154].

But there can be no *tertium datur* as far as the logical grammar of geometrical propositions are concerned: geometry cannot be both a 'body of conventions' *and* a 'body of factual statements'.

The tension surfacing here is the result of the Logical Positivists' attempt to salvage the distinction between 'schematic' and 'semantic' definitions. Were Waismann to remain faithful to his professed conventionalist outlook, the above argument would really entail that the *interpretation* of 'straight line' as the path of a light ray is nothing of the sort; rather, it is an *alternative definition* in which the path of a light ray serves as a paradigm for 'straight line' (in much the same way that the standard metre bar was originally used as the paradigm for 'standard metre'). In which case, the meaning of 'straight line' ('metre bar') would fluctuate according to the physical behaviour of the paradigm, which was not at all his intention. The heart of Waismann's dilemma was that there must be some sense in which we can be certain that the shortest distance between two points - whether these be tables, chairs, beer-mugs, or pulsars - is a Euclidean straight line. What he failed to grasp is that the source of this certainty lies not in epistemology but rather, in the logical grammar of mathematical propositions. For the propositions of geometry are norms for describing the spatial relations that hold between objects; they license the inferences that one can draw about these spatial relations. It is the rules of Euclidean grammar, for example, which determine the necessary fact that if the length of the segment between two points on the path of a light ray is a minimum, then the path of the light ray divides a plane into two identical halves (in all but position). The relation between mathematical concepts and their applications is thus internal: what the *path of a light ray* really signifies in Waismann's example is an *application*, not an interpretation of the Euclidean concept of straight line.

The basic problem with Hilbert's conception is that it rendered the relation between pure and practical geometry external. This left the Logical Positivists with the unanswerable question of how one could be certain that a given interpretation conformed to the 'logical structure of a theory'? [see 63]. To be sure, the platonist sees no difficulty accounting for 'the extensive coincidence between the mathematician's invented world and the natural world' [22, p. 62]; his response is to insist that 'the

mathematician's imagination is an extra sense with which we can perceive the natural world. And it is an extremely efficacious sense, because it often perceives reality long before our senses do' [22, p. 4]. In other words, the platonist resorts to a mystery in order to explain an enigma. The great appeal of Hilbert's axiomatic approach was that it claimed to do away with any need for such a faculty of *Urintuition*. But as became clear from the failure of logical positivist attempts to develop this conventionalist theory, it had left the underlying metaphysical problem inspiring platonist epistemology intact. On this argument, the mathematician tries to construct that pure geometry which will best satisfy the needs of science by conforming to the spatial reality under investigation. Thus in *Der Raum* Carnap covertly sought to reintroduce spatial relations into reality under the guise of the 'topological relations' with which the propositions of pure geometry are said to correspond [see 12, pp. 47f]. And therein lies the crux of the issue: by construing the meaning of mathematical propositions as supplied by semantic interpretation Hilbert had retained a descriptivist conception of mathematical propositions, and hence, a correspondence theory of mathematical truth. That is, he misconstrued the normative logico-grammatical character of mathematical propositions *ab initio*, and thus preserved the metaphysical presupposition that (true) mathematical propositions conform to how things are in the world.

Needless to say, it is hardly possible to do justice to so profound an issue in so short a compass [see 60, chaps. 7-8 and [7] chapter VI]. What concerns us here, however, is the manner in which this platonist assumption was preserved in the computationalist framework. Nelson drew attention to this point when he clarified the 'analogy between "Animals are automata" and "Physical space is euclidean"':

> An automaton is a mathematical object, and this statement is intended to say that certain physical things satisfy certain mathematical relations; or, more precisely - assuming one were to formalize automata theory along the lines of group theories or geometrical theories - to say that animals are models of automaton formalisms [50, p. 429].

We can now begin to understand why, as was observed at the outset of this paper, cognitive science has revitalized metaphysics in the philosophy of mind. In the case of automata, the platonist misconstrues the application of the formal system for what is termed a 'realization': i.e. the idea that any number of natural or artificial systems can be isomorphic with the formal system (or that the parts of a physical system can be isomorphic with myriad automata). In other words, the natural or artificial

system is said to be a *physical instantiation* of the formal system; hence its operations are literally governed by the same function-rules which define the automaton. From this was born the conception of the *embodied rules* which regulate the recursive universe. But the point of seeing the relationship between automata and physical things as that of system to application is to recognize that the fact that we can formulate rules for describing the operations of a natural phenomenon does not entail that those rules must somehow be present in that phenomenon *sub species aeternitatis*.

By overlooking this point automata theory soon found itself confronted with a similar dilemma to that which frustrated the Logical Positivists' attempts to reconcile conventionalism with the interrelatedness of mathematical truths. Let us suppose that it were possible to examine the operations of a neural assembly that had been mapped onto a neural net, and discrepancies were observed between the behaviour of the net and the rules of the automaton. Only two options would be open to the computationalist; either he could conclude that there *must* have been some error in the observations; or else he must accept that one can never be certain that the embodied rules governing the operations of a system have been discovered. But now, if he takes the first route he runs the risk of lapsing into armchair dogmatism; certainly no scientist wants to have the scope of his findings fixed in advance. But if he opts for the latter alternative he is condemned to suffer the sceptical consequence that one could never be certain that the algorithms which control nature have been correctly identified. In neither case is it clear what contribution automata theory has to make to neurophysiology, other than heuristic. Unless, of course, one were to return to the very picture which these developments were intended to displace, and maintain that 'the automata theorist's imagination is an extra sense with which he can perceive the natural world. And it is an extremely efficacious sense, because it often perceives reality long before our senses do.'

Such a move would not, in fact, be at all out of place, in so far as the underlying metaphysical presupposition inspiring platonist epistemology has also been preserved: here under the conflation of rules with causal mechanisms. For the mechanisms of rules embedded in the physical instantiations of natural automata are seen as operating on their own accord. That is, the rules determine on their own what shall constitute their applications. But this completely undermines the normative basis for the concept of rules: the ability to instruct, explain, correct, justify etc. one's actions by reference to the expression of the

rule. Without the possibility of possessing and displaying these abilities one has, not rule-governed actions but instead, causes and effects [7]. Rather than pursuing the logical grammar of *rule* and *rule-following* any further, however, our chief concern at this point must be to clarify the conceptual pressures which induced such a transgression. For as has been stressed throughout, automata theory represented the culmination, not the (mis)appropriation of Turing's Thesis. The fundamental premise operating here is not just that mathematicians can only develop (algorithmic) models of natural phenomena if the recursive rules which they employ are isomorphic with those embodied in the organism; more importantly, it was assumed that Turing's 'mechanical rules' explain exactly how this could be the case. It is no surprise that the two themes should have proved so complementary; for both had evolved from the same source. It is precisely because of this harmony that the theory has proved so attractive, and so damaging to our understanding of the nature of this mathematical achievement and the foundations of AI.

As Wittgenstein demonstrated in [79] and [81], the fact that rules can be mechanized does not entail that those rules are *mechanical* [see 61]. The danger here is to suppose that because recursive rules can be encoded in causal sequences this signifies that the latter constitute a 'representation' of the former. But the shift from *encoding* to *embodying* marks a categorial departure to causal domains from whence there can be no return to normativity. The point of emphasizing this distinction is to clarify that the relation between a rule and the actions which conform with it is internal, whereas in causal situations the relation between two events is strictly external. Hence an account of the machine can only explicate why it produced its results: not whether or not these were correct. Only the rules that have been encoded can establish this, and it is for that reason that they are *antecedent* to the machine's operations. That is, they establish the criteria which determine when one shall say that the machine is performing properly or malfunctioning. Turing had thus misconstrued rule-following as a (cybernetic) mechanism. The 'instructions' in his machine programs are certainly presented so as to look like rules, but they actually function as descriptions of the machine's printing and transit devices. And this has nothing whatsoever to do with rule-following; it simply shows how to break down a complex mechanical action - e.g. registering twice as many '1s' as were originally configured on a tape - into its sub-components. Moreover, the terms chosen here are entirely apposite, for the latter are indeed subject to 'breakdown', but not to negligence,

mental lapses, or misunderstandings. Hence they are immune from error, but not because they are infallible; for to be capable of making mistakes once again presupposes rule-following abilities.

One cannot emphasize enough the importance of Turing's demonstration that, given their (binary) encodability, recursive functions are ideally suited to mechanical implementation. But to mechanize rule-governed actions is to substitute, not subsume. It is to develop causal mechanisms that can greatly facilitate the much more tedious process of someone's applying those rules. But that is no basis for the assumption that, because those causal mechanisms can generate patterns that are isomorphic with natural structures, the latter - any more than the former - are recursive. None of the significance of these discoveries has been sacrificed by this critique, however; only the manner in which it is to be understood. For the essence of algorithms - as Turing did indeed show - is that they can be so easily mechanized. Hence, thanks to the aid of communications engineers, mathematicians have been able to play so prominent a role in modern physics and biology. Whether their impact on psychology has been quite so beneficial remains a moot point, to be pursued elsewhere. But no such doubts could be expressed about the relevance of the study of algorithms *vis-à-vis* computer science. It is the significance of this work for AI and thence philosophy *simpliciter* that is a far more problematic issue. For there are two different albeit interrelated questions involved here which post-computational mechanism has brought to the fore: viz. the nature of 'computer simulations' and of philosophical problems. The most pressing issue facing the philosophy of AI is to clarify why it is that the former cannot be used to resolve the latter. And the first step is to trace the origins of the premise that such could indeed be the case back to the assumption that machines are capable of *learning*.

As we saw in §1, the crux of Turing's version of the Mechanist Thesis was that the shift from 'brute force' to 'learning programs' signified the advance from 'slave' to 'intelligent' machines. But the basis for this argument was to be found in the continuum picture of learning/adaptation, which Turing undertook to reformulate by substituting computability theory's recursive networks of 'mechanical rules' for behaviorism's causal networks of stimulus-response connections. By proceeding from the same presuppositions, however, Turing was induced to commit the very conceptual transgression which he sought to avoid: viz. to misconstrue causal sequences for normative acts. For it is no more possible - on logical grounds - to organize all of nature

on the ability to follow or be guided by simple 'mechanical' rules than on the habituation of an organism to its environment. In the case of the latter thesis we are confronted with an initial misconception of *knowledge* from which ineluctably flows the violations of the logical grammar of *purpose*, *choice*, and *learning* which Turing inherited from behaviourism. In the case of the former it is the misconception of *rules* (and *information*) which leads to the violations of *purpose, choice*, and *learning* that were entrenched in the foundations of AI. This only served to embellish even further the Cartesian assumption that 'the mind of each human being forms a region inaccessible to all save its possessor. ... His neighbor's knowledge of each person's mind must always be indirect, a matter of inference' [77, p. 1]. To be sure, the tools may have changed, but the picture underpinning the shift from introspectivism to the modeling of internal representations has remained remarkably constant. Indeed, it is a picture which inspired the dreams of mathematicians long before Descartes fell under its spell.

York University (Canada)

NOTES

1. Peirce was to develop this theme in several important papers on 'logical machines' [see 35]. I am indebted to Kenneth Ketner for drawing this to my attention.
2. It is frustrating that Turing offers no clue to the source of his information on learning theory; the sole references listed are to Church's 'An Unsolvable Problem of Elementary Number Theory', Gödel's 'On Formally Undecidable Propositions of *Principia mathematica*', and 'On Computable Numbers', even though the paper has relatively little to do with recursion theory.
3. As this passage makes clear, there is no evidence to suggest that Turing was aware of Thorndike's findings in the 1930s that punishing a response in order to weaken it is less effective than rewarding it so as to strengthen it. As far as Turing was concerned, 'Pleasure interference has a tendency to fix the character, i.e., towards preventing it changing, whereas pain stimuli tend to disrupt the character, causing features which had become fixed to change, or to become again subject to, random variation' [72, p. 17].
4. The issue is more complicated than this suggests, but in general, one can say that 'to refer to an action as intelligent was in general understood as indicating that its performance

showed some beneficial effect of past experience' [9, p. 23].

5. In *The Organization of Behaviour* Hebb argued: 'In mammals even as low as the rat it has turned out to be impossible to describe behavior as an interaction directly between sensory and motor processes. Something like *thinking*, that is, intervenes. "Thought" undoubtedly has the connotation of a human degree of complexity in cerebral function and may mean too much to be applied to lower animals. But even in the rat there is evidence that behavior is not completely controlled by immediate sensory events: there are central processes operating also' [24, p.xvi].
6. That is, as far as the early history of AI is concerned; the theory remains a subject of lively interest, however, for learning theorists. Cf. in particular [21]. It is also making something of a comeback in computational studies of 'pre-programming'; cf. [59].
7. In the discussion which followed von Neumann's reading of 'The General and Logical Theory of Automata' at the Hixon Symposium, McCulloch revealed: 'it was not until I saw Turing's paper that I began to get going the right way around [in his efforts to develop a theory of human computation], and with Pitts' help formulated the required logical calculus. What we thought we were doing (and I think we succeeded fairly well) was treating the brain as a Turing machine. ... The delightful thing is that the very simplest set of appropriate assumptions is sufficient to show that a nervous system can compute any computable number. It is that kind of a device, if you like - a Turing machine' [32, p. 32].
8. It is interesting to note that, according to Granit, 'the more complex the function defined, the more neural space it seems to occupy in all dimensions' [20, p. 58].
9. Cf. Poggio and Koch: 'neurons are complex devices, very different from the single digital switches as portrayed by McCulloch and Pitts [44] type of threshold neurons. It is especially difficult to imagine how networks of neurons may solve the equations involved in visual algorithms in a way similar to digital computers' [52].
10. cf. Lashley's opening statement in 'The Problem of Serial Order in Behavior', which summarized the guiding spirit of the Hixon Symposium: 'My principal thesis today will be that the input is never into a quiescent or static system, but always into a system which is already actively excited and organized. In the intact organism, behavior is the result of interaction of this background of excitation with input from

any designated stimulus. Only when we can state the general characteristics of this background of excitation, can we understand the effects of a given input' [37, p. 112].

11. A precedent established by Donald McKay in [46]. In [47] McKay explained that, 'the term "behaviour" has a special psychological use which would here be slightly question-begging, and may have caused some confusion. I should make it clear, therefore, that I used it in ['Mindlike Behaviour in Artefacts'] to refer non-committally to *all that goes on* in an artefact, internally and externally - as in ordinary usage where one speaks of the "behaviour" of a physical or mathematical system' [p. 62].
12. Dennett explains: 'both strategies ought to work in principle, because both ought to end up having completed exactly the same task. In the end both sides want to understand the relationship between the brain and the mind. And you can either start with the mind and work down, or you can start with the bits of brain and work up. So if you compare this with the analogy of building a trans-continental railroad, you *do* start at both ends and plan to meet somewhere in the middle. I would bet, however, that most of the track is going to be laid by the people who are working from the top down, rather than from the bottom up. For a very simple reason: top-down, is much easier, as it turns out' [16, p. 69].
13. But cf. Nelson, who insists that 'a being has a mind if and only if its body or certain body parts are guided by formally distinct rules (essentially of a nondeterministic finite automaton) of a complexity sufficient to account for intentionality, and it is capable of conscious feeling' [81, p. 10].
14. Young explains: 'The making of each protein is organized by a section of the information in the DNA, or gene, a page in the instruction book, which is first 'transcribed' (as the biochemists say) into a copy written in a slightly different code in molecules called messenger RNA (ribonucleic acid). This transcription is done by special enzymes called RNA polymerases which move along the stretch of DNA "reading off" the bases to make the RNA copy molecule. This molecule then moves through the cell to one of the protein-making machines, called ribosomes. These have further special enzymes which 'translate' the information in the RNA to organize the making of a new protein molecule' [84, pp. 35-6].
15. Young further explains: 'This selective activation of certain of the genes is the fundamental basis of the process of adaptation to environment, which is an essential part of all living. The DNA carries the information about the various

things that the organism can do. By responding suitably to the surroundings each bacterial cell then selects which of the possible proteins it shall make. The particular combination chosen will depend on the surroundings and on the genetic make-up and past history of that particular cell. ... In this process of selection a cell can be said to be using the DNA code for communication about the environment, as we use the words of language. ... This process of learning is the response to the environment that is particularly characteristic of man' [84, pp. 37-8,42].

16. Cf. in particular Ulam's work on 'recursively defined geometric objects' and Conway's game of 'Life'. Both of these so-called 'simulation games', their relation to von Neumann's work in self-reproducing automata, and the connection with genetics are summarized in [53].
17. A development, interestingly, which Turing once again anticipated towards the end of his life; this time, in his studies in embryology.
18. Thus Brian Hayes 'used the computational metaphor to explain how molecules of water "know" to form "the elaborate symmetries of a snowflake."' On this account '"There is no architect directing the assembly ... and the molecules themselves carry within them no template for the crystalline form." Instead ... the snowflake works like a cellular automaton. "Pattern on a large scale emerges entirely from the short-range interactions of many identical units. Each molecule responds only to the influence of its nearest neighbors, but a consistent arrangement is maintained throughout a structure made up of perhaps 1020 molecules." To see how this can be explained computationally, "imagine that each site where a molecule might be emplaced is governed by a rudimentary computer. As the crystal grows, each computer surveys the surrounding sites and, depending on its findings, determines by some fixed rule whether its own site should be occupied or vacant"' [quoted in 34, p. 81].
19. Cf. [41, chapt. 5]. McCorduck recounts the interesting fact that, while 'In a logical genealogy, Turing wo[u]ld be central ... Turing's work had practically no influence on most people at the Dartmouth Conference. For instance, Minsky felt himself much more influenced by McCulloch and Shannon (especially Shannon's early chess paper); Simon considered Turing of no particular influence on his work' [p. 95 fn. 1]. What this actually reveals is the speed with which Turing's ideas were assimilated. But as the preceding sections (and [61]) should make clear, the real danger in such attempts to

locate the influence of key figures is that it encourages one to overlook the overriding importance of the conceptual pressures created by the frameworks in which they operated.

REFERENCES

1. Arbib, Michael A., *Brains, Machines, and Mathematics* (New York, McGraw-Hill Book Company, 1964).
2. Arbib, Michael A., 'Cognitive Science: The View from Brain Theory', *The Study of Information*, Fritz Machlup and Una Mansfield (eds.), (New York, John Wiley & Sons, 1983).
3. Ashby, W.R. 'The nervous system as physical machine', *Mind* vol. 56 (January 1947).
4. Ashby, W. Ross, *Design for a Brain* , 2nd ed. (London, Chapman & Hall, 1960).
5. Babbage, Charles, *On the Principles and Development of the Calculator* (New York, Dover Publications, 1961).
6. Baker, G.P. and P.M.S. Hacker, *Wittgenstein: Rules, Grammar and Necessity* (Oxford, Basil Blackwell, 1984).
7. Baker, G.P. and Hacker, P.M.S. *Wittgenstein, Rules, Grammar and Necessity* (Oxford, Basil Blackwell, 1984).
8. Bower, Gordon H. and Ernest R. Hilgard, *Theories of Learning*, 5th edn (Englewood Cliffs, Prentice-Hall, Inc., 1981).
9. Bruner, Jerome, *In Search of Mind* (New York, Harper and Row, 1983).
10. Burks, Arthur W. (ed.), *Essays on Cellular Automata* (Chicago, University of Illinois Press, 1970).
11. Campbell, Jeremy, *Grammatical Man* (Harmondsworth, Penguin Books, 1982).
12. Carnap, Rudolf, *Der Raum, Kantstudien*, vol. 56 (Berlin, 1922).
13. Chomsky, Noam, *Rules and Representations* (New York, Columbia University Press, 1980).
14. Cohen, John, *Human Robots in Myth and Science* (London, George Allen & Unwin, 1966).
15. Davis, Philip J. and Reuben Hersh, *Descartes' Dream* (Boston, Houghton Mifflin Company, 1986).
16. Dennett, Daniel, 'Artificial Intelligence and the Strategies of Psychological Investigation', in [49].
17. Frege, Gottlob, *Philosophical and Mathematical Correspondence*, G. Gabriel et. al (eds.), Hans Kaal (trans.) Oxford, Basil Blackwell, 1980).
18. Gardner, Howard, *The Mind's New Science* (New York, Basic Books, 1985).

19. George, F.H., *The Brain as a Computer (Oxford, Pergamon Press*, 1962).
20. Granit, Ragnar, *The Purposive Brain* (Cambridge, Mass., The MIT Press, 1979).
21. Grossberg, Stephen, 'A Theory of Human Memory: Self-organization and Performance of Sensory-motor Coders, Maps, and Plans', in *Progress in Theoretical Biology*, R. Rosen & F. Snell (eds.), vol. 5 (New York, Academic Press, 1978).
22. Guillen, Michael, *Bridges to Infinity* (Los Angeles, Jeremy P. Tarcher, Inc., 1983).
23. Harman, Gilbert, 'Linguistic Competence and Empiricism', in [25].
24. Hebb, D.O., *The Organization of Behavior* (London, John Wiley & Sons, 1949).
25. Hook, Sidney (ed.), *Language and Philosophy* (New York, New York University Press, 1969).
26. Hull, Clark L., 'Simple Trial-and-Error Learning: A Study in Psychological Theory', *Psychological Review*, vol. XXXVII (1930).
27. Hull, Clark L., 'A Mechanical Parallel to the Conditioned Reflex', *Science,* vol. 79 (1929).
28. Hull, Clark L., 'Knowledge and Purpose as Habit Mechanisms', *Psychological Review*, vol. 44 (1937).
29. Hull, Clark L., 'Mind, Mechanism, and Adaptive Behavior', *The Psychological Review,* vol. 37 (1930).
30. Hull, Clark L. *Principles of Behavior* (New York, Appleton-Century-Crofts, 1943).
31. Hull, Clark L., 'Principles of the Scientist: VI', *Perceptual and Motor Skills*, vol. 15 (1962).
32. Jeffress, L.A. (eds.), *Cerebral Mechanisms in Behavior* (New York, John Wiley, 1951).
33. Jefferson, G. *Selected Papers* (London, Pitman, 1960).
34. Johnson, George, *Machinery of the Mind* (Time Books, 1986).
35. Ketner, Kenneth Laine, 'Peirce and Turing: Comparisons and conjectures', *Semiotics*, vol. 68 (1988).
36. Lashley, Karl 'The Behavioristic interpretation of consciousness', *Psychological Review*, vol. 30, (1923) pp. 237-77, 329-353.
37. Lashley, K.S., 'The Problem of Serial Order in Behavior', in [32].
38. Lewes, G.H., *The Physical Basis of Mind* (Boston, 1877).
39. Loeb, Jacques, 'The Significance of Tropisms for Psychology', *The Mechanistic Conception of Life*, Donald Fleming (ed.), (Cambridge, Mass., The Belknap Press of Harvard University Press, 1964).

40. MacKenzie, Ann, 'Descartes on Life and Sense', *Canadian Journal of Philosophy*, forthcoming.
41. McCorduck, Pamela, *Machines Who Think* (New York, W.H. Freeman and Company, 1979.
42. McCulloch, Warren S., 'The Brain as a Computing Machine', *Electrical Engineering*, vol. 68 (1949).
43. McCulloch, Warren S., *Embodiments of Mind* (Cambridge, Mass., The MIT Press, 1965).
44. McCulloch, Warren S. and Walter Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of Mathematical Biophysics*, vol. 5 (1943).
45. McCulloch, Warren S. and Walter Pitts, 'How We Know Universals', in [43].
46. McKay, Donald, 'Mindlike Behaviour in Artefacts', *British Journal for the Philosophy of Science*, vol. II (1951).
47. McKay, Donald, 'Mentality in Machines', *Proceedings of the Aristotelian Society*, Supplement, vol. 26 (1952).
48. Miller, George A., 'The Background to Modern Psychology', in [49].
49. Miller, Jonathan, *States of Mind* (New York, Pantheon Books, 1983).
50. Nelson, R.J., 'Behaviorism is False', *The Journal of Philosophy*, vol. LXVI (1969).
51. Nelson, R.J., *The Logic of Mind* (Boston, D. Reidel Publishing Company, 1982).
52. Poggio, T. and C. Koch, 'Ill-posed problems in early vision from computational theory to analogue networks', *Proceedings of the Royal Society of London*, B 226 (1985).
53. Poundstone, William. *The Recursive Universe* (New York, William Morrow & Co., 1985).
54. Quine, W.V., 'Linguistics and Philosophy', in [25].
55. Quine, W.V., 'Methodological Reflections on Current Linguistic Theory', in *Semantics of Natural Language*, Donald Davidson and Gilbert Harman (eds.) (Dordrecht, D. Reidel Publishing Company, 1972).
56. Rosenblueth, Arturo, Norbert Wiener and Julian Bigelow, 'Behavior, Purpose and Teleology', *Philosophy of Science*, vol. 10 (1943).
57. Rosenblueth, Arturo and Norbert Wiener, 'Purposeful and Non-purposeful Behavior', in [21].
58. Shanker, S.G., 'The Decline and Fall of the Mechanist Metaphor', in *Artificial Intelligence: The Case Against*, Rainer Born (ed.) (London, Croom Helm Publishers, Ltd., 1986).
59. Shanker, S.G., 'Computer Vision or Mechanist Myopia?', in S.G. Shanker (ed.), *Philosophy in Britain Today* (London,

Croom Helm, 1987).
60. Shanker, S.G., *Wittgenstein and the Turing-Point in the Philosophy of Mathematics* (London, Croom Helm Publishers, Ltd., 1987).
61. Shanker, S.G., 'Wittgenstein versus Turing on the Nature of Church's Thesis', *Notre Dame Journal of Formal Logic*, vol. 28 (1987).
62. Shanker, S.G., 'AI at the Crossroads', in *Questions in Artificial Intelligence*, Brian Bloomfield (ed.) (London, Croom Helm Publishers, Ltd., 1987).
63. Shanker, S.G., 'Wittgenstein's Remarks on the Significance of Gödel's Theorem', in S.G. Shanker (ed.), *Gödels Theorem in Focus* (London, Croom Helm Publishers, Ltd., 1988).
64. Shyrock, Richard, *The Development of Modern Medicine* (New York, 1947).
65. Simon, Herbert, 'Why Should Machines Learn?', in *Machine Learning*, R.S. Michalski et. al (eds.) (Berlin, Springer-Verlag, 1983).
66. Skinner, B.F., *Contingencies of reinforcement* (Englewood Cliffs, N.J., Prentice-Hall, 1969).
67. Suppes, Patrick, 'From Behaviorism to Neobehaviorism', *Theory and Decision*, vol. 6 (1975).
68. Thomas, Stephen N., *The Formal Mechanics of Mind* (Ithaca, Cornell University Press, 1978).
69. Thorpe, W.H., *Purpose in a World of Chance* (Oxford, Oxford University Press, 1978).
70. Turing, Alan, 'Intelligent Machinery, A Heretical Theory', in Sarah Turing, *Alan M. Turing* (Cambridge, Heffers, 1959).
71. Turing, Alan, 'Computing Machinery and Intelligence,' in *Minds and Machines*, Alan Ross Anderson (ed.) (Englewood Cliffs, Prentice-Hall, Inc., 1964)
72. Turing, Alan, 'Intelligent Machinery', *Machine Intelligence*, vol. 5 (1969).
73. Turing, Alan, *A.M. Turing's ACE Report of 1946 and Other Papers*, B.E. Carpenter and R.W. Doran (eds.) (Cambridge, Mass., The MIT Press, 1986).
74. von Neumann, John, 'The General and Logical Theory of Automata', in *The World of Mathematics, vol. Four*, James R. Newman (ed.) (New York, Simon and Schuster, 1956).
75. Waismann, Friedrich, 'The Structure of Concepts', in *Lectures on the Philosophy of Mathematics* (Amsterdam, Rodopi, 1982).
76. Wang, Hao, *From Mathematics to Philosophy* (London, Routledge & Kegan Paul, 1974).
77. Washburn, Margaret Floy, *The Animal Mind* (London, The

Macmillan Company, 1926).
78. Wittgenstein, Ludwig, *Zettl*, G.E.M. Anscombe and G.H. von Wright (eds.), G.E.M. Anscombe (trans.) (Oxford, Basil Blackwell, 1967).
79. Wittgenstein, Ludwig, *Philosophical Investigations*, G.E.M. Anscombe (trans.), 3rd edition (Basil Blackwell, 1967).
80. Wittgenstein, Ludwig, *Lectures on the Foundations of Mathematics: Cambridge, 1939*, Cora Diamond (ed.) (The Harvester Press, Sussex, 1976).
81. Wittgenstein, Ludwig, *Remarks on the Foundations of Mathematics*, G.H. von Wright, R. Rhees, and G.E.M. Anscombe (eds.), G.E.M. Anscombe (trans.) 3rd edn (Basil Blackwell, Oxford, 1978).
82. Young, Robert, *Mind, Brain and Adaptation in the Nineteenth Century* (Oxford, Clarendon Press, 1970).
83. Young, J.Z., *Programs of the Brain* (Oxford, Oxford University Press, 1978).
84. Young, J.Z., *Philosophy and the Brain* (Oxford, Oxford University Press, 1987).