The Philosophical Foundations of Artificial Intelligence

Selmer Bringsjord and Konstantine Arkoudas
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

October 25, 2007

Contents

1 Introduction

The structure of this chapter is as follows. The next section (2) is devoted to explaining, in broad strokes, what artificial intelligence (AI) is, and that any plausible definition of the field must revolve around a key philosophical distinction between "strong" and "weak" AI.[1] Next (section 3), we take up the long-standing dream of mechanizing human reasoning, which in singular fashion continues to inextricably link philosophy, logic, and AI. The following section (4) summarizes the conceptual origins of the field, after which, in section 5, we discuss the computational view of the human mind that has underpinned most AI work to date, and the associated problem of mental content. Section 6 discusses a number of philosophical issues associated with classical AI, while section 8 discusses a relatively new approach to AI and cognitive science, and attempts to place that approach in a broader philosophical context. Finally (section 9), the future of AI, from a somewhat philosophical standpoint, is briefly considered.

2 What is AI?

This is itself a deep philosophical question, and attempts to systematically answer it fall within the foundations of AI as a rich topic for analysis and debate. Nonetheless, a provisional answer can be given: AI is the field devoted to building artifacts capable of displaying, in controlled, well-understood environments, and over sustained periods of time, behaviors that we consider to be intelligent, or more generally, behaviors that we take to be at the heart of what it is to have a mind. Of course this "answer" gives rise to further questions, most notably, what exactly constitutes intelligent behavior, what it is to have a mind, and *how* humans actually manage to behave intelligently. The last question is empirical; it is for psychology and cognitive science to answer. It is particularly pertinent, however, because any insight into human thought might help us to build machines that work similarly. Indeed, as will emerge in this article, AI and cognitive science have developed along parallel and inextricably interwoven paths; their stories cannot be told separately. The second question, the one that asks what is the mark of the mental, is philosophical. AI has lent significant urgency to it, and conversely, we will see that careful philosophical contemplation of this question has influenced the course of AI itself. Finally, the first challenge, that of specifying precisely what is to count as intelligent behavior, has traditionally been met by proposing particular behavioral tests whose successful passing would signify the presence of intelligence.

The most famous of these is what has come to be known as the *Turing Test* (TT), introduced by Turing (1950). In TT, a woman and a computer are sequestered in sealed rooms, and a human judge, ignorant as to which of the two rooms contains which contestant, asks questions by email (actually, by *teletype*, to use the original term) of the two. If, on the strength of returned answers, the judge can do no better than 50/50 when delivering a verdict as to which room houses which player, we say that the computer in question has *passed* TT. According to Turing, a computer able to pass TT should be declared a *thinking* machine.

His claim has been controversial, although it seems undeniable that linguistic behavior of the sort required by TT *is* routinely taken to be at the heart of human cognition. Part of the controversy stems from the unabashedly behaviorist presuppositions of the test. Block's "Aunt Bertha" thought experiment (1981) was intended to challenge these presuppositions, arguing that it is not only the behavior of an organism that determines whether it is intelligent. We must also consider *how* the organism achieves intelligence. That is, the internal functional organization of the system

[1] The distinction is due to Searle (1980).

must be taken into account. This was a key point of *functionalism*, another major philosophical undercurrent of AI to which we will return later.

Another criticism of TT is that it is unrealistic and may even have obstructed AI progress insofar it is concerned with *disembodied* intelligence. As we will see in the sequel, many thinkers have concluded that disembodied artifacts with human-level intelligence are a pipe dream—practically impossible to build, if not downright conceptually absurd. Accordingly, Harnad (1991) insists that sensorimotor capability is required of artifacts that would spell success for AI, and he proposes the *Total* TT (TTT) as an improvement over TT. Whereas in TT a bodiless computer program could, at least in principle, pass, TTT-passers must be robots able to operate in the physical environment in a way that is indistinguishable from the behaviors manifested by embodied human persons navigating the physical world.

When AI is defined as the field devoted to engineering artifacts able to pass TT, TTT, and various other tests,[2] it can be safely said that we are dealing with *weak* AI. Put differently, weak AI aims at building machines that *act* intelligently, without taking a position on whether or not the machines actually *are* intelligent.

There is another answer to the What is AI? question: viz., AI is the field devoted to building persons, period. As Charniak and McDermott (1985) put it in their classic introduction to AI:

> The ultimate goal of AI, which we are very far from achieving, is to build a person, or, more humbly, an animal. (Charniak & McDermott 1985, p. 7)

Notice that Charniak and McDermott don't say that the ultimate goal is to build something that *appears* to be a person. Their brand of AI is so-called *strong* AI, an ambitious form of the field aptly summed up by Haugeland:

> The fundamental goal [of AI research] is not merely to mimic intelligence or produce some clever fake. Not at all. AI wants only the genuine article: *machines with minds*, in the full and literal sense. This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely, we are, at root, *computers ourselves.* (Haugeland 1985b, p. 2)

This "theoretical conception" of the human mind as a computer has served as the bedrock of most strong-AI research to date. It has come to be known as the computational theory of the mind; we will discuss it in detail shortly. On the other hand, AI engineering that is itself informed by philosophy, as in the case of the sustained attempt to mechanize reasoning, discussed in the next section, can be pursued in the service of both weak and strong AI.

3 Philosophical AI: The Example of Mechanizing Reasoning

It would not be unreasonable to describe Classical Cognitive
Science as an extended attempt to apply the methods of proof
theory to the modelling of thought.
Fodor and Pylyshyn (1988, pp. 29-30)

This section is devoted to a discussion of an area that serves as an exemplar of AI that is bound up with philosophy (versus philosophy *of* AI). This is the area that any student of both philosophy and AI ought to be familiar with, first and foremost.[3] Part of the reason for this is that

[2] More stringent tests than TT and TTT are discussed by Bringsjord (1995). There has long been a tradition according to which AI should be considered the field devoted to building computational artifacts able to excel on tests of intelligence; see (Evans 1968, Bringsjord and Schimanski 2003).

[3] There are other areas that might be discussed as well (e.g., learning), but no other area marks the genuine marriage of AI and philosophy as deeply and firmly as that of mechanical reasoning.

other problems in AI of at least a partially philosophical nature[4] are intimately connected with the attempt to mechanize human-level reasoning.

Aristotle considered rationality to be an essential characteristic of the human mind. Deductive thought, expressed in terms of syllogisms, was the hallmark of such rationality, as well as the fundamental intellectual instrument ("organon") of all science. Perhaps the deepest contribution of Aristotle to artificial intelligence was the idea of *formalism*. The notion that certain patterns of logical thought are valid by virtue of their syntactic *form*, independently of their content, was an exceedingly powerful innovation, and it is that notion that remains at the heart of the contemporary computational theory of the mind (Pylyshyn 1989) and what we have called *strong* AI above, and which will be elaborated in section 5.

In view of the significance that was historically attached to deduction in philosophy (starting with Aristotle and continuing with Euclid, and later Bacon, Hobbes, Leibniz, and others), the very idea of an intelligent machine was often tantamount to a machine that can perform logical inference: one that can validly extract conclusions from given premises. *Automated theorem proving*, as the field is known today, has thus been an integral part of AI from the very beginning, although, as we will see, its relevance has been hotly debated, especially over the last two decades. Broadly speaking, the problem of mechanizing deduction can be couched in three different forms. Listed in order of increasing difficulty, we have:

- *Proof checking*: Given a deduction D that purports to derive a conclusion P from a number of premises $P_1, \ldots, P_n$, decide whether or not D is sound.
- *Proof discovery*: Given a number of premises $P_1, \ldots, P_n$ and a putative conclusion P, decide whether P follows logically from the premises, and if it does, produce a formal proof of it.
- *Conjecture generation*: Given a number of premises $P_1, \ldots, P_n$, infer an "interesting" conclusion P that follows logically from the premises, and produce a proof of it.

Technically speaking, the first problem is the easiest. In the case of predicate logic with equality, the problem of checking the soundness of a given proof is not only algorithmically solvable, but quite efficiently solvable. Nevertheless, the problem is pregnant with interesting philosophical and technical issues, and its relevance to AI was realized early on by McCarthy (1962), who wrote that "checking mathematical proofs is potentially one of the most interesting and useful applications of automatic computers." For instance, insofar proofs are supposed to express *reasoning*, we can ask whether the formalism in which the input proof D is expressed provides a good formal model of deductive reasoning. Hilbert-style formal proofs (long lists of formulas, each of which is either an axiom or follows from previous formulas by one of a small number of inference rules) were important as tools for metamathematical investigations, but did not capture deductive reasoning as practiced by humans. That provided the incentive for important research into logical formalisms that mirrored human reasoning, particularly as carried out by mathematicians. S. Jáskowski (1934) devised a system of *natural deduction* that was quite successful in that respect. (Gentzen (1969) independently discovered similar systems, but with crucial differences from Jáskowski's work.[5])

[4] Many examples can be given. One is the *frame problem*, which we discuss in section 6. Another is *defeasible reasoning*, which is the problem of how to formalize inference in the face of the fact that much everyday reasoning only temporarily commits us to conclusions, in light of the fact that newly arrived knowledge often defeats prior arguments. E.g., you no doubt currently believe that your domicile is perfectly intact (and could if pressed give an argument in defense of your belief), but if you suddenly learned that a vicious tornado had recently passed through your town (city, county, etc.), you might retract that belief, or at least change the strength of it. Defeasible reasoning has been studied and (to a degree) mechanized by the philosopher John Pollock (1992), whose fundamental approach in this area aligns with that of Chisholm (1966) to defeasible inference.

[5] The system of Jáskowski dealt with hypotheses by introducing a crucial notion of *scope*. Gentzen worked with *sequents* instead. See Pelletier (1999) for a detailed discussion.

The ideas of natural deduction introduced by Jáskowski and Gentzen later played a key role, not only in theorem proving and AI, but in computational cognitive science as well. Mental logic (Osherson 1975, Braine and O'Brien 1998, Rips 1994), in particular, a family of computational cognitive theories of human deductive reasoning, was heavily influenced by natural deduction.

The second problem is considerably harder. Early results in recursive function theory (Turing 1936, Church 1936) established that there is no Turing machine which can decide whether an arbitrary formula of first-order logic is valid (that was Hilbert's *Entscheidungsproblem*). Therefore, by Church's thesis, it follows that the problem is algorithmically unsolvable—there is no general mechanical method that will always make the right decision in a finite amount of time. However, humans have no guarantee of always solving the problem either (and indeed often fail to do so). Accordingly, AI can look for conservative approximations that are as good as they can possibly get: programs that give the right answer as often as possible, and otherwise do not give an answer at all (either failing explicitly, or else going on indefinitely until we stop them). The problem was tackled early on for weaker formalisms with seemingly promising results: The Logic Theorist (LT) of Newell, Simon, and Shaw, presented at the inaugural 1956 AI conference at Dartmouth mentioned earlier, managed to prove 38 out of the 52 propositional-logic theorems of *Principia Mathematica*. Other notable early efforts included an implementation of Presburger arithmetic by Martin Davis in 1954 at Princeton's Institute for Advanced Studies (Davis 2001), the Davis-Putnam procedure (M. Davis and H. Putnam 1960), variations of which are used today in many satisfiability-based provers, and an impressive system for first-order logic built by Wang (1960). It should be noted that whereas LT was intentionally designed to simulate human reasoning and problem-solving processes, the authors of these other systems believed that mimicking human processes was unnecessarily constraining, and that better results could be achieved by doing away with cognitive plausibility. This was an early manifestation of a tension that is still felt in the field and which parallels the distinction between strong and weak forms of AI: AI as science, particularly as the study of human cognition, vs. AI as engineering—the construction of intelligent systems whose operation need not resemble human thought.

Robinson's discovery of unification and the resolution method (Robinson 1965) provided a major boost to the field. Most automated theorem provers today are based on resolution.[6] Other prominent formalisms include semantic tableaux and equational logic (Robinson and Voronkov 2001). While there has been an impressive amount of progress over the last 10 years, largely spurred by the annual CADE ATP system competition,[7] the most sophisticated ATPs today continue to be brittle, and often fail on problems that would be trivial for college undergraduates.

The third problem, that of conjecture generation, is the most difficult, but it is also the most interesting. Conjectures do not fall from the sky, after all. Presented with a body of information, humans—particularly mathematicians—regularly come up with interesting conjectures and then often set out to prove those conjectures, usually with success. This discovery process (along with new concept formation) is one of the most *creative* activities of the human intellect. The sheer difficultly of simulating this creativity computationally is surely a chief reason why AI has achieved rather minimal progress here. But another reason is that throughout most of the previous century (and really beginning with Frege in the nineteenth century), logicians and philosophers were concerned almost exclusively with *justification* rather than with *discovery*. This applied not only to deductive reasoning but to inductive reasoning as well, and indeed to scientific theorizing in general (Reichenbach 1938). It was widely felt that the discovery process should be studied by psychologists,

[6] For instance, systems such as Spass (Weidenbach 2001), Vampire (Voronkov 1995), and Otter (Wos, Overbeek, Lusk and Boyle 1992).

[7] CADE is an acronym for Conference on Automated Deduction; for more information on the annual CADE ATP competition, see Pelletier, Sutcliffe and Suttner (2002).

not by philosophers and logicians. Interestingly, this was not the case prior to Frege. Philosophers such as Descartes (1988), Bacon (2002), Mill (1874), and Peirce (1960) had all attempted to study the discovery process rationally and to formulate rules for guiding it. Beginning with Hanson (1958) in science and with Lakatos (1976) in mathematics, philosophers started re-emphasizing discovery.[8] AI researchers also attempted to model discovery computationally, both in science (Langley, Simon, Bradshaw and Zytkow 1987) and in mathematics (Lenat 1976, Lenat 1983), and this line of work has led to machine-learning innovations in AI such as genetic programming (Koza 1992) and inductive logic programming (Muggleton 1992). However, the successes have been limited, and fundamental objections to algorithmic treatments of discovery and creativity in general—e.g., such as put forth by Hempel (1985)—remain trenchant. A major issue is the apparently holistic character of higher cognitive processes such as creative reasoning, and the difficulty of formulating a rigorous characterization of relevance. Without a precise notion of relevance, one that is amenable to computational implementation, there seems to be little hope for progress on the conclusion generation problem, or on any of the other similar problems, including concept generation and abductive hypothesis formation.

Faced with relatively meager progress on the hard reasoning problems, and perhaps influenced by various other critiques of symbolic AI (see Section 6), some AI researchers have launched serious attacks on formal logic, which they have criticized as an overly rigid system that does not provide a good model of human reasoning mechanisms, which are eminently flexible. They have accordingly tried to shift the field's attention and efforts away from rigorous deductive and inductive reasoning, turning them toward "commonsense reasoning" instead. For instance, Minsky (1986, p. 167) writes:

> For generations, scientists and philosophers have tried to explain ordinary reasoning in terms of logical principles with virtually no success. I suspect this enterprise failed because it was looking in the wrong direction: common sense works so well not because it is an approximation of logic; logic is only a small part of our great accumulation of different, useful ways to chain things together.

A good deal of work has been made toward developing *formal*, rigorous logical systems for modeling commonsense reasoning (Davis and Morgenstern 2004). However, critics charge that such efforts miss the greater point. For instance, Winograd (1990) writes that "Minsky places the blame for lack of success in explaining ordinary reasoning on the rigidity of logic, and does not raise the more fundamental questions about the nature of all symbolic representations and of formal (though possibly non-logical) systems of rules for manipulating them. There are basic limits to what can be done with symbol manipulation, regardless of how many 'different, useful ways to chain things together' one invents. The reduction of mind to decontextualized fragments is ultimately impossible and misleading." As we will see in the sequel, similar points have been made in the criticisms of Dreyfus (1992) and others, who have argued that symbol manipulation cannot account for such essential human traits as intuition, judgment, and imagination, all of which can play a key role in inference and problem-solving in general; and that human reasoning will never be matched by any decontextualized and unembodied (or "unsituated") system that works by formally representing and manipulating symbolic information.

4 Historical and Conceptual Roots of AI

AI officially started in 1956, launched by a small but now-famous summer conference at Dartmouth College, in Hanover, New Hampshire. (The 50-year celebration of this conference, AI@50, was held

[8] In mathematics, Pólya (1945) was instrumental in that respect (and had a significant influence on Lakatos).

in July 2006 at Dartmouth, with five of the original participants making it back. Some of what happened at this historic conference figures in the final section of this chapter.) Ten thinkers attended, including John McCarthy (who was working at Dartmouth in 1956), Claude Shannon, Marvin Minsky, Arthur Samuel, Trenchard Moore (apparently the youngest attendee, and the lone note-taker at the original conference), Ray Solomonoff, Oliver Selfridge, Allen Newell, and Herbert Simon. From where we stand now, at the start of the new millennium, the Dartmouth conference is memorable for many reasons, including this pair: one, the term 'artificial intelligence' was coined there (and has long been firmly entrenched, despite being disliked to this day by some of the attendees, e.g., Moore); two, Newell and Simon revealed a program — Logic Theorist (LT) — agreed by those at the conference (and, indeed, by nearly all those who learned of and about it soon after the Dartmouth event) to be a remarkable achievement. LT was capable of proving elementary theorems in the propositional calculus, and was regarded to be a remarkable step toward the rendering of human-level reasoning in concrete computation. For example, LT was able to find a proof of the fact that if p implies (truth-functionally) q, then the denial of q (i.e., $\neg q$) (truth-functionally) implies the denial of p (i.e., $\neg p$).

From the standpoint of philosophy, neither the 1956 conference nor the aforementioned Turing *Mind* paper of 1950 come close to marking the start of AI. Hobbes had already anticipated strong AI back in the seventeenth century, when he famously proclaimed that "ratiocination is computation." Roughly in that same era, Leibniz dreamed of a "universal calculus" in which all disputes could be settled by rote calculation. And Descartes had already considered something like the Turing test long before Turing, albeit adopting a rather pessimistic view of the matter in a somewhat glib manner:

> If there were machines which bore a resemblance to our body and imitated our actions as far as it was morally possible to do so, we should always have two very certain tests by which to recognise that, for all that, they were not real men. The first is, that they could never use speech or other signs as we do when placing our thoughts on record for the benefit of others. For we can easily understand a machine's being constituted so that it can utter words, and even emit some responses to action on it of a corporeal kind, which brings about a change in its organs; for instance, if it is touched in a particular part it may ask what we wish to say to it; if in another part it may exclaim that it is being hurt, and so on. But it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do. And the second difference is, that although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by which means we may discover that they did not act from knowledge, but only for the disposition of their organs. For while reason is a universal instrument which can serve for all contingencies, these organs have need of some special adaptation for every particular action. From this it follows that it is morally impossible that there should be sufficient diversity in any machine to allow it to act in all the events of life in the same way as our reason causes us to act. (Descartes 1911, p. 116)

But while the ceremonial inauguration of AI might have been the 1956 Dartmouth conference, and while philosophers might have ruminated on machines and intelligence for centuries, the key conceptual origins of AI can be found at the intersection of two of the most important intellectual developments of the twentieth century:

- the "cognitive revolution"[9] that started in the mid-1950s and which overthrew behaviorism and rehabilitated mentalistic psychology;

[9] Speaking of the cognitive revolution has become somewhat of a banality, particularly after the 1980s saw the publication of two books on the subject (Baars 1986, Gardner 1985). The term is certainly romantic and radical, but some psychologists (Leahey 1992) have questioned whether there was a revolution at all (at least in the sense of Kuhn).

- the theory of computability that had been developed over the preceding couple of decades by pioneers such as Turing, Church, Kleene, and Gödel.

The significance of each for AI will be briefly discussed below.

The cognitive revolution is typically associated with the work of George Miller and Noam Chomsky in the 1950s, particularly with the latter's notorious review of Skinner's theory of language (Chomsky 1996 [1959]). It had been anticipated in the 1940s by McCulloch and Pitts (1943) and other cybernetics pioneers who had already been pointing out the similarities between human thought and information processing, as well as by experimental results obtained by psychologists such as Tolman (1948), who, studying maze navigation by rats, presented evidence for the existence of "cognitive maps." Particularly influential was Chomsky's famous "poverty of stimulus" argument to the effect that the efficiency and rapidity of language acquisition during childhood cannot be explained solely by appeal to the meager data to which children are exposed in their early years; rather, they compel the postulation of innate mental rules and representations that encode linguistic competence. Strong cases for the existence of mental representations were also made by experimental findings pertaining to memory, such as the results of Sperling (1960), which indicated that humans typical *store* more information than they can report. Memory, after all, provides perhaps the clearest case of mental representation; it seems absurd to deny that people store information, i.e., that we have some sort of internal representation of information such as the year we were born or the names of our parents. That much is commonsensical to the point of triviality, as is the claim that people routinely talk as if they really have beliefs, hopes, desires, etc.; and indeed most behaviorists would not have denied these claims. What they denied was the *theoretical legitimacy* of explaining human behavior by positing unobservable mental entities (such as memories), or that intentional terminology had any place in a *science* of the mind. Essentially a positivist doctrine, behaviorism had a distrust of anything that could not be directly observed and a general aversion for theory. It had been the dominant paradigm in psychology for most of the twentieth century, up until the mid-1950s, until it was finally dethroned by the new "cognitive" approach.

5 Computational Theories of Mind and the Problem of Mental Content

Once the first steps were taken and mental representations were openly allowed in scientific theorizing about the mind, the "computer metaphor"—with which researchers such as Newell and Simon had already been flirting—became ripe for explosion. After all, computers were known to store structured data in their memories and to solve interesting problems by manipulating that data in systematic ways, by executing appropriate instructions. Perhaps a similar model could explain—and eventually help to duplicate—human thought. Indeed, the postulation of mental representations would not by itself go far if their causal efficacy could not be explained in a mechanistic and systematic manner. Granted that structured mental representations are necessary for higher-order cognition; but *how* do such representations actually *cause* rational thought and action? The theory of computation was enlisted precisely in order to meet this important theoretical need. The result became known as the *computational theory of mind* (CTM for short), a doctrine that has been inextricably linked with strong AI. In the next paragraph we briefly discuss the main tenets of CTM.

The first core idea of CTM is to explain intentional mental states by giving a computational spin to Russell's analysis (1940) of intentional sentences such as "Tom believes that 7 is a prime number" as *propositional attitudes* that involve a psychological attitude A (in this case believing) towards a proposition P (in this case, that 7 is prime). More precisely, to be in a mental state involving an attitude A and proposition P is to be in a certain relationship R_A with a mental representation M_P whose meaning is P. To put it simply, to have a belief that 7 is a prime number is to have a mental representation in your "belief box" which means that 7 is prime. The representation itself is symbolic. That is, your "belief box" contains a *token of a symbolic structure* whose meaning (or "content") is that 7 is prime. Thus mental representations have both syntax and semantics, much like the sentences of natural languages. They constitute a "language of thought," so to speak, or *mentalese.* But it is only their syntax—syntax being ultimately reducible to physical shape—that makes them causally efficacious. This is a plausible story because as work in logic and computability had shown, there exist purely syntactic transformations of symbolic structures that are nevertheless sensitive to semantics. Deductive proofs provide perhaps the best example: by manipulating formulas exclusively on the basis of their syntactic properties, it is possible to extract from them other formulas which *follow logically* from them. Syntax can thus mirror semantics, or, as Haugeland (1985a, p. 106) put it, "if you take care of the syntax, the semantics will take care of itself." On this model, a mental process is a sequence of tokenings of mental representations which express the propositional content of the corresponding thoughts. The causes and effects of each mental representation, what it can actually *do*, is determined by its syntax "in much the way that the geometry of a key determines which locks it will open" (Fodor 1987, p. 19). And the entire process is orchestrated by an algorithm, a set of instructions that determines how the representations succeed one another in the overall train of thought. That is the second core idea of CTM. The mind is thus viewed as a "syntactic engine" driving a semantic engine, and, at least in principle, its operation can be duplicated on a computer.

A natural extension of CTM is Turing-machine functionalism, which was first adumbrated by Putnam (1960) in an influential paper that helped to drive the cognitive revolution forward (at least in philosophical circles), to undermine behaviorism, and to shape the outlook of strong AI.[10] Functionalism in general is, roughly, the idea that the essence of a mental state is not to be found in the biology of the brain (or in the physics that underwrites the hardware of its CPU, in the case of a machine) but rather in the *role* that the state plays in one's mental life (or computations), and particularly in the causal relations that it bears to stimuli (inputs), behavior (outputs), and other mental (computational) states. Turing-machine functionalism, in particular, is the idea that the mind is essentially a giant Turing machine whose operation is specified by a set of instructions dictating that if the mind is in a certain state s and receives a certain input x, a transition is made to a state s' and an output y is emitted. The most popular—and least implausible—versions of Turing-machine functionalism allow for probabilistic transitions.

Also closely related to CTM (in fact stronger than it) is the *physical symbol system hypothesis* (PSSH) put forth by Newell and Simon (1976). According to it, a physical symbol system "has the necessary and sufficient means for general intelligent action" (1976, p. 116), where a physical symbol system is "a machine that produces through time an evolving collection of symbol structures," a symbol structure being a collection of symbol tokens "related in some physical way (such as one token being next to another)" and subject to a variety of syntactic operations, most notably "creation, modification, reproduction, and destruction." Newell and Simon regarded machines executing list-processing programs of the Lisp variety as the prototypical examples of physical symbol systems. While there have been various internal disagreements (e.g., pertaining to questions of innateness), in some form or other CTM, PSSH, and Turing-machine functionalism together loosely characterize "classical" or "symbolic" AI, or what Haugeland (1985a, p. 112) has dubbed GOFAI ("good old-fashioned AI"). All three were posited as substantive empirical theses,

[10] Note that subsequently Putnam (1988) had a volte-face and renounced functionalism.

CTM and Turing-machine functionalism about the human mind and PSSH about intelligence in general; (GOFAI too, was explicitly characterized by Haugeland as an empirical doctrine of cognitive science). They set the parameters and goals for most AI research for at least the first three decades of the field. They continue to be a dominant influence, although, as we will see, they are no longer the only game in town, having suffered considerable setbacks as a result of forceful attacks that have elaborated serious conceptual and empirical problems with the GOFAI approach.

According to CTM, complex thoughts are represented by complex symbolic structures in much the same way that, in natural languages and formal logics alike, complex sentences are recursively built up from simpler components. Thus the mental representation of a complex thought such as "All men are mortal" contains component mental representations for concepts such as "mortal" and "men," as well as "all" and "are." These components are somehow assembled together (and eventually science should be able to spell out the details of how such symbolic operations are carried out in the brain) to form the complex thought whose content is that all men are mortal. That is how complex thoughts attain their meaning, by combining the meanings of their components. Now, a compositional story of this sort—akin to the compositional semantics championed by Frege and Tarski—is only viable if there is an inventory of primitives that can be used as the ultimate building blocks of more complex representations. The central question for CTM, which has a direct analogue in AI, is the question of how these primitives acquire meaning. More precisely, the question is how mentalese primitives *inside* our brains (or inside a robot's CPU) manage to be about objects and state of affairs *outside* our brains—objects that may not even exist and state of affairs that may not even obtain. This is also referred to as the *symbol grounding* problem (Harnad 1990). It is not merely a philosophical puzzle about the human mind, or even a protoscientific question of psychology. It has direct engineering implications for AI, since a plausible answer to it might translate into a methodology for building a robot that potentially averts some of the most devastating objections to CTM (these will be discussed in the next section); i.e., a robot that "thinks" by performing computations over formal symbolic structures (as we presumably do, according to CTM), but is nevertheless sufficiently grounded in the real world that it can be said to attain extra-symbolic understanding (as we do). Clearly it is not tenable to suggest that evolution has endowed us with all the right primitive symbols having all the right meanings built in, since evolution could not have foreseen thermostats, satellites, or Bill Clinton.

A number of theories have been expounded in response, all falling under the banner of "naturalizing content," or "naturalizing semantics," or "naturalizing intentionality." The objective is to provide a physicalistic account of how mentalese symbol tokens in our heads manage to be about things that are external to us (or, framing the issue independently of CTM, how mental states can achieve meaning). The type of account that is sought, in other words, is reductive and materialistic; it should be expressed in the non-intentional vocabulary of pure physical science.[11] In what follows we will briefly review three of the most prominent attempts to provide naturalized accounts of meaning: informational theories, evolutionary theories, and conceptual-role semantics.

The gist of informational theories is the notion of covariance. The idea is that if a quantity x covaries systematically with a quantity y, then x carries information about y. A car's speedometer covaries systematically with the car's speed and thus carries information about it. Accordingly, we can view the speedometer as an intentional system, in that its readings are *about* the velocity of the car. Likewise, we say that smoke "means" fire in that smoke carries information about fire. This is a sense of the word "means" that Grice called *natural meaning*, a notion that presaged theories of informational semantics. Again, because smoke and fire are nomologically covariant, we can say that one is about the other. Here, then, we have the beginnings of a naturalized treatment

[11] Whether or not this is even possible is not discussed as much as it should.

of meaning that treats intentionality as a common natural phenomenon rather than a peculiarly mental one. Concerning mentalese semantics, the core insight of such theories—put somewhat simplistically—is that the meaning of a symbol is determined by whatever the tokenings of that symbol systematically (nomologically) covary with. If a token of a certain mentalese symbol $\mathcal{H}$ pops up in our brains whenever a horse appears in front of us, then $\mathcal{H}$ carries information about horses and thus means *horse*. The idea has a straightforward appeal but is vulnerable to several objections; we will only mention three. The first is non-existent objects such as unicorns, and abstract objects such as the square root of two. How can these actually *cause* anything, and how can they nomologically covary with brain states? The second is the so-called disjunction problem. It cannot possibly be the case that the meaning of a symbol is whatever causes its tokenings (whatever such tokenings systematically covary with), because the aforementioned symbol $\mathcal{H}$, for instance, might be systematically caused not just by actual horses but also by cows that appear in front of us at night or under other suitably misleading environmental conditions. So if the preceding thesis were true we would have to say that $\mathcal{H}$ does not have a univocal meaning; it does not only mean *horse*. Rather, it means *horse* or *cow* or *zebra* or *TV image of a horse* or *stuffed animal that looks like a horse* or $\cdots$, a potentially gigantic and seemingly open-ended disjunction. But that seems clearly wrong. In addition, it would have the absurd consequence of making misrepresentation impossible. A principal characteristic of intentionality, at least of the usual mental variety, is that any system which represents must also be able to *misrepresent*, to entertain mistaken thoughts about reality. Disjunctive causal definitions of the above sort would preclude that. Another issue with misrepresentation is that sometimes an actual horse fails to cause $\mathcal{H}$ (and if the conditions are right this will be systematic, not an one-off failure); so it cannot be the case that $\mathcal{H}$ is tokened iff a horse appears before us. For attempts to meet these and other objections see Fodor (1987), Fodor (1990), and Dretske (1988).

Evolutionary theories maintain, roughly, that intentional states are adaptations, in the same way that livers and thumbs are adaptations, and that the content (meaning) of an intentional state is the function for which it was selected, i.e., the purpose that it serves. Like all adaptationist theories, this too is susceptible to charges of Panglossianism (Gould and Lewontin 1979). Nevertheless, the basic story is not implausible for beliefs to the effect that there is a predator nearby or for desires to get and eat some bananas that appear in one's visual field. The content of such a belief would be something like "there is a tiger under that tree over there," which would presumably be the function for which such beliefs were selected (to correlate with tigers under trees), and the content of such a desire would be "I really want to eat those bananas over there," which would again coincide with the purpose served by such desires (obtaining food and surviving). However, the going gets a lot tougher for teleological accounts when it comes to believing in the independence of the continuum hypothesis, the desire to drink a martini and listen to Miles Davis while flying over the Atlantic, or the desire to commit harakiri. In addition, such accounts make intentionality exceedingly contingent on having the proper evolutionary history. Suppose, for instance, that a molecule-per-molecule replica of Dennett is created in some futuristic synthetic-biology lab; or perhaps that some such replica already exists in an alternative universe where it was *created* (say, by a god) rather than evolved. Surely the latter would be a horrifying possibility for both Dennetts to contemplate, and pun aside, that is precisely the point. It seems counterintuitive to deny that such replicas would have beliefs and other mental states with the exact same contents as Dennett's, even though they are not the products of evolution. It should also be noted that such biological accounts of meaning do not bode well for getting the semantics right when *building* AI agents. See Millikan (1993) for a subtle evolutionary account of meaning that discusses these and other issues.

Conceptual role semantics (CRS) takes its cue from Wittgenstein's famous "use" theory of meaning, according to which the meaning of a linguistic item (expression, sentence, etc.) is the way in which that item is used by speakers of the language. The main thesis of CRS is that the meaning of a mentalese symbol S is fixed by the role that S plays in one's cognitive life, and particularly by the relations that it bears to other symbols, to perception, and to action. It is thus very similar to functionalism about mental states. Logical connectives such as *and* provide the standard illustrations for the use theory of linguistic meaning, and, likewise, the meaning of a mentalese symbol equivalent to *and* would be the role it plays in our heads, e.g., the set of mentalese inferences in which it participates. The theory has notable similarities to the *operational semantics* of programming languages in theoretical computer science. In cognitive science, CRS has been known as *procedural semantics.* It was primarily advocated by Johnson-Laird (1977), and roundly criticized by Fodor, as we will see below. CRS is an essentially holistic theory, as no symbol is said to have a meaning in and of itself; meaning is acquired only by relation to the totality of the meanings of all other symbols. A rather obvious problem is that no two people ever have the exact same mental states. In fact any two random people are likely to have quite different sets of beliefs, and therefore the role that a certain symbol plays in one's network cannot possibly be the same role that it plays in the other's network. It would seem to follow that meaning will vary from person to person, a result that would make it harder to explain the fact that people are quite capable of understanding one another. We do not view this as a decisive objection. A symbol might well have different shades of meaning for each of us at the peripheries of our respective networks, but the core uses that we make of it can be sufficiently invariant to enable successful communication. Nevertheless, the objection does point to a more general issue with CRS, that of correctness. If different people can make diverging uses of one and the same symbol and think different thoughts with it, as indeed they often do, then some of them will surely be erroneous. But what constitutes error in CRS? The question is not easy to answer if meaning is identified with use. Presumably meaning has a normative aspect: some uses of a symbol are correct and others are not. In computer science this is not an issue because the formal semantics of an artificial language are laid down with explicit normative force, and at any rate in the case of both artificial and natural languages we are dealing with derived rather than intrinsic intentionality. But CRS does not seem to have the wherewithal to distinguish right from wrong. Invoking counterfactual uses in addition to one's actual uses does not help, for such uses would only tell us how the symbol *would* be used in different circumstances, not how it *should* be used in present or different circumstances.[12] Related to this is the whole issue of externalism. CRS makes the meaning of a symbol a thoroughly internal matter, contingent only on the relations it bears to other symbols and states (including perceptual and behavioral states). But clearly there is more to meaning than that. At least on the face of it, meaning seems to hook symbols to the world, not just to other symbols. The meaning of the term *dog* must have something to do with its *reference*, i.e., with actual dogs, and the meaning of *Aristotle* must have something to do with the actual Aristotle. More sophisticated externalist challenges were presented by Putnam and Burge, arguing respectively that meaning is a function of the overall physical and social environment in which one is embedded. So-called *two-factor* CRS were developed in response, in an attempt to distinguish between *narrow* and *wide* mental content. Narrow meaning is "in the head" and does not depend on the surrounding circumstances, while wide meaning hinges on reference and has truth conditions. An entire industry has grown around this distinction, and we do not have the space to delve into it here.

Incidentally, it was precisely the lack of a connection between mentalese and the world (or between a computer's symbols and the world) that was the chief objection of Fodor (1978) when he argued against the AI-style "procedural semantics" advocated by Johnson-Laird. The latter had

[12] Normativity in general is problematic across the board for so-called "naturalized" philosophical theories, i.e., theories which demand that the analysans be couched exclusively in the vocabulary of physical science. In epistemology, for example, naturalized theories of knowledge have had a very difficult time coping with justification.

written that the "artificial languages which are used to communicate the programs of instructions to computers, have both a syntax and semantics. Their syntax consists of rules for writing well-formed programs that a computer can interpret and execute. *The semantics consists of the procedures that the computer is instructed to execute*" (Johnson-Laird 1977, p. 189, our italics). This computer model was applied to human sentence processing, where

> the first step that a person must perform is to compile the sentence—to translate it into a program in his or her internal mental language. This is generally an automatic and involuntary process for anyone who knows the language well. If someone says, *Pass the salt* or *When did you last see your father?* or *I'm leaving for Chicago tonight*, the listener usually compiles a corresponding program. ··· Once a program is compiled, the question arises as to whether the listener should run it. Should he pass the salt? Should he tell the speaker when he last saw his father? (Johnson-Laird 1977, pp. 191-192)

In an important critique of Johnson-Laird's article that in some ways presaged Searle's Chinese room argument, Fodor protested that:

> the computer models provide no semantic theory *at all*, if what you mean by semantic theory is an account of the relation between the language and the world. In particular, procedural semantics doesn't supplant classical semantics, it merely begs the questions that classical semanticists set out to answer. ··· a machine can compile 'Did Lucy bring dessert?' and have not the foggiest idea that the sentence asks about whether Lucy brought dessert. For the machine-language "translation" of that sentence—the formula which the machine does, as it were, understand—*isn't* about whether Lucy brought dessert. It's about whether a certain formula appears at a certain address (Fodor 1978, p. 204, his italics).

6 Philosophical Issues

The three principal philosophical criticisms of strong AI that helped to change the tide in the AI community and point to new research directions are the following:

1. The critique of Hubert Dreyfus;
2. Block's critique of machine functionalism via the *China brain* thought experiments; and
3. Searle's *Chinese room* thought experiment.

All three surfaced within 10 years of one another. There had been several other philosophical criticisms of strong AI before these (e.g., the ones by Lucas and Penrose; see the article by Robinson in the present volume[13] computationalism, and there have been others since.[14] But these three generated the most debate and have had the greatest impact.

Dreyfus's critique was the first. It was a mixture of empirical and philosophical arguments. Empirically, his main charge was that AI researchers had simply failed to deliver the goods; despite exceedingly optimistic—and often grandiose—early forecasts, they had not managed to build general-purpose intelligent systems. This line of criticism was generally dismissed as invalid and unfair. Invalid because at best it showed that AI had not succeeded *yet*, not that it could not *ever* succeed; and unfair because AI was a very young field, and revolutionary technological breakthroughs could not be expected from a field in its infancy, despite the overly enthusiastic proclamations of some of its pioneers. Philosophically, Dreyfus argued that AI is an ill-conceived attempt to implement a rationalist programme that goes back at least to Leibniz and Hobbes, a project

[13] See Bringsjord (1992) for a sustained, detailed updating of all these criticisms.

[14] For instance, see Bringsjord and Zenzen (1997).

that rests on the misguided "Cartesian" tenet which holds that human understanding consists in forming and manipulating symbolic representations. In contradistinction, he maintained that our ability to understand the world and other people is a non-declarative type of *know-how skill* that is not amenable to propositional codification. It is inarticulate, preconceptual, and has an indispensable phenomenological dimension which cannot be captured by any rule-based system. People do not achieve intelligent behavior in their daily lives by memorizing large bodies of facts and by following explicitly represented rules. Being born, equipped with a body and with the ability to feel, and growing up as part of a society are essential elements of intelligence and understanding. Dreyfus also stressed the importance of capacities such as imagination, ambiguity tolerance, and the use of metaphor, as well as phenomena such as fringe consciousness and gestalt perception, all of which were—and continue to be—resistant to computational treatment. Most importantly, in our view, Dreyfus stressed the importance of *relevance*, emphasizing the ability of humans to distinguish the essential from the inessential, and to effortlessly draw on relevant aspects of their experience and knowledge in accordance with the demands of their current situation, as required by their ongoing involvement with the world.[15] He correctly felt that imparting the same ability to a digital computer would be a major stumbling block for AI—what he called the "holistic context" problem. The problem of relevance remains, in our view, the key technical challenge to AI, both strong and weak, and to computational cognitive science as well.

The claim that people do not go about their daily activities by following rules points to a concern that has been a recurrent issue for strong AI and CTM, and even for general mentalistic theories such as Chomsky's generative linguistics, and merits a brief discussion here before we move on to Block's thought experiment. The objection has been made under somewhat different guises by many philosophers, from Wittgenstein and Quine to Dreyfus, Searle, and others. It has to do with the so-called *psychological reality* of rule-based explanations of cognition, and particularly with computerized simulations of mental processes. The issue hinges on the distinction between description and causation, and also between prediction and explanation. A set of rules (or a fortiori a computer program) might adequately *describe* a cognitive phenomenon, in that the rules might constitute a veridical model of the gross observational regularities associated with that phenomenon. They might fit all the available experimental data and make all the right predictions. But this does not mean that there is actually an encoded representation of the rules (or the program) inside our heads that is causally implicated in the production of the phenomenon. A set of grammatic rules R might correctly describe certain constraints on English syntax, for instance, but that does not mean that English speakers have an encoding of R inside their brains which causes them to produce speech in accordance with R. So even though R might correctly predict behavior,[16] it does not necessarily explain it.[17] The distinction is also known in the terminology of Pylyshyn (1991, p.

[15] It may be of historical interest to note that much of what Dreyfus had to say was couched in the language of continental phenomenology and existentialism, heavily influenced by thinkers such as Heidegger and Merleau-Ponty. That was not exactly conducive to facilitating communication with AI researchers, or with analytic philosophers for that matter. For instance, using an example—and the terminology—of Heidegger, Dreyfus wrote that when one is using a hammer one is not experiencing the hammer as an object with properties but rather as "in-order-to-drive-in-the-nail." The hammer has a "way of being" that is called "readiness-to-hand," signifying availability, a way of being different from "unreadiness-to-hand," which signifies unavailability; both of these should be distinguished from the "occurrent," also known as the "present-at-hand" mode of being, which is the mode of being of "stable objects." It is hardly surprising that such verbal acrobatics failed to resonate with people who were brought up on *Principia Mathematica*. This was unfortunate, as Dreyfus had several prescient observations to contribute.

[16] In fact rules do not even do that, at least in Chomskyan linguistics, as they are supposed to model idealized human *competence* rather than actual performance.

[17] Note that prediction via computerized simulation is normally very different from nomological prediction as the latter is often understood in the philosophy of science. It is particularly different from the notion of scientific prediction found in the seminal deductive- and inductive-nomological models of Hempel (1965). Most notably,

233) as the difference between *explicit* and *implicit* rules. Implicit rules merely describe behavioral regularities, whereas explicit rules have encoded representations, presumably in our brains, which play a causal role in the production of the regularities. The issue of psychological reality gives rise to serious epistemological problems. What evidence would count as substantiating the claim that certain rules are explicitly encoded in our brains? How do we distinguish between different sets of rules or different computer programs that are nevertheless descriptively equivalent? What parts of a computerized model should be ascribed psychological significance and what parts should be ignored? Those who are sympathetic to Quine's arguments about the radical indeterminacy afflicting the study of language are likely to entertain similar misgivings about computational approaches to cognitive science and to conclude that the above difficulties are insurmountable. (Although it is not necessary to accept Quine's indeterminacy arguments or his behaviorism in order to reach these conclusions). Chomsky views such fears as manifestations of empirical prejudices about the human mind and of a deep-seated but unwarranted methodological dualism which presupposes a sharp distinction between the physical and mental realms. To him, the foregoing epistemological problems amount to nothing more than the usual inductive underdetermination issue that regularly confronts all sciences. Computational cognitive scientists such as Newell, Pylyshyn and others have responded more concretely by developing the notion of different levels of system description; championing the use of production systems in order to avoid commitment to control flow and other implementation details that would inevitably be specified in conventional programming languages; and trying to facilitate theory choice by encouraging rigorous testing with very different sets of experimental results, such as chronometric data, eye-movement data, verbal protocols, etc. Nevertheless, serious issues with computational cognitive modeling remain, and many continue to feel that the epistemological difficulties faced by such modeling do not stem from the usual underdetermination problem that is commonly found in the physical sciences, but from a fundamentally different sort of problem that is much more challenging.

A second influential criticism was directed specifically against machine functionalism. It was delivered by Block (1978) in the form of a thought experiment which asks us to imagine the entire population of China simulating a human mind for one hour. The citizens of China are all supplied with two-way radios that connect them to one another in the right way. We can think of the individual Chinese citizens as neurons, or whatever brain structures we care to regard as atomic. The people are also connected, via radio, to an artificial body, from which they can receive sensory stimuli and to which they can deliver output signals for generating physical behavior such as raising an arm. According to machine functionalism, one would have to conclude that if the Chinese simulated the right transition table faithfully, then by virtue of being properly related to one another and to inputs and outputs, they would in fact amount to a conscious mind. But this strikes us as counterintuitive, if not patently absurd. The resulting system might well be isomorphic to the brain, on some level of description, but it would not seem to harbor any sensations, pains, itches, or beliefs and desires for that matter. For similar reasons it would follow that no purely computational AI system could ever be said to have a genuine mind. Some functionalists have chosen to bite the bullet and concede that the "China brain" (or a properly programmed robot) would in fact possess genuine mental contents, chalking up our contrary intuitions to brain chauvinism, our propensity to regard only neurological wetware as capable of sustaining a mental life. But this is hard to swallow, and the thought experiment convinced many that unabashed functionalism is too liberal and must be either abandoned or significantly circumscribed.

The third seminal philosophical attack on strong AI was launched by Searle (1980) with his now-famous Chinese room argument (CRA). CRA has generated a tremendous amount of discussion

computerized simulations violate the thesis of structural identity of prediction and explanation, precisely because successful predictions on the basis of a computer program (or a set of rules) do not necessarily constitute explanations.

and controversy, and we will only provide a very cursory review of it here; for a detailed discussion the reader is referred to Cole (2004). CRA is based on a thought experiment in which Searle himself stars. He is inside a room; outside the room are native Chinese speakers who don't know that Searle is inside it. Searle-in-the-room, like Searle-in-real-life, doesn't know any Chinese, but is fluent in English. The Chinese speakers send cards into the room through a slot; on these cards are written questions in Chinese. The box, courtesy of Searle's secret work therein, returns cards to the native Chinese speakers as output. Searle's output is produced by consulting a rulebook: this book is a lookup table that tells him what Chinese to produce based on what is sent in. To Searle, the Chinese is all just a bunch of—to use Searle's language—squiggle-squoggles. The gist of the argument is rather simple: Searle inside the room is supposed to be everything a computer can be, and because he doesn't understand Chinese, no computer could have such understanding. Searle is mindlessly moving squiggle-squoggles around, and, according to the argument, that's all computers do, fundamentally. Searle has given various more general forms of the argument. For example, he summarizes the argument on page 39 of (Searle 1984) as one in which from

1. Syntax is not sufficient for semantics.
2. Computer programs are entirely defined by their formal, or syntactical, structure.
3. Minds have mental contents; specifically, they have semantic contents.

it's supposed to follow that

> No computer program by itself is sufficient to give a system a mind. Programs, in short, are not minds, and they are not by themselves sufficient for having minds.

Many replies have been given to CRA, both in its original incarnation and in the general form expressed above; perhaps the two most popular ones are the *systems reply* and the *robot reply.* The former is based on the claim that though Searle inside the room doesn't understand Chinese, the overall system that includes him as a proper part does. This means that the premise to the effect that Searle-in-the-room is everything a computer can be is called into question. The latter objection is based on the claim that though, again, Searle-in-the-room doesn't understand Chinese, this deficiency stems from the fact that Searle is not causally connected to the outside environment in the right manner. The claim is that in a real robot, meaning would be built up on the basis of the robot's causal transactions with the real world. So, though Searle may in some sense be functioning in the room as a computer, he's not functioning as a full-fledged robot, and strong AI is in the business of aiming at building persons as full-fledged robots. Searle has put forth replies to the replies, and the controversy continues. Regardless of one's opinions on CRA, the argument has undeniably had a tremendous impact on the field.

At the same time that philosophical criticisms like the above were being made, serious technical problems with classical AI began to emerge. One of them was the *frame problem.* By now the term has become quite vague. Sometimes it is understood as the relevance problem that was mentioned earlier (how to tell whether a piece of information might be relevant in a given situation); sometimes it is understood to signify the apparent computational intractability of holistic thought processes; and occasionally it is even misunderstood as a generic label for the infeasibility of symbolic AI. Perhaps the widest and least inaccurate reading of it is this: it is the problem of spelling out the conditions under which a belief should be updated after an action has been undertaken. In its original incarnation the problem was more technical and narrow, and arose in the context of a very specific task in a very specific framework: reasoning about action in the *situation calculus.* The latter is a formal system, based on first-order logic, for representing and reasoning about action, time, and change. Its basic notion is that of a *fluent*, which is a property whose value can change over time, such as the temperature of a room or the position of a moving object. Fluents are reified, and can thus be quantified over. Importantly, boolean properties of the world are themselves treated as fluents. Such a *propositional fluent* might represent whether or not an object is to the left of another object, or whether the light in a room is on. The world at any given point in time can be exhaustively described by a set of formulas stating the values of all fluents at that point; such a description is said to represent the *state* of the world at that point in time. Actions are also reified. Each action has a set of preconditions and effects, both of which are described in terms of fluents. If the preconditions of an action are satisfied in a given state, then the action can be carried out and will result in a new state. Starting from an initial state, which presumably represents the world when a robot first enters it, many different sequences of states are possible depending on the different courses of action that may be undertaken.

Many AI problems, such as planning, have a natural formulation in this setting. For instance, devising a plan to achieve a certain goal amounts to discovering a sequence of actions that will transform the current state to a state that satisfies the goal at hand. As it turns out, however, it is not sufficient to describe the effects of each action; one must also specify what fluents are *not* affected by an action. Suppose, e.g., that an action *liftBox* is available, by which a robot can lift a box off the floor. Its precondition is that the box must be on the floor and its effect is that the box is off the floor. And suppose that there is another action, *turnLightOn*, by which the robot can turn the light on in the room, its precondition being that the light is off and the effect being that the light is on. If the light is off initially and the box is on the floor, it would seem that performing the *turnLightOn* action followed by *liftBox* should result in a state where the box is off the floor and the light is on. But in fact neither conclusion follows, because it *could* be, for all we have said, that turning the light on lifts the box off the floor and that lifting the box off the floor turns the light off. In the former case the plan would not even work because the precondition of the second action (*liftBox*) would not be satisfied after the first action, while in the latter case the light would be off after the second action. To rule out such outlandish models, we need to explicitly specify the non-effects of each action via so-called "frame axioms." While succinct ways of stating frame axioms have been devised, the computational complexity of reasoning with them remains a challenge. Several other proposed solutions have been put forth, ranging from circumscription to altogether different formalisms for representing and reasoning about action and change. It is noteworthy that none of the proposed solutions so far comes anywhere near approaching the efficiency with which young children reason about action. It has been suggested that humans do not run into the problem of reasoning about the non-effects of actions because they take it for granted that an action does not affect anything unless they have evidence to the contrary.[18] However, the real problem, to which philosophers such as Fodor have latched, is this: How can we tell whether or not a piece of information constitutes "evidence to the contrary"? There are at least two separate issues here. First we need to be able to determine whether or not a piece of information is potentially relevant to some of our beliefs. That is again the relevance problem. And second, we need to be able to determine whether or not the information falsifies the belief. These are both engineering problems for GOFAI and general philosophical problems. On the engineering front, it is not too difficult to build a symbolic system that reaches a reasonable verdict *once the right background beliefs have been identified*. The major practical difficulty is quickly zeroing in on relevant information. Many have come to believe it highly unlikely that any symbol-manipulating system can overcome this difficulty.

[18] Of course a simpler explanation for why humans do not face this problem is that they do not go about their daily lives by representing the world in logical formulas and performing deduction over them. As we will explain shortly, that is the key conclusion reached by proponents of so-called "situated intelligence."

7 Connectionism and dynamical systems

Conceptual and engineering problems such as the above, combined with the disillusionment that followed a brief period of excitement over expert systems and the grand "fifth-generation" project launched in Japan during the 1980s, helped to pave the way for a backlash against GOFAI approaches, both in AI and in cognitive science. To a large extent that backlash was manifested in the very rapid ascension of *connectionism* during the 1980s. Connectionism had been around at least since the 1940s (the foundations had been laid by McCulloch and Pitts (1943)), but it did not begin to emerge as a serious alternative to GOFAI until the 1980s, largely due to the efforts of Rumelhart and McClelland (1986).

The basic conceptual and engineering tool of connectionists is the neural network. A neural network consists of a number of nodes (or "units") that resemble brain neurons.[19] Each node receives a number of input signals and delivers an output signal. The nodes are connected to one another so that the output of one node becomes an input to another node. Input and output values are typically represented by real numbers. The connections have weights attached to them, which are also represented by real numbers. Intuitively, the weight of a connection represents the influence that one node has on the output of another. The output of each node is a simple linear function of the inputs. A typical function computed by neural-network nodes is the linear-threshold function: the weighted sum of the input values is calculated, and an output of 1 or 0 is produced depending on whether or not the sum exceeds a certain threshold, respectively. If the output is 1 the node is said to be activated, or to fire; otherwise it is inhibited. Certain units are designated as the input and output nodes of the entire network; typically there is only one output node. Neural networks are capable of a certain type of learning; they can be trained to compute—or approximate—a target function. General-purpose learning algorithms exist, such as back-propagation, which, starting with random weights, repeatedly expose the network to different inputs in a training set and adjust the weights so as to bring the output closer to the correct value. Neural networks have been constructed that perform well on various non-trivial cognitive tasks, such as learning the past tense of English verbs or synthesizing speech from written text.

Neural networks have a number of remarkable features that set them apart from GOFAI systems. One of them is the absence of a central processing unit, or of any explicitly coded instructions that determine the behavior of the system. There are only individual nodes, and an individual node has only a small amount of entirely local information: the input values it receives from its neighbors. Owing to this massive locality and interconnectedness, neural networks are capable of graceful degradation, meaning that if some parts of the network are damaged, the network as a whole continues to function, with a performance drop that is more or less proportional to the amount of damage. In contrast, symbol-manipulating systems are usually brittle; a small deviation from the programmed course of events can lead to catastrophic failure. Such brittleness is atypical of human intelligence. Like the performance of neural networks, human cognition will suffer a continuous and graceful degradation under adverse conditions, instead of an abrupt general failure. Second, representation is distributed, in that pieces of information are not encoded by concrete symbolic structures; rather, a piece of information is essentially represented as a pattern of activity over the entire network—the firings of the various nodes. And the overall "knowledge" encoded by a neural network essentially resides in the weights of the various connections; it is sub-symbolic and

[19] From a biological perspective, the similarities are actually rather superficial. For instance, actual neurons are continuously operating at different intensity levels rather than being discretely on or off. And the learning algorithms that are typically used in AI and cognitive applications of neural networks are exceedingly implausible from a cognitive standpoint, typically requiring many thousands of training rounds. It is more accurate to think of neural nets as based on a neurologically inspired metaphor rather than a credible scientific model of neurons.

highly distributed. An important corollary of distributed representation is that neural networks end up sidestepping the vexing question of content that arises for classical CTM. The question of how atomic symbols manage to acquire their meaning does not arise—because there are no atomic symbols.

These interesting features of neural networks, in combination with the fact that they appear to be more biologically plausible than digital computers, continue to appeal to many cognitive scientists and AI workers, and intensive research in the field is continuing unabated, although so far there have been relatively few outstanding achievements. The problem of common sense, however, resurfaces in the setting of neural networks in a different guise. The intelligence exhibited by a (supervised) neural network is pre-built into the system by the human modeler who trains the network. But this is not enough to sufficiently circumscribe the space of possible hypotheses so as to rule out generalizations which are legitimate from the perspective of the training data but inept and inappropriate from the human perspective. There are legions of stories about neural networks which, after intensive training, came up with generalizations which had learned to distinguish features which were entirely irrelevant to the human modeler (indeed, features which had not even been noticed by the modeler). Moreover, in terms of computational power, anything that can be done by neural networks can be done by Turing machines, and therefore, by Church's thesis, there is nothing that neural networks can do which cannot also be done, say, by LISP programs.[20] This entails that even if brains turned out to be giant neural networks of sorts, it would be possible in principle to simulate them with perfect precision using classical GOFAI techniques. It would follow, for instance, that there do exist rule-based systems capable of passing the Turing test, even if those systems are so incredibly vast and unwieldy that it is practically impossible to build them. (Although, should brains turn out to be nothing but neural networks, that would certainly prove that it is not *necessary* for a system to form a symbolically encoded rule-based theory of a domain in order to achieve competence in that domain.) There are other issues related to the question of whether neural networks could ever manage to achieve general intelligence, including the famous *systematicity debate*, which was started by Fodor and Pylyshyn (1988) and is still ongoing, but we will not take these up here.

Closely related to connectionism is the *dynamical-systems* approach to intelligence (Port and Gelder 1995). That approach draws on the general theory of non-linear dynamical systems, conceiving of the mind as a continuous dynamical system, essentially a set of variables whose values evolve concurrently over time. The evolution of the system is typically described by a set of laws, usually expressed by differential or difference equations. The state of the system at a given moment in time is described by the values of the variables at that moment. The values of the variables at subsequent times (i.e., the dynamic trajectory of the system through the space of all possible states) is determined by the present state and the dynamical laws. Dynamical systems theory, however, is used to do pure cognitive science, not AI. That is, it provides a set of conceptual resources for understanding cognition and for modeling aspects of it as dynamical systems, but not for *building* intelligent systems. Accordingly, we will not have much to say on it here. Nevertheless, the advocates of the dynamical-systems approach to cognition typically emphasize the importance of time, context, interaction, embodiment, and the environment, and have thus been natural allies of situated and embedded AI, to which we turn next.

[20] Mathematically speaking, it is possible to define neural networks, such as those of Siegelmann and Sontag (1994), which can compute functions that are not Turing-computable. But this is accomplished simply by providing uncomputable inputs to the networks (more precisely, infinite-precision real numbers which encode uncomputable functions), and should not be taken to imply that neural networks are more powerful than Turing machines. Turing machines (or Lisp programs, or Java programs, etc.) can also compute uncomputable functions if they are provided with uncomputable inputs.

8 AI From Below: Situated Intelligence

As disillusionment with GOFAI began to take hold in the 1980s, AI researchers such as Rod Brooks of MIT were coming to the conclusion that systems which relied on detailed symbolic representations of the world, baked in ahead of time by the engineers, and on action generation via logical planning, conceived as search of the entire space of possible action sequences and taking into account all the preconditions and effects of each action, were infeasible, brittle, and cognitively implausible. They encouraged a shift of focus from higher-order symbolic tasks such as deductive reasoning to lower-level, ostensibly "simple" perceptual and motor tasks, such as sensing, moving, turning, grasping, avoiding obstacles, etc. They maintained that only fully embodied agents capable of carrying out these tasks adeptly can be truly validated as artificial agents dealing with their surrounding environments, and that only full embodiment has any hopes of properly "grounding" an artificial agent in the real world. GOFAI had typically either completely ignored or minimized the importance of such activities. Perceptual and motor faculties were seen as mere "transducers," peripherally useful and relevant only inasmuch they delivered symbolic representations of the world to the central thought processes or deployed effectors to translate the outcomes of such processes into bodily movements. Brute sensation and action were also seen as uninteresting in that, as Aristotle had pointed out, lower animals were also capable of them; surely what differentiates humans from insects, in the minds of GOFAI advocates, is our capacity for rational thought.

Brooks and his coworkers argued that bodily capacities were far from trivial—indeed, GOFAI had proved inept at building systems that had them. Moreover, they held that the study of such capacities could lend us valuable insights on how higher-order cognition can possibly emerge from them. Language and the capacity for symbolic thought have emerged very recently in human history, a consideration which suggests that evolution has put most of its effort into building up our sensory and motor systems. Once we understand the seemingly simple and mundane workings of such systems, the puzzle of intelligence will begin to dissolve. Language and reasoning will become simple once we know how to build a robot that can successfully navigate the physical world, for, according to Brooks, the "prime component of a robot's intellect" is not to be found in reasoning but rather in "the dynamics of the interaction of the robot and its environment." Essentially, the AI research program pursued by Brooks and his followers, which became known as *situated* AI,[21] amounted to "looking at simpler animals as a bottom-up model for building intelligence" (Brooks 1991, p. 16). To borrow a phrase from historiography, this was, in a sense, AI "from below." [22]

A key point made by Brooks and his team was that intricate symbolic representations of the world are simply not necessary for solving a wide variety of problems. Many problems can be more efficiently tackled by doing away with representations and exploiting the *structure of the surrounding environment*, an idea captured in the slogan "the world is its own best representation." Continuously sensing and interacting with the world in a closed feedback loop was thought to be a much more promising approach than building a static symbolic "model" of it (described, say, as a state in the situation calculus) and reasoning over that. Brooks and his team demonstrated their approach by building a robot named Herbert, whose task was to roam the halls of the MIT AI lab with the goal of identifying and disposing of empty soda cans. Herbert was built on a so-called subsumption architecture, consisting of a number of independent modules, each of them specialized

[21] Also known as *embedded* or *embodied* AI.

[22] History from below is a method of historiography, primarily associated with British historians such as Thompson (1963), that shifts attention from the classical a-few-great-men model to the unmentioned masses, to the "commoners" whose lives had typically been regarded as too uninteresting or unimportant for history—much as "low-level" activities such as perception and movement had been regarded as too uninteresting for intelligence.

for performing a specific task, such as moving forward. At any given time, a module might become activated or suppressed depending on the stimuli dynamically received by Herbert. The overall system relied on little or no internal representations and symbolic manipulation, but managed to exhibit surprisingly robust behavior.

By turning the spotlight away from internal representations and processes towards external behavior and continuous interaction with the environment, the work of Brooks and other situated-AI researchers marked a reverse shift away from the cognitive revolution and back towards behaviorism. Indeed, some have spoken about a "counter-revolution" in AI and cognitive science. We believe that such claims are exaggerated; the majority of researchers in these fields are not willing to renounce the scientific legitimacy of representations in explaining the mind or their usefulness as engineering tools. Nor should they, in our view. The points made by the situation theorists have been well-taken, and AI as a whole now pays considerably more attention to environmental context and embodiment. That is a positive development, and the trend is likely to persist. But the existence of mental representations seems as undeniable now as ever. People can simply close their eyes, shut their ears, and do some very non-trivial thinking about the possible effects of their actions, the tenth prime number, logical consequences of their beliefs, the structure of our solar system or water molecules, various properties of unicorns, and so on. The premise that such "classical" thinking will become straightforward once we understand how we tie shoelaces is dubious, and has been indeed been called into question by many (Kirsh 1991). In summary, the following provides a rough graphical depiction of the main differences between GOFAI and situated AI:

Classical	**Embodied**
representational	non-representational
individualistic	social
abstract	concrete
context-independent	context-dependent
static	dynamic
atomistic	holistic
computer-inspired	biology-inspired
thought-oriented	action-oriented

It is noteworthy that the advent of situational theories in AI and cognitive science had already been mirrored—or has since been mirrored—by similar movements in several areas of analytic philosophy. In philosophy of language, for instance, the ordinary-language turn that started to occur in Oxford in the 1950s, primarily as a reaction against logical positivism, can be seen as a precursor to the behavioral backlash against cognitivism. Speech-act theory, in particular, initiated by Austin and then taken up by Strawson, Searle, and others, was the first to emphasize that they key unit of linguistic meaning was not an abstract sentence but rather an *utterance*, thereby shifting attention from an isolated theoretical entity (the sentence) to a concrete *act* carried out by real people in real time. The trend has continued and strengthened, particularly as seen in the subsequent development of pragmatics and the growing recognition of the extensive and very intricate dependence of meaning on deictic and other contextual factors. The following passage is telling, coming as it does from a champion of the classical truth-conditional approach to the study of language:

> ⋯ We have yet to address ourselves to the second- and third-person pronouns, tense, spatial indexicals, discourse indexicals, and so forth. ⋯ How many further parameters must be added to our truth predicate? Montague, Scott, and others have solved this proliferation of parameters by relativizing truth simply to an "index," viz., to an *n*-tuple whose members are all the objects determined by a context that can affect the truth values of sentences relative to that context. An "index" was originally taken to be an octuple whose members are a world, a time, a place, a speaker, an audience, a sequence of indicated or demonstrated objects, a discourse-segment, and

> a sequence of assignments to free variables; other contextual factors may be added as needed. (Lycan 1984, p. 50)

Shortly afterwards Lycan confesses that “it is a bit disturbing that an n-tuple of this type has no definite number of members”, and notes that Lewis has pointed out that the eight aforementioned members will not suffice—a “sequence of prominent objects” is also needed, along with “a causal-history-of-acquisition-of-names coordinate, and a delineation coordinate for resolving vagueness.” The situational semantics of Barwise and Perry is one approach that was developed with the purpose of sensitizing the semantics of natural language to context; but it is still a classical approach, developed within a realist, objectivist framework. A more radical development that took place in the 1970s in the west coast was *cognitive linguistics*, an approach that was largely based on Eleanor Rosch’s work on categorization (itself presaged by Wittgenstein’s famous remarks on games and family resemblances), and carried out primarily by University of California linguists such as Charles Filmore, Ronald Langacker, and George Lakoff. It started out as a reaction against the traditional cognitivist theories of Chomsky, i.e., classical generative linguistics, but it went much further in that it rejected the realist metaphysical framework in which traditional linguistic theories were couched. It specifically renounced Aristotelian categorization (essentially, set theory), metaphysical realism (the existence of an objective mind-independent external reality, which we internally represent in our minds), and correspondence theories of truth. According to most cognitive linguists, reality is a construction of the mind.

Similar developments took place in the philosophy of science and mathematics, again to a large extent as a reaction against positivism. In science, philosophers such as Kuhn and Agassi emphasized that abstract systems of justification (essentially inductive and deductive logics) are poor models of scientific practice, which is, above all, a human *activity* that is highly contingent on social interaction and cultural and political factors. In mathematics, philosophers such as Lakatos and social constructivists such as Bloom launched vigorous attacks against “Euclideanism,” the formal style of doing mathematics whereby a seemingly indubitable set of axioms is laid down at the foundation and deductive consequences are thereby derived cumulatively and monotonically. That style, it was claimed, is too neat to reflect “real” mathematics. In fact it completely disregarded the *process* of *doing* mathematics (actually *coming up* with results), the dynamic and live aspects of the field.

We do not claim that all these threads necessarily influenced one another, or that there was anything inexorable about any of them. Nevertheless, the existence of certain salient common points of reference and underlying similarities is undeniable. There has been an overall trend away from statics and towards dynamics, from the abstract and decontextualized to the concrete and context-bound, from justification to discovery, from isolated contemplation to social interaction, and from thinking to doing. A dominant and recurrent theme has been the conviction that genuine understanding will never be attained by taking something that is dynamic and evolving, reactive, plastic, flexible, informal, highly nuanced, textured, colorful, and open-ended; and modeling it by something static, rigorous, unbending, and inflexible—i.e., essentially by replacing something alive by something that is dead. Everything depends on context, and it will never do to try to prepackage context into some “n-tuple of factors” that could be manipulated by machine. The trends in question are not unrelated to the increasing prominence of social sciences, cultural anthropology, feminist studies, etc., and to a large extent the conflict between GOFAI and situated AI can be seen as a reflection of the infamous science wars, the clash between traditional “objectivist” metaphysics and social constructivism, and more recently, in the case of cognitive science, the rationality wars, where situational theorists have been questioning the “ecological validity” of classical cognitive science and psychology laboratory experiments, and calling for a greater focus on ethology and on “real” behavior exhibited in the “real” world (as opposed to the supposedly artificial and highly constrained conditions of the laboratory). Our remarks here have not been intended as a commentary on the science wars or as an attempt to take sides, but merely as an effort to provide a greater context for understanding the backlash against GOFAI and the emergence of situational approaches to AI and the study of the mind.

9 The Future of AI

If past predictions are any indication, the only thing we know today about tomorrow’s science and technology is that it will be radically different than whatever we predict it will be like. Arguably, in the case of AI, we may also specifically know today that progress will be much slower than what most expect. After all, at the 1956 kickoff conference at Dartmouth College, Herb Simon predicted that thinking machines able to match the human mind were “just around the corner” (the relevant quotes and informative discussion, see the first chapter of Russell and Norvig 2002). As it turned out, the new century would arrive without a single machine able to converse at even the toddler level. (Recall that when it comes to the building of machines capable of displaying human-level intelligence, Descartes, not Turing, seems today to be the better prophet.) Nonetheless, astonishing though it may be, people do continue today to issue incredibly optimistic predictions regarding the progress of AI. For example, Moravec (1999), in his *Robot: Mere Machine to Transcendent Mind*, informs us that because the speed of computer hardware doubles every 18 months (in accordance with Moore’s Law, which has apparently held in the past and shows no sign of failing), “fourth generation” robots will soon enough exceed humans in all respects, from running companies to writing novels. These robots, so the story goes, will evolve to such lofty cognitive heights that we will stand to them as single-cell organisms stand to us today.

Moravec is by no means singularly Pollyannaish: Many others in AI predict the same sensational future unfolding on about the same rapid schedule. In fact, at the 50^{th} anniversary celebration at Dartmouth of the original 1956 AI conference at this university, host and philosopher Jim Moor posed the question “Will human-level AI be achieved within the next 50 years?” to five thinkers who attended the original 1956 conference: John McCarthy, Marvin Minsky, Oliver Selfridge, Ray Solomonoff, and Trenchard Moore. McCarthy and Minsky gave firm, unhesitating affirmatives, and Solomonoff seemed to suggest that AI provided the one ray of hope in the face of the fact that our species seems bent on destroying itself. (Selfridge’s reply was a bit cryptic. Moore returned a firm, unambiguous negative, and declared that once his computer is smart enough to interact with him conversationally about mathematical problems, he might take this whole enterprise more seriously.)

Moor’s question is not just for scientists and engineers; it is also a question for philosophers. This is so for two reasons. One, research and development designed to validate an affirmative answer must include philosophy — for reasons expressed above. (E.g., as section 3 makes plain, philosophy and logic must be turned to if one is to have a hope of capturing human reasoning in computation.) Two, philosophers might be able to provide arguments that answer Moor’s question now, definitively. If any of the strong critiques of AI that we have discussed is fundamentally correct, then of course AI will not manage to produce machines having the mental powers of persons. At any rate, time marches on, and will tell.

References

Baars, B. J.: 1986, *The Cognitive Revolution in Psychology*, Guilford Press.

Bacon, F.: 2002, *The New Organon*, Cambridge University Press. Originally published in 1620.

Block, N.: 1978, Troubles with Functionalism, *in* C. W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology*, Vol. 9 of *Minnesota Studies in the Philosophy of Science*, University of Minnesota Press, pp. 261–325.

Block, N.: 1981, Psychologism and behaviorism, *Philosophical Review* **90**, 5–43.

Braine, M. D. S. and O'Brien, D. P. (eds): 1998, *Mental Logic*, Lawrence Erlbaum Associates.

Bringsjord, S.: 1992, *What Robots Can and Can't Be*, Kluwer, Dordrecht, The Netherlands.

Bringsjord, S.: 1995, Could, how could we tell if, and why should–androids have inner lives?, *in* K. Ford, C. Glymour and P. Hayes (eds), *Android Epistemology*, MIT Press, Cambridge, MA, pp. 93–122.

Bringsjord, S. and Schimanski, B.: 2003, What is artificial intelligence? Psychometric AI as an answer, *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI–03)*, Morgan Kaufmann, San Francisco, CA, pp. 887–893.

Bringsjord, S. and Zenzen, M.: 1997, Cognition Is Not Computation: The Argument from Irreversibility, *Synthese* **113**, 285–320.

Brooks, R. A.: 1991, Intelligence Without Reason, *AI Memo 1293*, MIT AI Lab.

Charniak, E. and McDermott, D.: 1985, *Introduction to Artificial Intelligence*, Addison-Wesley, Reading, MA.

Chisholm, R.: 1966, *Theory of Knowledge*, Prentice-Hall, Englewood Cliffs, NJ.

Chomsky, N.: 1996 [1959], A Review of B. F. Skinner's 'Verbal Behavior', *in* H. Geirsson and M. Losonsky (eds), *Readings in Language and Mind*, Blackwell, pp. 413–441.

Church, A.: 1936, An Unsolvable Problem of Elementary Number Theory, *American Journal of Mathematics* **58**, 345–363.

Cole, D.: 2004, The chinese room argument, *in* E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University. `http://plato.stanford.edu/entries/chinese-room`.

Davis, E. and Morgenstern, L.: 2004, Progress in formal commonsense reasoning, Introduction to the Special Issue on Formalization of Common Sense, *Artificial Intelligence Journal* **153**, 1–12.

Davis, M.: 2001, The early history of automated deduction, *in* A. Robinson and A. Voronkov (eds), *Handbook of Automated Reasoning*, Vol. I, Elsevier Science, chapter 1, pp. 3–15.

Descartes, R.: 1911, *The Philosophical Works of Descartes, Volume 1. Translated by Elizabeth S. Haldane and G.R.T. Ross*, Cambridge University Press, Cambridge, UK.

Descartes, R.: 1988, *Descartes: Selected Philosophical Writings*, Cambridge University Press. Translated by J. Cottingham and R. Stoothoff and D. Murdoch.

Dretske, F.: 1988, *Explaining Behavior: Reason in a World of Causes*, MIT Press.

Dreyfus, H. L.: 1992, *What Computers Still Can't Do: A Critique of Artificial Reason*, MIT Press.

Evans, G.: 1968, A program for the solution of a class of geometric-analogy intelligence-test questions, *in* M. Minsky (ed.), *Semantic Information Processing*, MIT Press, Cambridge, MA, pp. 271–353.

Fodor, J. A.: 1978, Tom Swift and his procedural grandmother, *Cognition* **6**, 229–247.

Fodor, J. A.: 1987, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, MIT Press.

Fodor, J. A.: 1990, *A Theory of Content and Other Essays*, MIT Press.

Fodor, J. A. and Pylyshyn, Z. W.: 1988, Connectionism and Cognitive Architecture: A Critical Analysis, *Cognition* **28**, 139–196.

Gardner, H.: 1985, *The mind's new science: A history of the cognitive revolution*, Basic Books.

Gentzen, G.: 1969, *The collected papers of Gerhard Gentzen*, North-Holland, Amsterdam, Holland. English translations of Gentzen's papers, edited and introduced by M. E. Szabo.

Gould, S. J. and Lewontin, R.: 1979, The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Paradigm, *Proceedings of the Royal Society*, Vol. B205, pp. 581–598.

Hanson, N. R.: 1958, *Patterns of Discovery*, Cambridge University Press.

Harnad, S.: 1990, The Symbol Grounding Problem, *Physica D* **42**, 335–346.

Harnad, S.: 1991, Other bodies, other minds: A machine incarnation of an old philosophical problem, *Minds and Machines* **1**(1), 43–54.

Haugeland, J.: 1985a, *AI: The Very Idea*, MIT Press.

Haugeland, J.: 1985b, *Artificial Intelligence: The Very Idea*, MIT Press, Cambridge, MA.

Hempel, C. G.: 1965, *Aspects of Scientific Explanation*, The Free Press.

Hempel, C. G.: 1985, Thoughts on the Limitations of Discovery by Computer, *Logic of Discovery and Diagnosis in Medicine*, University of California Press.

Johnson-Laird, P. N.: 1977, Procedural Semantics, *Cognition* **5**, 189–214.

Kirsh, D.: 1991, Today the earwig, tomorrow man?, *Artificial Intelligence* **47**, 161–184.

Koza, J.: 1992, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press.

Lakatos, I.: 1976, *Proofs and refutations: the logic of mathematical discovery*, Cambridge University Press.

Langley, P., Simon, H. A., Bradshaw, G. L. and Zytkow, J. M.: 1987, *Scientific Discovery: Computational Explorations of the Creative Process*, MIT Press.

Leahey, T. H.: 1992, The Mythical Revolutions of American Psychology, *American Psychologist* **47**(2), 308–318.

Lenat, D.: 1983, EURISKO: A program that learns new heuristics and domain concepts: the nature of heuristics iii: Program design and results, *Artificial Intelligence* **21**(1 & 2), 61–98.

Lenat, D. B.: 1976, *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search*, PhD thesis, Stanford University.

Lycan, W. G.: 1984, *Logical Form in Natural Language*, MIT Press.

M. Davis and H. Putnam: 1960, A computing procedure for quantification theory, *Journal of the Association for Computing Machinery* **7**(3), 201–215.

S. Jáskowski: 1934, On the rules of suppositions in formal logic, *Studia Logica* **1**.

McCarthy, J.: 1962, Computer programs for checking mathematical proofs, *Proceedings of the Symposium in Pure Math, Recursive Function Theory*, Vol. V, American Mathematical Society, Providence, RI, pp. 219–228.

McCulloch, W. S. and Pitts, W. A.: 1943, A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics* **5**, 115–133.

Mill, J. S.: 1874, *System of logic*, eight edn, Harper & Brothers, New York.

Millikan, R. G.: 1993, *White Queen Psychology*, MIT Press.

Minsky, M.: 1986, *The Society of Mind*, Simon and Schuster.

Moravec, H.: 1999, *Robot: Mere Machine to Transcendant Mind*, Oxford University Press, Oxford, UK.

Muggleton, S.: 1992, Inductive logic programming, *in* S. Muggleton (ed.), *Inductive Logic Programming*, Academic Press, London, pp. 3–27.

Newell, A. and Simon, H. A.: 1976, Computer Science as Empirical Inquiry: Symbols and Search, *Communications of the ACM* **19**, 113–126.

Osherson, D. N.: 1975, *Logical Abilities in Children, volume 3, Reasoning in Adolesescence: Deductive Inference*, Lawrence Erlbaum Associates.

Peirce, C.: 1960, *The collected papers of C. S. Peirce*, Harvard University Press.

Pelletier, F. J.: 1999, A Brief History of Natural Deduction, *History and Philosophy of Logic* **20**, 1–31.

Pelletier, F., Sutcliffe, G. and Suttner, C. B.: 2002, The Development of CASC, *AI Communications* **15**(2-3), 79–90.

Pollock, J. L.: 1992, How to reason defeasibly, *Artificial Intelligence* **57**(1), 1–42.

Pólya, G.: 1945, *How to solve it*, Princeton University Press.

Port, R. and Gelder, T. V.: 1995, *Mind as Motion: Explorations in the Dynamics of Cognition*, MIT Press.

Putnam, H.: 1960, Minds and machines, *in* S. Hook (ed.), *Dimensions of Mind*, New York University Press, pp. 138–164.

Putnam, H.: 1988, *Representation and Reality*, MIT Press.

Pylyshyn, Z.: 1989, Computing in Cognitive Science, *in* M. Posner (ed.), *Foundations of Cognitive Science*, MIT Press.

Pylyshyn, Z.: 1991, Rules and Representations: Chomsky and Representational Realism, *in* A. Kasher (ed.), *The Chomskyan Turn*, Blackwell, pp. 231–251.

Reichenbach, H.: 1938, *Experience and Prediction*, University of Chicago Press.

Rips, L. J.: 1994, *The Psychology of Proof*, MIT Press.

Robinson, A. and Voronkov, A. (eds): 2001, *Handbook of Automated Reasoning*, Vol. 1, North-Holland.

Robinson, J. A.: 1965, A machine-oriented logic based on the resolution principle, *Journal of the ACM* **12**, 23–41.

Rumelhart, D. E. and McClelland, J. L.: 1986, *Parallel Distributed Processing*, MIT Press.

Russell, B.: 1940, *An inquiry into meaning and truth*, George Allen and Unwin.

Russell, S. and Norvig, P.: 2002, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, NJ.

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

Searle, J.: 1984, *Minds, Brains, and Science*, Harvard University Press, Cambridge, MA.

Siegelmann, H. T. and Sontag, E. D.: 1994, Analog Computation via Neural Networks, *Theoretical Computer Science* **131**, 331–360.

Sperling, G.: 1960, The information available in brief visual presentations, *Psychological Monographs* **74**(11).

Thompson, E. P.: 1963, *The Making of the English Working Class*, Victor Gollancz, London.

Tolman, E. C.: 1948, Cognitive maps in rats and men, *Psychological Review* **55**, 189–208.

Turing, A.: 1950, Computing machinery and intelligence, *Mind* **LIX (59)**(236), 433–460.

Turing, A. M.: 1936, On Computable Numbers with Applications to the Entscheidungsproblem, *Proceedings of the London Mathematical Society* **42**, 230–265.

Voronkov, A.: 1995, The anatomy of Vampire: implementing bottom-up procedures with code trees, *Journal of Automated Reasoning* **15**(2).

Wang, H.: 1960, Toward mechanical mathematics, *IBM Journal of Research and Development* **4**, 2–22.

Weidenbach, C.: 2001, Combining superposition, sorts, and splitting, *in* A. Robinson and A. Voronkov (eds), *Handbook of Automated Reasoning*, Vol. 2, North-Holland.

Winograd, T.: 1990, Thinking machines: Can there be? Are we?, *in* D. Partridge and Y. Wilks (eds), *The Foundations of Artificial Intelligence*, Cambridge University Press, pp. 167–189.

Wos, L., Overbeek, R., Lusk, E. and Boyle, J.: 1992, *Automated Reasoning, Introduction and Applications*, McGraw-Hill, Inc.