Peirce and the Philosophy of Artificial Intelligence

James Fetzer

The Commens Encyclopedia

The Digital Encyclopedia of Peirce Studies

New Edition

Edited by Mats Bergman and João Queiroz

URL	http://www.commens.org/encyclopedia/article/fetzer-james-peirce-and-philosophy-artificial-intelligence
Retrieved	29.05.2026
ISSN	2342-4257
License	Creative Commons Attribution-NonCommercial-ShareAlike

Abstract:

A philosophical appraisal of historical positions on the nature of thought, mentality, and intelligence, this survey begins with the views of Descartes, Turing, and Newell and Simon, but includes the work of Haugeland, Fodor, Searle, and other major scholars. The underlying issues concern distinctions between syntax, semantics, and pragmatics, where physical computers seem to be best viewed as mark-manipulating or syntax-processing mechanisms. Alternative accounts have been advanced of what it takes to be a thinking thing, including being Turing machines, symbol systems, semantic engines, and semiotic systems, which have the ability to use signs in the sense of the Charles S. Peirce. Reflections regarding the nature of representations and existence of mental algorithms suggest that the theory of minds as semiotic systems should be preferred to its alternatives, where digital computers can still qualify as "intelligent machines" even without minds.

Keywords: Thinking Things, Symbol Systems, Semantic Engines, Semiotic Systems, Sign, Mind, Consciousness, Cognition

Historical Background

Prior to the advent of computing machines, theorizing about the nature of mentality and thought was predominantly the province of philosophers, among whom perhaps the most influential historically has been Rene´ Descartes (1596-1650), commonly called "the father of modern philosophy". Descartes advanced an *ontic* (or ontological) thesis about the kind of thing minds are as features of the world and an *epistemic* (or epistemological) thesis about how things of that kind could be known. According to Descartes, who advocated a form of dualism for which mind and body are mutually exclusive categories, "minds" are things that can think, where access to minds can be secured by means of a faculty known as "introspection", which is a kind of inward perception of a person's own mental states.

Descartes' approach exerted enormous influence well into the 20th century, when the development of digital computers began to captivate the imagination of those who sought a more scientific and less subjective conception of the nature of thinking things. The most important innovations were introduced by Alan Turing (1912-54), a British mathematician whose (posthumous) entitlement to the title of "the father of computer science" and even of artificial intelligence (or "AI") appears difficult to deny. Turing's

most important research concerned the limitations of proof within mathematics, where he proposed that the boundaries of the computable (of mathematical problems whose solutions were obtainable on the basis of finite applications of logical rules) were the same as those that can be solved using a specific kind of problem-solving machinery.

Things of this kind, which are known as *Turing machines*, consist of an arbitrarily long segmented tape and a device capable of four operations upon that tape, namely: making a mark, removing a mark, moving the tape one segment forward, and moving the tape one segment backward. (The state of the tape before a series of operations is applied can be referred to as "input", the state of the tape after it has been applied as "output", and the series of instructions as a "program".) From the perspective of these machines, it became obvious that there are mathematical problems for which no finite or computable solutions exist. Similar results that relate effective procedures to computable problems were concurrently obtained by the American logician, Alonzo Church.

The Turing Test

Church's work was based on purely mathematical assumptions, while Turing's work appealed to a very specific kind of machine, which provided an abstract model for the physical embodiment of the procedures that suitably define "(digital) computers" and laid the foundation for the theory of computing. Turing argued that such procedures impose limits upon human thought, thereby combining the concept of a program with that of a mind in the form of a machine which in principle could be capable of having many types of physical implementation. His work thus introduced what has come to be known as *the computational conception of the mind*, which inverts the Cartesian account of machines as mindless by turning minds themselves into special kinds of machines, where the boundaries of computability define the boundaries of thought.

Turing's claim to have fathered AI rests upon the introduction of what is known as *the Turing test*, where a thing or things of one kind are pitted against a thing or things of another kind. Adapting a party game where a man and a woman might compete to see whether the man could deceive a contestant into mistaking him for the woman (in a context that would not give the game away), he proposed pitting a human being against an inanimate machine (equipped with a suitable program and mode of communication). Thus, if an interlocutor could not differentiate between them on the basis of the answers they provided to questions that they were asked, then those systems should be regarded as equal (or equipotent) with respect to (what he took to be) *intelligence* (Turing, 1950).

This represented a remarkable advance over Cartesian conceptions in three different respects. First, it improved upon the vague notion of a thinking thing by introducing the precise notion of a Turing machine as a device capable of mark manipulation under the control of a program. Second, it implied a solution to the mind/body problem according to which hardware is to software as bodies are to minds that was less metaphorical and more scientific than the notion of bodies with minds. Third, it appealed to a behavioral criterion in lieu of introspection for empirical evidence supporting inference to the existence of thinking things, which made the study of the mind appear much less subjective.

Physical Machines

Thus, Descartes' conception of human minds as thinking things depends upon actually having thoughts, which might not be the case when they are unconscious (say, asleep, drugged, or otherwise incapable of thought), since their existence as things that think would not then be subject to introspective verification, and supports hypothesis (h1):

(h1) (Conscious) human minds are thinking things (Descartes);

Analogously, Turing's conception of these machines as thinking things depends upon the exercise of the capacity to manipulate marks as a sufficient condition for the possession of intelligence, which could be compared with that of humans, suggesting hypothesis (h2):

(h2) Turing machines manipulating marks possess intelligence (Turing);

where the identification of *intelligence* with *mentality* offers support for the conclusion that suitably programmed and properly functioning Turing machines might qualify as man-made thinking things or, in the phrase of John McCarthy, as "artificial intelligence".

As idealized devices that are endowed with properties physical systems may not possess, including segmented tapes (or "memories") of arbitrary length and perfection in performance, however, Turing machines are abstract entities. Because they do not exist in space/time, they are incapable of exerting any causal influence upon things in space/time, even though, by definition, they perform exactly as intended (Fetzer, 1988). The distinction is analogous to that between numbers and numerals, where numbers are abstract entities that do not exist in space/time, while numerals that stand for them are physical things that do exist in space/time. Roman numerals, Arabic numerals, and such have specific locations at specific times, specific shapes and sizes, come into and go out of existence, none which is true of numbers as timeless and unchanging abstract entities.

These "machines", nevertheless, might be subject to at least partial implementations as physical things in different ways employing different materials, such as by means of digital sequences of 0s and 1s, of switches that are "on" or "off", or of higher and lower voltage. Some might be constructed out of vacuum tubes, others transistors, and others silicon chips. They then become instances of physical things with the finite properties of things of their kinds. None of them performs exactly as intended merely as a matter of definition: all of them have the potential for malfunction and variable performance like aircraft, automobiles, television sets, and other physical devices. Their memories are determined by specific physical properties, such as the size of their registers; and, while they may be enhanced by the addition of more memory, none of them is infinite.

Symbol Systems

While (some conceptions of) God might be advanced as exemplifying a timeless and unchanging thinking thing, the existence of entities of that kind falls beyond the scope of empirical and scientific inquiries. Indeed, within computer science, the most widely accepted and broadly influential adaptation of Turing's approach has been by means of *the physical symbol system conception* Alan Newell and Herbert Simon have advanced, where symbol systems are physical machines–possibly human–that process physical symbol structures through time (Newell and Simon, 1976). These are special kinds of digital machines that qualify as serial processing (or von Neumann) machines. Thus, they implement Turing's conception by means of a physical machine hypothesis (h3),

(h3) Physical computers manipulating symbols are intelligent (Newell and Simon);

where, as for Turing, the phrase "intelligent thing" means the same as "thinking thing".

There is an ambiguity about the words "symbol systems" as systems that process symbols and as the systems of symbols which they process, where Newell and Simon focused more attention on the systems of symbols that machines process than they did upon the systems that process those symbols. But there can be no doubt that they took for granted that the systems that processed those symbols were physical. It therefore becomes important, from this point hence, to distinguish between "Turing machines" as abstract entities and "digital computers" as physical implementations of such machines, where digital computers, but not Turing machines, possess finite memories and potential to malfunction. Newell and Simon focused upon computers as physical machines, where they sought to clarify the status of the "marks" that computers subject to manipulation.

They interpreted them as sets of physical patterns they called "symbols", which can

occur in components of other patterns they called "expressions" (or "symbol structures"). Relative to sets of alphanumerical (alphabetical and numerical) characters (ASCII or EBCDIC, for example), expressions are sequences of symbols understood as sequences of characters. Their "symbol systems" as physical machines that manipulate symbols thus qualify as necessary and sufficient for intelligence, as formulated by hypothesis (h4):

> (h4) (Being a) symbol system is both necessary and sufficient for intelligence (Newell and Simon);

which, even apart from the difference between Turing machines as abstractions and symbol systems as physical things, turns out to be a much stronger claim than (h2) or even (h3). Those hypotheses do not imply that every thinking thing has to be a digital computer or a Turing machine. (h2) and (h3) are both consistent with the existence of thinking things that are not digital computers or Turing machines. But (h4) does not allow for the existence of thinking things that are not digital machines.

The Chinese Room

The progression of hypotheses from (h1) to (h2) to (h3) and perhaps (h4) appears to provide significant improvement on Descartes' conception, especially when combined with the Turing test, since they not only clarify the nature of mind and elucidate the relation of mind to body, but even explain how the existence of other minds might be known, a powerful combination of ontic and epistemic theses that seems to support the prospects for artificial intelligence. As soon as computing machines were designed with performance capabilities comparable to those of human beings, it would be appropriate to ascribe to those inanimate entities the mental properties of thinking things. Or so it seemed, when the philosopher John Searle advanced a critique of the prospects for AI that has come to be known as "the Chinese Room" and cast it all in doubt (Searle, 1980).

Searle proposed a thought experiment involving two persons, call them "C" and "D", one (C) fluent in Chinese, the other (D) not. Suppose C were locked in an enclosed room into which sequences of marks were sent on pieces of paper, to which C might respond by sending out other sequences of marks on other pieces of paper. If the marks sent in were questions in Chinese and the marks sent out were answers in Chinese, then it would certainly look as though the occupant of the room knew Chinese, as, indeed, by hypothesis, he does. But suppose instead D were locked in the same room with a table that allowed him to look up sequences of marks to send out in response to sequences of marks sent in. If he were very proficient at this activity, his performance might be the

equal of that of C, who knows Chinese, even though D, by hypothesis, knows no Chinese.

Searle's argument was a devastating counterexample to the Turing test, which takes for granted that similarities in performance indicate similarities in intelligence. In the Chinese Room scenario, the same "inputs" yield the same "outputs", yet the processes or procedures that produce them are not the same. This suggests that a distinction has to be drawn between "simulations", where systems *simulate* one another when they yield the same outputs from the same inputs, and "replications", where systems replicate one another when they yield the same outputs from the same inputs by means of the same processes or procedures. In this language, Searle shows that, even if the Turing test is sufficient for comparisons of input/output behavior (simulations), it is not sufficient for comparisons of the processes or procedures that yield those outputs (replications).

Weak AI

The force of Searle's critique becomes apparent in asking which scenario, C or D, is more like the performance of a computer executing a program, which might be implemented as an automated look-up table: in response to inputs in the form of sequences of marks, a computer processes them into outputs in the form of other sequences of marks on the basis of its program. So it appears appropriate to extend the comparison to yet a third scenario, call it "E", where a suitably-programmed computer takes the same inputs and yields the same outputs. For just as the performance of D might simulate the performance of C, even though D knows no Chinese, so the performance of E might simulate the performance of D, even though E possesses no mentality. Mere relations of simulation thus appear too weak to establish that systems are equal relative to their intelligence.

Searle also differentiated between what he called "strong AI" and "weak AI", where *weak AI* maintains that computers are useful tools in the study of the mind, especially in producing useful models (or simulations), but *strong AI* maintains that, when they are executing programs, computers properly qualify as minds (or replications). Weak AI thus represents an epistemic stance about the value of computer-based models or simulations, while strong AI represents an ontic stance about the kinds of things that actually are instances of minds. Presumably, strong AI implies weak AI, since actual instances of minds would be suitable subjects in the study of mind. Practically no one objects to weak AI, of course, while strong AI remains controversial on many grounds.

That does not mean it lacks for passionate advocates. One of the most interesting

introductions to artificial intelligence has been co-authored by Eugene Charniak and Drew McDermott (Charniak & McDermott, 1985). Already in their first chapter, the authors define "artificial intelligence" as the study of mental faculties through the use of computational models. The tenability of this position, no doubt, depends upon the implied premise that mental faculties operate on the basis of computational processes, which, indeed, they render explicit by similarly postulating that what brains do "may be thought of at some level as a kind of computation" (Charniak & McDermott, 1985, p. 6). The crucial distinction between "weak" and "strong" AI, however, depends upon whether brains actually qualify as computers, not whether they may be thought to be.

Strong AI

They go further in maintaining that "the ultimate goal of research in AI is to build a person or, more humbly, an animal". Their general conception is that the construction of these artificial things must capture key properties of their biological counterparts, at least with respect to kinds of input, kinds of processing, and kinds of output. Thus, the "inputs" they consider include vision (sights) and speech (sounds), which are processed by means of internal modules for learning, deduction, explanation, and planning, which entail search and sort mechanisms. These combine with speech and motor capabilities to yield "outputs" in the form of speech (sounds) and behavior (motions), sometimes called "robotics". The crucial issue thus becomes whether these "robots" are behaving like human beings as (mindless) simulations or instead embody (mindful) replications.

Their attention focuses upon what goes on in "the black box" between stimulus and response, where those with minds depend upon and utilize *internal representations* as states of such systems that describe or otherwise represent various aspects of the world. Indeed, some of these aspects could be internal to the system itself and thus represent its own internal states as internal representations of aspects of itself. But, while self-awareness and self-consciousness are often taken to be important kinds of intelligence or mentality, they do not appear to be essential to having intelligence or mentality in general as opposed to having intelligence or mentality of specific kinds. There may be various kinds of mentality or intelligence–mathematical, verbal, and artistic, for example–but presumably they share certain core or common properties.

There would seem to be scant room for doubt that, if artificial machines are going qualify as comparable to human beings relative to their mental abilities, they must have the same or similar capacities to use and manipulate internal representations, at least with respect to some specified range–presumably, alphanumeric–of tasks. They

must take the same or similar external inputs (or "stimuli"), processes them by means of the same or similar "mental" mechanisms, and produce the same or similar external outputs (or "responses"). While Charniak & McDermott may aspire to build an artificial animal, the AI community at large, no doubt, would settle for building an artificial thinking thing, presuming that it is possible to create one without the other.

Folk Psychology

There is an implied presumption that different systems that are subject to comparison are operating under the same or similar causally-relevant background conditions. No one would suppose that a computer with a blown mother board should yield the same outputs from the same inputs as a comparable computer with no hardware breakdown, even when they are loaded with the same programs. Analogously, no one would assume that a human being with a broken arm, for example, should display the same behavior in response to the same stimuli (say, a ball coming straight toward him while seated in the bleachers at a game) as another person without a broken arm. But that does not mean that they are not processing similar stimuli by means of similar representations.

Human beings are complicated mechanisms, whether or not they properly qualify as "machines" in the sense that matters to AI. Indeed, the full range of causally-relevant factors that make a difference to human behavior appears to include motives, beliefs, ethics, abilities, capabilities, and opportunities (Fetzer, 1996). Different persons with the same or similar motives and beliefs, for example, but who differ in their morals, may be expected to display different behavior under conditions where ethics makes a difference, even though they may have similar abilities and are not incapacitated from the exercise of those abilities. As we all know, human beings consume endless hours endeavoring to explain and predict the behavior of others and themselves employing a framework of causally-relevant factors of this kind, which has come to be known as "folk psychology".

No doubt, when appraised from the perspective of, say, the conditions of adequacy for scientific theories–such as clarity and precision of language, scope of application for explanation and prediction, degree of empirical support, and the economy, simplicity, or elegance with which these results are attained–folk psychology appears to enjoy a high degree of empirical support by virtue of its capacity to subsume a broad range of cases within the scope of its principles. Some of that apparent success, however, may be due to the somewhat vague and imprecise character of the language upon which it depends, where there would appear to be opportunity for revision and refinement to enhance or

confine its scope of application. Yet some students argue for its elimination altogether.

Eliminative Materialism

Paul Churchland, for example, maintains that folk psychology is not only incomplete but also inaccurate as a "misrepresentation" of our internal states and mental activities. He goes so far as to suggest that progress in neuroscience should lead, not simply to the refinement of folk psychology, but to its wholesale elimination (Churchland, 1984, p. 43). The model Churchland embraces thus follows the pattern of elimination of "phlogiston" from the language of chemistry and of "witches" from the language of psychology. He thus contends that the categories of *motives* and *beliefs*, among others, are destined for a similar fate as neuroscience develops. Churchland admits he cannot guarantee that this will occur, where the history of science in this instance might instead simply reflect some adjustment in folk-psychological principles or dispensing with some of its concepts.

The deeper problem that confronts eliminative materialism, however, appears to be the same problem confronting classic forms of reductionism, namely, that without access to information relating brain states to mind states, on the one hand, and mind states to behavioral effects, on the other, it would be impossible to derive predictive inferences from brain states to behavioral effects. If those behavioral effects are manifestations of dispositions toward behavior under specific conditions, moreover, then it seems unlikely that a "mature" neuroscience could accomplish its goals if it lacked the capacity to relate brain states to behavioral effects by way of dispositions, because there would then be no foundation for relating mind states to brain states and brain states to human behavior.

In the case of jealousy (hostility, insincerity, and so on) as causal factors that affect our behavior in the folk-psychological scheme of things, if we want to discover the brain states that underlie these mind states as dispositions to act jealous (to act hostile, and so forth) under specific conditions, which include our other internal states, then a rigorous science of human behavior might be developed by searching for and discovering some underlying brain states, where those dispositions toward behavior were appropriately (presumably, lawfully) related to those brain states. Sometimes brain states can have effects upon human behavior that are not mediated by mind states, as in the case of brain damage or mental retardation. For neurologically normal subjects, mind states are able to establish connections between brain states and their influence on behavior.

Processing Syntax

The predominant approach among philosophers eager to exploit the resources provided by the computational conception, however, has been in the direction of refining what it takes to have a mind rather than the relationship between minds, bodies, and behavior. While acknowledging these connections are essential to the adequacy of any account, they have focused primarily upon the prospect that language and mentality might be adequately characterized on the basis of purely formal distinctions of the general kind required by Turing machines–the physical shapes, sizes, and relative locations of the marks they manipulate–when interpreted as the alphanumeric characters that make up words, sentences, and other combinations of sentences as elements of a language.

Jerry Fodor, for example, has observed that computational conceptions of language and mentality entail the thesis that, "... mental processes have access only to formal (nonsemantic) properties of the mental representations over which they are defined" (Fodor, 1980, p. 307). He elaborates upon the relationship between the form (syntax) and the content (semantics) of thoughts, maintaining (a) that thoughts are distinct in content only if they can be identified with distinct representations, but without offering an explanation of how it is (b) that any specific thoughts can be identified with any specific representations, a problem for which he elsewhere offers a solution known as "the language of thought". But any account maintaining that the same syntax always has the same semantics or that the same semantics always has the same syntax runs afoul of problems with ambiguity, on the one hand, and with synonymy, on the other.

Nevertheless, the strongest versions of computational conceptions tend to eschew concern for semantics and focus instead on the centrality of syntax. Stephen Stich has introduced *the syntactic theory of the mind (STM)* as having an agnostic position on content, neither insisting that syntactic state types (as repeatable patterns of syntax) have no content nor insisting that syntactic state tokens (specific instances of syntactic state types) have no content: "It is simply silent on the whole matter... (T)he STM is in effect claiming that psychological theories have no need to postulate content or other semantic properties" (Stich, 1983, p. 186). STM is thereby committed to hypothesis (h7):

(h5) Physical computers processing syntax possess minds (STM);

which may initially appear much stronger than (h3). But Newell and Simon's notion of "symbol" is defined formally and their "symbol systems" are also computing machines. Both approaches run the risk of identifying "thinking things" with mindless machines.

Semantic Engines

Systems of marks with rules for their manipulation are examples of (what are known as) *formal systems*, the study of which falls within the domain of pure mathematics. When those formal systems are subject to interpretations, especially with respect to properties and objects within the physical world, their study falls within the domain of applied mathematics. A debate has raged within computer science over whether that discipline should model itself after pure mathematics or applied (Colburn et al., 1993). But whatever the merits of the sides to that dispute, there can be scant room for doubt that mere mark manipulation, even in the guise of syntax processing, is not enough for thinking things. Thoughts possess content as well as form, where it is no stretch of the imagination to suggest that, regarding thought, content dominates form.

The STM, which makes syntax-processing sufficient for the possession of mentality, thus appears to be far too strong, but a weaker version might still be true. The ability to process syntax might be necessary for mentality instead, as, indeed, hypothesis (h3) implies, when Newell and Simon's "symbols" are properly understood as marks subject to manipulation. Thus, a more plausible version of (h5) should maintain instead (h6):

(h6) (Conscious) minds are physical computers processing syntax;

where syntax consists of marks and rules for their manipulation that satisfy constraints that make them meaningful. But since there are infinitely many possible interpretations of any finite sequence of marks, some specific interpretation (or class of interpretations) requires specification as "the intended interpretation". Marks can only qualify as syntax relative to specific interpretations in relation to which those marks become meaningful.

From this point of view, a (properly functioning) computing machine can be qualified as *an automatic formal system* when it is executing a program, but becomes meaningful only when its syntax satisfies the constraints of an intended interpretation. Indeed, an automatic formal system where "the semantics follows the syntax" has been designated "a semantic engine" by Daniel Dennett. This supports the contention some have called the basic idea of cognitive science–*that intelligent beings are semantic engines*, that is, automatic formal systems under which they consistently make sense (Haugeland 1981, p. 31). (h6) thus requires qualification to incorporate the role of interpretation as (h7):

(h7) Semantic engines are necessary and sufficient for intelligence;

where, as in the case of Newell and Simon, "intelligent things" are also "thinking things" and "(conscious) minds", understood as physical computers processing syntax under an interpretation. The problem is to "pair up" the syntax and the semantics the

right way.

Cognitive Science

John Haugeland has emphasized that the conception of (conscious) minds as semantic engines places cognitive psychology and artificial intelligence more or less on an equal footing, where human beings and intelligent machines turn out to be simply different manifestations of the same underlying phenomena. Indeed, he notes, we can see why from this perspective, artificial intelligence can be regarded as psychology in a particularly pure and abstract form. The same fundamental structures are under investigation, but in AI, all the relevant parameters are under direct experimental control (in the programming), without any messy physiology or ethics to get in the way (Haugeland, 1981, p. 31).

This complements the computational conception of language and the mind according to which thinking is reasoning, reasoning is reckoning, reckoning is computation, and the boundaries of computability define the boundaries of thought, as Turing envisioned.

Haugeland submits there are at least two strategies for attacking cognitive science, thus understood, where (what he calls) "the hollow shell" strategy maintains that, no matter how successfully a semantic engine performs, it still lacks understanding; and (what he calls) "the poor substitute" strategy maintains that mere semantic engines are never going to perform as well as human beings. The second thus contends that no semantic engine will ever be capable of passing the Turing test, whereas the first maintains that, even if one did, it would still not properly qualify as a thinking thing. Given the distinction between simulation and replication, the second contends that no semantic engine will ever equal human beings in its capacity for simulation, and the first maintains that semantic engines will never attain equivalence as replications.

It should come as no surprise when digital machines turn out to be semantic engines for which their semantics follows their syntax, since programmers design programs to insure precisely that result. Thus, it would be more appropriate to suggest that, in this case at least, the syntax follows the semantics, since the marks that they employ and the operations that are performed upon them are intended for precisely that purpose. Not all programmers are equally adept at bring about such results, of course, as the process of compiling and executing programs–including running and debugging them–displays. In the case of human beings, however, the situation is more complex, since it would be inappropriate within a "science of cognition" to appeal to God as an intelligent designer, which suggests that human beings become semantic engines by evolution or

by learning.

The Language of Thought

Jerry Fodor (1975) has advanced an argument hypothesizing the existence of an innate language, which is species-specific and possessed by every neurologically normal human being. He calls it mentalese (or "the language of thought"). He contends the only way to learn a language is to learn the truth conditions for sentences that occur in that language: "... learning (a language) L involves learning that *'Px' is true if and only if x is G* for all substitution instances. But notice that learning that could be learning P (learning what P means) only for an organism that already understood G" (Fodor, 1975, p. 80). Given the unpalatable choice between an *endless hierarchy* of successively richer and richer meta-languages for specifying the meaning of lower level languages and a *base language* that is unlearned, Fodor opts for the existence of an innate and inborn language of thought.

The process of relating a learned language to the language of thought turns human beings into semantic engines, which may be rendered by hypothesis (h8) as follows:

(h8) Human beings are semantic engines with a language of thought (Fodor).

Fodor commits a mistake in his argument, however, by overlooking the possibility that the kind of prior understanding which is presupposed by language learning might be *non-linguistic*. Children learn to suck nipples, play with balls, and draw with crayons long before they know that what they are doing involves "nipples", "balls", or "crayons". Through a process of interaction with things of those kinds, they acquire habits of action and habits of mind concerning the actual and potential behavior of things of those kinds. Habits of action and habits of mind that obtain for various kinds of things are concepts. Once that non-linguistic understanding has been acquired, the acquisition of linguistic dispositions to describe them appears to be relatively unproblematical (Fetzer 1990).

One of the remarkable features of Fodor's conception is that the innate and inborn language of thought possesses a semantic richness such that this base language has to be sufficiently complete to sustain correlations between any natural language (French, German, Swahili, and such) at any stage of historical development (past, present, and future). This means that mentalese not only has to supply a foundation for everyday words, such as "nipple", "ball", and "crayon" in English, for example, but also those for more advanced notions, such as "jet propulsion", "polio vaccine", and "color television", since otherwise the language of thought could not fulfill its intended role. Among the less plausible consequences of this conception turn out to be that, since every human

has the same innate language, which has to be complete in each of its instantiations, unsuccessful translations between different languages and the evolution of language across time are both impossible, in principle, which are difficult positions to defend.

Formal Systems

Fodor's approach represents an extension of the work of Noam Chomsky, who has long championed the conception of an innate syntax, both inborn and species-specific, to which Fodor has added a semantics. Much of Chomsky's work has been predicated upon a distinction between *competence* and *performance*, where differences between the grammatical behavior of different language users, which would otherwise be the same, must be accounted for by circumstantial differences, say, in physiological states or psychological context. In principle, every user of language possesses what might be described as *(unlimited) computational competence*, where infinitely many sentences can be constructed from a finite base by employing recursive procedures of the kind that were studied by Church and Turing in their classic work on effective procedures.

Fodor and Zenon Plyshyn (1988), for example, adopt conditions for the production of sentences by language users implying that the semantic content of syntactic wholes is a function of the semantic content of their syntactic parts as a *principle of the compositionality of meaning* and that molecular representations are functions of other molecular or atomic representations as a *principle of recursive generability*. Such conditions are obvious corollaries of distinctions between structurally atomic and structurally molecular representations as a precondition for a language of thought modeled on formal systems, such as sentential calculus. The principles of formal systems–automated or not–however, may or may not transfer from abstract to physical contexts, not least of all because physical systems, including digital machines, are limited in their capacities.

Turing machines with infinite tapes and infallible performance are clearly abstract idealizations compared to digital machines with finite memories that can malfunction. The physical properties of persons and computers are decidedly different than those of automated formal systems as another case of abstract idealization. By comparison, digital machines and human beings appear to possess no more than *(limited) computational competence* (Fetzer, 1992). Thus, the properties of formal systems, such as incompleteness proofs in higher order logic established by Kurt Godel, which might be supposed to impose limits on mental processes and have attracted the interest of scholars, including J. R. Lucas (1961) and Douglas Hofstadter (1979), appear to have

slight relevance to understanding the nature of cognition. Formal systems are useful in modeling reasoning, but reasoning is a special case of thinking. And if we want to understand the nature of thinking, we have to study thinking things rather than the properties of formal systems. Thinking things and formal systems are not the same.

Connectionism

The acquisition of different habits of action and habits of mind as an effect of our own life histories thus tends to determine our capacity to subsume various experiences by means of corresponding concepts. Every neurologically normal human being may have the same innate potential to learn language and acquire other mental abilities, yet only actually acquire them under specific environmental conditions. Fodor accommodates language learning but precludes partial languages and linguistic evolution. In order for a meta-language to have the resources necessary to subsume any lower level languages, arbitrarily selected, it would have to have the resources to subsume them all. And, by the same token, an innate language can only accommodate learning arbitrarily selected alternative languages if it possesses the (innate and inborn) resources to learn them all.

An approach known as *connectionism* attempts to understand the mind by means of more realistic models of the brain, which can accommodate both learning and evolution (Rumelhart et al., 1986). The number of neurons in the human brain appears to be on the order of 1012 (or one trillion), where the number of connections that those neurons can establish with other neurons is approximately 1,000. This means there are 1015 (or one quadrillion) possible states of the brain defined by distinct arrangements of patterns of activation. The activation of some of these patterns is supposed to bring about the activation of others as innate properties no normal brain could be without, while others may come about as acquired properties normal brains could be without. The capacity to learn a language (within the range of humanly learnable languages), for example, is inborn, while actually knowing French, German, or Swahili is acquired.

Some brain states interact with other internal states to bring about transitions to successive brain states, while others activate specific sequences of motor behaviors, which may include speech. The factors that affect these effects include every property that makes a difference to these transitions, which constitute its neural context. Our interest in these patterns of activation derives from their causal contribution to speech and other behavior, as noted above. Motives and beliefs are variables for the energizing and directive factors that influence behavior and speech, while ethics refers to a special subset that tends to inhibit certain kinds of behavior and speech under suitable

conditions. Certain abilities may have specific modular neural locations.

Mental Propensities

Roger Penrose has suggested that thinking may be a quantum phenomenon and thereby qualify as *non-algorithmic* (Penrose, 1989, pp. 437-439). The importance of this prospect is that algorithms are commonly understood as functions that map single values within some domain onto single values within some range. If mental processes are algorithmic (functions), then they must be *deterministic*, in the sense that the same mental-state cause (completely specified) invariably brings about the same mental-state effect or behavioral response. Since quantum phenomena are not deterministic, if mental phenomena are quantum processes, they are not functions–not even partial functions, for which, when single values within a domain happen to be specified, there exist single values in the corresponding range, but where some of the values in the domain and range of the relevant variables might not be specified.

Systems for which the presence or the absence of every property that makes a difference to an outcome is completely specified are said to be "closed", while those for which the presence or absence of some properties that make a difference to the outcome are unspecified are said to be "open". The distinction between deterministic and (in this case) probabilistic causation is that, for closed systems, for *deterministic* causal processes, the same cause (or complete set of conditions) invariably (or with universal strength u) brings about the same effect, whereas for *probabilistic* causal processes, the same cause variably (with probabilistic strength p) brings about one or another effect within the same fixed class of possible outcomes. A polonium218 atom, for example, has a probabililty for decay during a 3.05 minute interval of 1/2.

The determination that a system, such as an atom of polonium218, is or is not a closed system, of course, poses difficult epistemic problems, which are compounded in the case of human beings, precisely because they are vastly more complex causal systems. Moreover, probabilistic systems have to be distinguished from (what are called) *chaotic systems*, which are deterministic systems with "acute sensitivity to initial conditions", where the slightest change to those conditions can bring about previously unexpected effects. A tiny difference in hundreds of thousands of lines of code controlling a space probe, for example, consisting of the occurrence of only one wrong character, a single misplaced comma, caused Mariner 1, the first United States' interplanetary spacecraft, to veer off course and then have to be destroyed.

The Frame Problem

Indeed, there appear to be at least three contexts in which probabilistic causation may matter to human behavior, namely: in processing sensory data into patterns of neural activation; in transitions between one pattern of activation and another; and in producing sounds and other movement as a behavioral response. Processes of all three kinds might be governed by probabilistic or by chaotic deterministic processes and therefore be more difficult to explain or predict, even when the kind of system under consideration happens to be known. These concerns also arise in the context of the study of mental models or representations of the world, specifically, what has be known as the *frame problem*, which Charniak and McDermott describe as follows:

> The need to infer explicitly that a state will not change across time is called the frame problem. It is a problem because almost all states fail to change during an event, and in practical systems there will be an enormous number of them, which it is impractical to deal with explicitly. This large set forms a "frame" within which a small number of changes occur, hence the phrase. (Charniak and McDermott, 1985, p. 418)

While the frame problem has proven amenable to many different characterizations–a variety of which may be found, for example, in Ford and Hayes (1991)–one important aspect of the problem is the extent to which a knowledge base permits the prediction and the explanation of systems when those systems are not known to be open or close.

Indeed, from this point of view, the frame problem even appears to instantiate the classic *problem of induction* encountered in attempting to predict the future based upon information about the past identified by David Hume (1711-76), a Scottish philosopher of considerable influence. Hume observed that there are no deductive guarantees that the future will resemble the past, since it remains logically possible that, no matter how uniformly the occurrence of events of one kind have been associated with events of another, they may not continue to be. If the laws of nature persist through time, however, then, in the case of systems that are closed, it should be possible to predict–invariably or probabilistically–precisely how those systems will behave over intervals of time t* - t so long as the complete initial conditions and laws of systems of that kind are known.

Minds and Brains

Because connectionism appeals to patterns of activation of neural nodes rather than to individual nodes as features of brains that function as representations and affect

behavior, it appears to improve upon computationally-based conceptions in several important respects, including perceptual completions of familiar patterns by filling in missing portions, the recognition of novel patterns even in relation to previously unfamiliar instances, the phenomenon known as "graceful degradation", and related manifestations of mentality (Rumelhart et al. 1986, pp. 18-31). Among the most important differences is that connectionist "brains" are capable of what is known as *parallel processing*, which means that, unlike (sequential) Turing machines, they are capable of (concurrently) processing more than one stream of data at the same time.

This difference, of course, extends to physical computers, which can be arranged to process data simultaneous, but each of them itself remains a sequential processor. The advantages of parallel processing are considerable, especially from the point of view of evolution, where detecting the smells and the sounds of predators before encountering the sight of those predators, for example, would afford adaptive advantages. Moreover, learning generally can be understood as a process of increasing or decreasing activation thresholds for specific patterns of nodes, where classical and operant conditioning may be accommodated as processes that establish association between patterns of activation and make their occurrence, under similar stimulus conditions, more (or less) probable, where the activation of some patterns tends to bring about speech and other behavior.

Those who still want to defend computational conceptions might hold that, even if their internal representations are distributed, human beings are semantic engines (h9):

(h9) Human beings are semantic engines with distributed representations;

but the rationale for doing so becomes less and less plausible and the mechanism–more and more "independent but coordinated" serial processors, for example–appears more and more ad hoc. For reasons that arose in relation to eliminative materialism, however, no matter how successful connectionism as a theory of the brain, it cannot account for the relationship between bodies and minds without a defensible conception of the mind that should explain why symbol systems and semantic engines are not thinking things.

The Total Turing Test

The difficulties posed in relating behavioral evidence to mentalistic hypotheses raise problems suggesting the possibility that the Turing test simply does not go far enough. Stevan Harnad, for example, has explored its character and ramifications in a series of articles in which he emphasizes the importance of infusing otherwise purely syntactical strings with semantic content, if they are to be meaningful symbols rather than merely meaningless marks, as a suitable theory of the mind requires. The necessity to locate a

suitable mechanism for imparting meaning to symbols he calls "the symbol grounding problem" (Harnad, 1990). Harnad therefore offers a new and improved version of the Turing test (TT) in the form of the total Turing test (TTT), encompassing non-verbal as well as verbal behavior, where symbols are "grounded" by the behavior they display.

Harnad's approach blends Cartesian dualism with Turing behaviorism. Because a robot might display verbal and non-verbal behavior arbitrarily similar to that of a human being–a real "thinking thing"–and still not possess mentality, however, there is a permanently unbridgeable gulf between our public behavior and our private minds:

> Just as immunity to Searle's [Chinese Room] argument cannot guarantee mentality, so groundedness cannot do so either. It only immunizes against the objection that the connection between the symbol and what the symbol is about is only in the mind of the [external] observer. A TTT-indistinguishable system could still fail to have a mind; there may still be no meaning in there. Unfortunately, that is an ontic state of affairs that is forever epistemically inaccessible to us: We cannot be any the wiser. (Harnad, 1993, p. 30)

Harnad thus not only rejects the Turing test as a sufficient condition for mentality (h10):

(h10) Systems that pass the Turing test (TT) possess mentality (Turing);

but also rejects his own total Turing test as a sufficient condition for mentality (h11):

(h11) Systems that pass the total Turing test (TTT) possess mentality;

because of which he concludes that the problem of other minds–whether anyone else besides ourselves possesses a mind–can never be resolved. We are forever ignorant.

Harnad, however, may underestimate the epistemic resources at our disposal. The pattern of reasoning known as *inference to the best explanation* supplies a methodology that might transcend his conceptions. This approach involves selecting one member of the set of available alternatives as the hypothesis that provides the best explanation for the available evidence. Any hypotheses that explain more of the available evidence are preferable to those that explain less. Those that are incompatible with the evidence are rejected as false. Those that are preferable when sufficient evidence becomes available are also acceptable. Hypotheses that are acceptable may be false, which makes inference of this kind fallible, but they remain the most rational among the available alternatives. If the hypothesis of the existence of mentality turns out to provide a better explanation for the available evidence than its alternatives, its acceptance could still be warranted.

Semiotic Systems

The conception of minds as *semiotic* (or as sign-using) systems advances an alternative to computational accounts that appears to fit the connectionist model of the brain like hand in glove. It provides a non-computational framework for investigating the nature of mind, the relation of mind to body, and the existence of other minds. According to this approach, minds are things for which something can stand for something else in some respect or other (Fetzer, 1990; 1996; 2001). The semiotic relation itself, which was elaborated by the American philosopher, Charles S. Peirce, is triadic (or three-placed), insofar as it involves a relation of *causation* between signs and their users, a (crucial) relation of *grounding* between signs and that for which they stand, and an *interpretant* relation between signs, what they stand for, and the users of signs.

There are three branches of the theory of semiotic, which include *syntax* as the study of the relations between signs and how they can be combined to create new signs, *semantics* as the study of the relations between signs and that for which they stand, and *pragmatics* as the study of the relations between signs, what they stand for, and sign users. Different kinds of minds can then be classified on the basis of the kinds of signs they are able to utilize, such as *icons*, which resemble that for which they stand (similar in shapes, sizes, and such); *indices*, which are causes or effects of that for which they stand (ashes, fires, and smoke), and *symbols*, which are merely habitually associated with that for which they stand (words, sentences, and things) as iconic, indexical, and symbolic varieties of mentality, respectively.

Meanings are identified with the totality of possible and actual behavior that a sign user might display in the presence of a sign as a function of context, which is the combination of motives, beliefs, ethics, abilities, and capabilities that sign-users bring to their encounters with signs. And patterns of neural activation can function as internal signs, where (all and only) thinking things are semiotic systems, (h12):

(h12) Thinking things, including human beings, are semiotic systems.

This approach can explain what it is to be conscious relative to a class of signs, where a system is conscious with respect to signs of that kind when it has the ability to utilize signs of that kind and is not inhibited from the exercise of that ability. And it supports the conception of cognition as an effect that is brought about (possibly probabilistically) by interaction between signs and sign-users when they are in suitable causal proximity.

Critical Differences

Among the most important differences between semiotic systems and computational accounts becomes apparent at this point, because the semantic dimension of mentality has been encompassed by the definition of systems of this kind. Observe, for example, the difference between symbol systems and semiotic systems in Figures 1 and 2, where semiotic systems reflect a grounding relationship that symbol systems lack, as follows:

This difference applies even when these systems are processing marks by means of the same procedures. A computer processing a tax return can yield the same outputs from the same inputs, yet they mean nothing to that system as, for example, income, deductions, or taxes due. A distinction must be drawn between marks that are meaningful for use by a system and marks that are meaningful for the users of that system. They can function as signs for their users and not function as signs for those systems.

"Symbols" in this sense of semiotic systems must therefore be clearly distinguished from "symbols" in the sense of symbol systems, which can be meaningless marks, lest one mistake symbol systems in Newell and Simon's sense for (symbol-using) semiotic systems, as has John McCarthy (McCarthy, 1996, Ch. 12). This reflects (what might be called) the *static difference* between computer systems and thinking things. Another is that digital machines are under the control of programs as causal implementations of algorithms, where "algorithms" in turn are effective decision procedures. Effective decision procedures are completely reliable in producing solutions to problems within appropriate classes of cases that are invariably correct and they do in a finite number of steps. If these machines are under the control of algorithms but minds are not, then there is a *dynamic difference* that may be more subtle but is not less important as well.

Indeed, there are many kinds of thinking–from dreams and daydreams to memory and perception as well as ordinary thought–that do not satisfy the constraints imposed by effective decision procedures. They are not reliable problem-solving processes and need not yield definitive solutions to problems in a finite number of steps. The causal links that affect transitions between thoughts appear to be more dependent upon our life histories and associated emotions (our pragmatic contexts) than they do on syntax and semantics per se. Even the same sign, such as a red light at an intersection, can be taken as an icon (because it resembles other red lights), as an index (as a traffic control device that is malfunctioning), or as a symbol (where drivers should apply the breaks and come to a complete halt) as a function of a sign user's context at the time. Anyone

else in the same context, presumably, would have interpreted that sign the same way.

The Hermeneutic Critique

Whether or not the semiotic conception ultimately prevails, current research makes it increasingly apparent that an adequate account of mentality will have to satisfy many of the concerns raised by the hermeneutic critique advanced by Hubert Dreyfus (1979). Dryfus not only objected to the atomicist conception of representation that became the foundation for the compositionality of meaning and recursive generability theses that Fodor and Pylyshyn embraced but also emphasized the importance of the role of bodies as vehicles of meaning, especially though interactions with the world, very much in the spirit of Harnad and even of Peirce, which whom he shares much in common. The very idea of creating artificial thinking things that are not inextricably interwined with their bodies and capable of interacting with the world thus becomes increasingly implausible.

Indeed, it now appears clear that differences between Turing machines, digital computers, and human beings even go beyond those addressed above, where the semiotic conception of consciousness and cognition, for example, offers the capacity to make a mistake as a general criterion of mentality, where making a mistake involves taking something to stand for something else, but doing so wrongly, which is the right result.

From this point of view, there appear to be three most important differences, namely:

	(Abstract) Turing Machines	(Physical) Digital Computers	(Actual) Human Beings
Infinite Capacities:	Yes	No	No
Subject to Malfunction:	No	Yes	Yes
Capable of Mistakes:	No	No	Yes

Figure 3. *Three Distinctly Different Kinds of Things*

Even apart from a specific theory of representation intended to account for the meaning of the marks that machines can manipulate, it appears evident from Figure 1 that these are three distinctly different kinds of things where thinking things are unlike machines.

Ultimately, of course, the adequacy of a theory of mind hinges upon the adequacy of the theory of meaning it provides that relates brains, minds, and behavior. The crucial consideration appears to be that, whether bodies and minds are deterministic, chaotic,

or probabilistic systems, it must provide a completely causal account of how the signs that minds employ make a difference to the behavior of those systems that is sufficient to sustain an inference to the existence of mentality as the best explanation for the data. One way in which that may occur emerges from the different ways in which sensations affect behavior, where the dog barked at the bush when the wind blew, *because* he mistook it for a stranger; where Mary rushed to the door at the sound of the knock, *because* she thought her friend had come; or where Bob slowed down when the light turned red, *because* he knew that he should apply the breaks and bring the car to a complete halt.

Conventions and Communication

Because different uses can use different signs with the same meaning and the same signs with different meaning, it is even possible for a sign user to use signs in ways that, in their totality, are not the same as those of any other user. This implies that social conceptions of language, according to which private languages are impossible, are not well-founded from the perspective of semiotic systems. A person who found himself abandoned on a deserted island, for example, might while away the time by constructing an elaborate system of classification for its flora and fauna. Even though that system of signs might therefore have unique significance for that individual user, that system of signs, presumably, would still be learnable in the sense that there is no reason why it could not be taught to others. It would simply be the case it never had.
In communication situations, whether spoken, written, or otherwise, different sign users tend to succeed when they use signs the same way or to the extent to which they mean the same things by them. The question that arises is whether the same sign s stands for the same thing x for different sign users z1 and z2 under specific conditions:

When z1 and z2 speak different languages, such as English and German, the success of a translation can be difficult to ascertain. But it can also be difficult when very similar sounds are associated with meanings that may not mean the same thing for every user. There are circumstances under which we may prefer for our signs to be confidential.

Turing himself, for example, spent time successfully cracking the Enigma cipher during World War II, enabling the English to understand the German's coded messages. Other circumstances, however, encourage the use of the same signs in the same ways, such as in the case of a community of members with common objectives and goals. Systems of public schools, for example, are commonly financed with the purpose, among others, of

instilling the members of the community with a common understanding of the language they use, which promotes communication and cooperation between them. Great nations such as the United States have benefited immeasurably from their standing as "melting pots" where people from many countries come together and are united by reliance upon English, in an absence of which this country like others could tend toward Balkanization.

Other Minds

The conception of minds as semiotic systems also clarifies and illuminates distinctively mental aspects of various kinds of causal processes. When causal relations occur (when causes such as inputs bring about effects such as outputs) and those inputs and outputs do not serve as signs for a system, they may then be classified as *stimuli*. When effects are brought about by virtue of their grounding (because they stand for those things in those respects) for the systems that use them, they may properly be classified as *signs*. And when semiotic relations occur (when signs being used by one user are interpreted by another) between systems that use them, they may be further classified as *signals*. Sometimes the signals we send are intentional (successful, and so on), sometimes not. Every sign must be a stimulus and every signal must also be a sign, but not vice versa.

Every human being, (other) animal, and inanimate machine capable of using signs thereby qualifies as a thinking thing on the semiotic conception. This realization thus explains why dreams and daydreams, memory and perception, and ordinary thought are mental activities, while tooth decay, prostate cancer, and free fall, by comparison, are not. Whether or not the semiotic conception emerges as the most adequate among the alternative conceptions, it has become apparent that an adequate account ought to be one that is at least very much like it, especially in accommodating crucial differences between Turing machines, digital computers, and human beings. It has become equally apparent, I surmise, that minds are not machines. If thinking were governed by mental algorithms, as such accounts imply, then minds simply follow instructions mechanically, like robots, and have no need for insight, ingenuity, or invention. Perhaps we deny that we are nothing but robots because our mental activities involve so much more. Indeed, some of the most distinctive aspects of thought tend to separate minds from machines.

Simulations are clearly too weak and *emulations*, which yield the same inputs from the same outputs by means of the same processes and are made of the same matter, are clearly too strong. But the shoals are treacherous. David Chalmers, for example, has argued that, for some systems, *simulations are replications*, on the presumption that the

same psychophysical laws will be operative. Thus, if the transition from an initial state S1 at time t1 yields a final state Sn at tn, where the intermediate steps involved in the transition between them, say, S2 at t2 through Sn-1 at tn-1, are the same, then properties that are lawfully related to them, such as consciousness, must come along with them, even when they are made of different stuff (Chalmers, 1996). But that will be true only if the difference in matter does not affect the operation of those laws themselves. In cases where it does, *replications may require emulations*.

Intelligent Machines

If these reflections are well-founded, then a conception of minds along these lines would have the capacity to explain why symbol systems and semantic engines are not thinking things. Either they can account for the form of thoughts but not their content, or they cannot account for the transitions between thoughts themselves. Turing machines, with which we began, are not even physical things and cannot sustain the existence of finite minds that can malfunction and can make mistakes. The connectionist conception of brains as (wet) neural networks supplies a crucial foundation for rethinking the nature of the mind, but requires supplementation by an account of the nature of the mind that is non-computational. The hypothesis of minds as semiotic systems blends with the connectionist conception of the brain to support a wholly causal, non-computational account of consciousness and cognition.

Not the least of the benefits that are thereby derived is an account of mentality that can be reconciled with biology and evolution. Primitive organisms must have had extremely elementary semiotic abilities, such as sensitivity to light by means of single cells with flagella to bring about motion. If moving toward the light promotes survival and reproduction, then that behavior would have adaptive benefits for such simple systems. Under the combined influence of genetic mutation, natural selection, genetic drift, sexual reproduction, sexual selection, group selection, artificial selection and genetic engineering, of course, biological evolution, including of our own species, continues to this day, bringing about more complex forms of semiotic systems with abilities to use more signs of similar kinds and other signs of various different kinds, as a consequence of the sort of probabilistic progress that evolution makes possible.

As man-make connectionist systems of (dry) neural networks are developed, it should not be too surprising if they reach a point where they can be appropriately classified as *artificial thinking things*. Whether that point will ever come depends upon advanced in science and technology over which philosophers have no control. While the conception

of symbol systems and even semantic engines appear to fall short of capturing the character of thinking things, this does not mean that they fail to capture the character of intelligent machines. To the extent to which machines properly qualify as "intelligent" when they are able to process complex tasks in a reliable fashion, it is clear that the advent *intelligent machines* arrived long ago. The conceptual confusion has simply been to confound intelligent machines with thinking things, which the history of this subject has now substantially clarified.

References

Chalmers, D. (1996). *The Conscious Mind*. New York, NY: Oxford University Press.

Charniak, E. & McDermott, D. (1985). *Introduction to Artificial Intelligence*. Reading, MA: Addison-Wesley Publishing Company.

Churchland, P. (1984). *Matter and Consciousness*. Cambridge, MA: The MIT

Colburn, T., et al. (Eds.) (1993). *Program Verification: Fundamental Issues in Computer Science*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Dreyfus, H. (1979). *What Computer's Can't Do: The Limits of Artifical Intelligence* (Rev. ed.). New York: Harper & Row.

Fetzer, J. H. (1988). Program Verification: The Very Idea. *Communications of the ACM 31*, 1048-1063.

Fetzer, J. H. (1990). *Artificial Intelligence: Its Scope and Limits*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Fetzer, J. H. (1992). Connenctionism and Cognition: Why Fodor and Pylyshyn are Wrong. In A. Clark and R. Lutz (Eds.), *Connectionism in Context* (pp. 37- 56). Berlin, Germany: Springer-Verlag.

Fetzer, J. H. (1996). *Philosophy and Cognitive Science* (2nd ed.). St. Paul, MN: Paragon House.

Fetzer, J. H. (2001). *Computers and Cognition: Why Minds are Not Machines*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: The MIT Press.

Fodor, J. (1980). Methodological Solipsism as a Research Strategy in Cognitive Psychology. In J. Haugeland (Ed.), *Mind Design* (pp. 307-338). Cambridge, MA: The MIT Press.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition, 28*, 3-71.

Ford, K. M. & Hayes, P. (Eds.) (1991). Reasoning Agents in a Dynamic World:

Harnad, S. (1990). The symbol grounding problem. *Physica D 42*, 335-346.

Harnad, S. (1993). Gounding symbols in the analog world with neural nets: A Hybrid Model. *THINK, 2*, 12-20.

Haugeland, J. (1981). Semantic Engines: An Introduction to Mind Design. In J. Haugeland (Ed.), *Mind Design* (pp. 1-34). Cambridge, MA: The MIT Press.

Hofstadter, D. (1979). *Godel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

Lucas, J. R. (1961). Minds, Machines, and Godel. *Philosophy, 36*, 112-127.

McCarthy, J. (1996). *Defending AI Research*. Stanford, CA: CSLI Lecture Notes.

Newell, A. & Simon, H. (1976). Computer Science as Empirical Inquiry: Symbols and Search. In J. Haugeland (Ed.), *Mind Design* (pp. 35-66). Cambridge, MA: The MIT Press, 1981.

Penrose, R. (1989). *The Emperor's New Mind*. New York: Oxford University Press.

Rumelhart, D., et al. (1986). *Parallel Distributed Processing* (Vols. 1 and 2). Cambridge, MA: The MIT Press.

Searle, J. (1980). Minds, Brains, and Programs. *Behavior and Brain Sciences, 3*, 417-457.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: The MIT Press.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind LIX*, 433-460.

- This chapter is an expanded and revised version of "The Philosophy of AI and Its Critique", in Luciano Floridi, ed., The Blackwell Guide to the Philosophy of Computing and Information (Malden, MA: Blackwell Publishers, forthcoming). By permission of the author.