Searle's Chinese Box: Debunking the Chinese Room Argument*

LARRY HAUSER
507 N. Francis Av., Lansing, MI 48912, U.S.A. (lshauser@aol.com)

Abstract. John Searle's Chinese room argument is perhaps the most influential and widely cited argument against artificial intelligence (AI). Understood as targeting AI proper – claims that computers *can think* or *do think* – Searle's argument, despite its *rhetorical* flash, is *logically* and *scientifically* a dud. Advertised as effective against AI proper, the argument, in its main outlines, is an *ignoratio elenchi*. It musters persuasive force fallaciously by indirection fostered by equivocal deployment of the phrase "strong AI" and reinforced by equivocation on the phrase "causal powers' (at least) equal to those of brains." On a more carefully crafted understanding – understood just to target *metaphysical identification* of thought with computation ("Functionalism" or "Computationalism") and not AI proper the argument is *still unsound,* though more interestingly so. It's unsound in ways difficult for high church – "someday my prince of an AI program will come" – believers in AI to acknowledge without undermining their *high church* beliefs. The *ad hominem* bite of Searle's argument against the high church persuasions of so many cognitive scientists, I suggest, largely explains the undeserved repute this really quite disreputable argument enjoys among them.

Key words: Artificial intelligence, cognitive science, computation, Functionalism, Searle's Chinese room argument.

> Anyone who wishes to challenge the central theses owes us a precise specification of which 'axioms' and which derivations are being challenged. (Searle 1988, p. 232)

1. The Chinese Room and Vicinity

John Searle's Chinese room argument (Searle 1980a; 1984, pp. 38f; 1988; 1989a; 1990a) is perhaps the most influential and widely cited argument against claims of artificial intelligence (AI). This "infamous Chinese room argument" (Fisher 1988, p. 279) has been described by Stevan Harnad as having "already achieved the status of a minor classic" and as "having shook [*sic*] up the entire AI field" so considerably that "things still have not settled down since" (Harnad 1991, p. 47). Even critics of the argument extol its importance: William Rapaport, for instance, deems it "a rival to the Turing Test as a touchstone of philosophical inquiries into the foundations of AI" (Rapaport 1988, p. 83). On the other hand, Searle's argument has been decried, by Dennett as "sophistry" (Dennett 1980, p. 428), and, by Hofstadter, as a "religious diatribe against AI masquerading as a serious scientific argument" (Hofstadter 1980, p. 433). I'm with the masquerade party. When AI pioneer Patrick J. Hayes says the core of cognitive science could be summed up as "a careful and detailed explanation of what's really silly about Searle's Chinese room argument" (Hayes 1982, p. 2) I agree with Hayes about the *silliness,* especially. No doubt

Minds and Machines **7:** 199–226, 1997.

"the argument raises critical issues about the nature and foundations of AI" (Moor 1988, p.35): it raises *and muddles* them. With Georges Rey, I think "this argument has commanded more respect than it deserves." (Rey 1986, p.169).

Searle invariably styles his argument to be "directed at strong AI" (1980a, p.417), a view he characterizes variously. On Searle's initial characterization,

> according to strong AI, the computer is not merely a tool in the study of the mind; rather the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. (Searle 1980a, p. 417).

Searle promptly adds,

> In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, they themselves are the explanations. (Searle 1980a, p. 417)

And yet a while later, we're told,

> According to Strong AI, instantiating a formal program with the right input and output is a sufficient condition of, indeed is constitutive of, intentionality. As Newell (1979) puts it, the essence of the mental is the operation of a physical symbol system. (Searle 1980a, p. 421)

It is best to distinguish what is here run together: low-level empirical or existence claims about the mental properties and prospects of computers (as in Searle's initial characterization), epistemological claims about the nature of psychological explanation (as in the prompt addendum), metaphysical or high-level theoretic claims about the nature of mind (as in the third characterization). Let us mark these distinctions. Call the claim that computers can or someday will think *Possible AI* (PAI) and the claim that some already do think *Actual AI* (AAI). Since these two claims speak most directly to the root AI question "Can a machine think" (Turing 1950, p. 433), I refer to them, collectively, as *AI proper* (AIP).[1] Over against these, distinguish the doctrine that *identifies* programs with minds. This claim – also known as "Computationalism" or "Turing Machine Functionalism" (Searle 1980c) – is *Essential* AI (EAI). Essential AI holds that anything that (rightly) computes *must* be thinking (since that's what thinking is), and that everything that's thinking *must,* conversely, be (rightly) computing.[2] These first three (in a broad sense) *metaphysical* claims need, further, to be distinguished from *epistemological* or *methodological* claims that "programs. . . explain human cognition" (Searle 1980a, p. 417) or that "AI concepts of some sort must form part of the substantive content of psychological theory" (Boden 1990, p.2). Call *these* claims "cognitivism." The view Searle dubs "weak AI" which, curiously, combines *denial* of AI proper with *allegiance* to cognitivism will concern us below only in its negative *metaphysical* asseveration, not in its positive methodological one. Metaphysically speaking, *weak AI,* in claiming that computers (can) only *simulate* thinking, is *no AI*. It's no AI plus a cognitivistic sop.[3]

Besides these main distinctions and the terminology just introduced to mark them, some minor terminological stipulations to facilitate discussion also need to be noted at the outset. Call the right programs – those identified with or claimed to be sufficient for mentality – "Programs" (with a capital 'P').[4] Additionally – "cognitive" or "intentional" mental attributes being the sorts of mental properties computers generally are thought to have (if they have any) or to have the best prospects for acquiring (if they've any such prospects) – understand the mental states at issue to be these: propositional attitudes and their intentional kith and kin (detecting keypresses, trying to initialize the printer, etc.). Finally, to further facilitate discussion, following Searle, take "think" and "have a mind" to mean "have mental properties": understand "mind," then, to "abbreviate" something like "mental processes" (Searle 1984, p. 39); understand "brains cause minds" to be used "as a slogan" for something like "the brain causes mental states" (Searle *et al.* 1984, p.153); etc.

My first aim is to show that as an argument against AI proper Searle's would-be "refutation of strong AI" is an *ignoratio elenchi* wherein two crucial equivocations – on the expression "strong AI" and on the phrase "causal powers (at least) equivalent to those of brains" – tempt the reader to pass invalidly from the argued for *nonidentity of thought with computation* (from not EAI), to the claim that extant *computers don't think* (to not AAI), and thence, by inductive extrapolation, to the claim than *computers can't think* or *probably never will* (to not PAI).[5] Furthermore, I maintain, the advertisedly "brutally simple" (Searle 1989a, p. 703) part of Searle's *formal* Chinese room argument targeting *Essential* AI, is invalid on its face and unsound on every plausible reconstruction. For all its rhetorical brilliance and polemical force, Searle's "minor classic" is *logically* and *scientifically* a dud. Mainly of historical interest, the Chinese room argument survives as a sociological phenomenon. The penultimate section of this paper, consequently, ponders how something so logically and scientifically negligible as Searle's argument could have "shook up the entire AI field" so considerably. I condude by showing how Searle's Chinese room example, by marshalling "ill gotten gains" (Dennett 1980, p. 429) from "impressive pictures and dim notions capturing old prejudices" (Weiss 1990, p. 180) masks the unsoundness of his "derivation."

2. Experiment and Argument

We need also to distinguish the *gedankenexperiment* in which Searle imagines himself locked in a room, "blindly" hand tracing a natural language understanding program (in the form of written instructions, in English) capable of generating appropriate Chinese replies to Chinese queries – the Chinese room *experiment*, or *example* – from the formal "derivation from axioms" or Chinese room *argument* that Searle elaborates in several later presentations. According to one, fairly recent, detailed exposition of it (Searle 1990a, pp. 26–31), this "derivation" proceeds from the following three "axioms" (p. 27):[6]

(AI) Programs are formal (syntactic).

(A2) Minds have mental contents (semantics).

(A3) Syntax by itself is neither constitutive of nor sufficient for semantics.

to the following, conclusion (p. 27):

(C1); Programs are neither constitutive of nor sufficient for minds.

Searle then adds a fourth axiom (p. 29):

(A4) Brains cause minds.

from which we are supposed to "immediately derive, trivially" the conclusion (p. 29):

(C2) Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains.

Finally, from the preceding, Searle claims to derive two further conclusions (p. 29):

(C3) Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program.

and

(C4) The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program.

For now, I credit the usual understanding of the relation of experiment to argument, according to which "Searle's thought experiment is devoted to shoring up axiom 3 specifically" (Churchland & Churchland 1990, p. 34). On this understanding, the "point of the parable about the Chinese room" is to show that "from syntax alone you can't get the mental [semantic] content" (Searle *et al.*, 1984, p. 147).[7] Thus crediting the usual understanding of the "point of the parable" – since I don't dispute A3 – the *example*, for now, won't concern me. My first focus is the "derivation." On any interesting understanding of C2, Searle's would-be derivation of *it* is *invalid*. Furthermore, his "derivation" of C1 is *unsound*. . . due to invalidity (on one construal) or to the falsity of A1 (on another).[8]

3. The Chinese Room and Artificial Intelligence Proper

Searle, from the first, has advertised his Chinese room argument as effective against claims of AI proper – against claims that computers *do* or at least *can* (or probably someday will) *think*. Thus he touts it as being "directed at" the claim that "the appropriately programmed computer literally has cognitive states" (Searle 1980a, p. 4l7); as an argument to "demonstrate the falsity of" the claim that "appropriately programmed computers. . . literally have thought processes" (Searle *et al.* 1984, p.

146); as evidencing that it's not the case that "a machine [can] have. . . thoughts in exactly the same sense that you and I have" (Searle 1990a p. 26); etc. Though he hedges – explaining, "I have not tried to *prove* that a computer cannot think" (Searle 1990a p. 27: my emphasis), that there is no "*logically* compelling reason why [computers] could not also give off [thought or] consciousness" (Searle 1990a p. 31: my emphasis) – Searle continues to maintain that "scientifically, the idea [of AI] is out of the question" (Searle 1990a p. 31); presumably that the Chinese room at least *evidences* the falsity of AI; and, perhaps, that the Chinese room is what *puts* AI "scientifically. . . out of the question." At any rate, Searle's continuing use of the phase "strong AI" both to designate Computationalism (Essential AI) and in contrast to "weak AI" maintains the widely credited impression that his argument is inimical to claims of AI proper. This is misleading. The only construal that makes C2 an immediate trivial consequence of A4 (either alone or in conjunction with A1–A3) makes it trivial in its own right, an instance of the truism that whatever suffices to cause some effect has the same causal powers as whatever else does *insofar as each has the causal power to bring about this very effect*. Thus understood, C2 is too trite to be of interest and too weak for use in arguing against AI proper – the use, I take it, to which Searle would have it put. On the other hand, taken in such a way as to suit it to figure in the implied argument against AI proper that Searle's texts suggest, C2 is not a consequence (obvious or otherwise) of A4 alone or of A4 "in conjunction with [the] earlier derivation [of C1]" (Searle 1980a, p. 417 abstract; 1990a, p. 29). The "derivation from axioms" of C2 *nontrivially understood* is invalid. Furthermore, claims *that computers can think* and even *already do think* are consistent with the premises of Searle's "derivation" (were C2 *trivially understood*, even with its would-be *conclusions*). What, then, underwrites Searle's continuing advertisement of the argument as targeted against the claim that "appropriately programmed computers. . . literally have thought processes"?

Lawrence Carleton suggests that in styling the Chinese Room argument to be "directed at" AI proper – as in the passages just cited – Searle implicitly relies on argumentation along the following lines:

> "Certain brain process equivalents produce intentionality" and "X [a digital computer] does not have these equivalents," therefore "X does not have intentionality." (Carleton 1984, p. 221)

Perhaps, in styling the Chinese Room argument to be aimed at AI proper, then, Searle can be taken to implicitly appeal to something like the following argument:

(C2) Any other system capable of causing minds would have to have causal powers (at least) equivalent to a brain's.

(S) No (presently existing) computer has causal powers (at least) equivalent to a brain's.

∴ (-AAI) No (presently existing) computers are capable of causing minds.

Whether or not Searle himself has some such argument as this in mind when he proclaims AI proper to be "scientifically. . . out of the question," many who take the Chinese room argument for an argument against claims of Al proper, I surmise, yield (given C2) to the temptation of something like this argument.[9] C2, *in a sense*, is an immediate and trivial consequence of A4. Supplemental thesis S just seems an obvious empirical truth about the powers of existing computers. Not AAI. Q.E.D.

Appearances of soundness here are deceiving. They derive from equivocation on the phrase "causal powers (at least) equivalent to a brain's." In the sense of that phrase in which the claim "Any other system capable of causing mind would have to have causal powers (at least) equivalent to a brain's" follows immediately and trivially from the statement "Brains cause minds," it is not at all obvious that presently existing computers don't have causal powers equivalent to brains. The sense that "equivalent causal powers" must bear for C2 to be a trivial consequence of A4 is just that of equivalence *with respect to causing minds or mental states*, the very thing at issue. It merely begs the question against AAI to assume computers lack *this*. On the other hand, in the sense in which it *is* empirically obvious that presently existing computers fall short of brains in their causal powers, "causal powers (at least) equivalent to those of brains" must mean something like *equivalent in all respects*. C2 being understood in this sense, the *inference* from A4 to C2 is anything but obvious and trivial. You might just as well argue, "Since atomic bomb explosions cause death, to cause death requires causal powers (at least) equal to an A-bomb. Dynamite lacks causal powers equal to an A-bomb. So, dynamite explosions can't cause death." In sum, if "causal powers (at least) equivalent" just means "equally capable of causing the specified effect," C2 follows trivially enough from A4, but the truth of S is no more obvious than the falsity of AAI : premising that computers lack causal powers equivalent to brains in this sense just begs the disputed question. On the other hand, if "causal powers (at least) equivalent" means "equivalent in all respects," then the inference from A4 to C2 seems guilty of denying the antecedent (in effect: as Carleton notes) in its seeming assumption that if *A*'s (e.g., brains) cause *B*'s (e.g., mental states), then only *A*'s can cause *B*'s; that what suffices for one is necessary for all. Thus construed, the argument is little better than "Human brains cause mental states"; "No digital computers are human brains"; therefore, "No digital computers cause mental states": not much of an argument.

It may be thought that the argument requires an understanding of "equal causal powers" stronger than merely "equal with respect to being able to cause the specified effect" (the question begging alternative), yet weaker than "equivalent in all respects" (the invalidating alternative): something like "equivalent in the *causally relevant* respects." Apropos of this, Searle suggests that just as a certain amount of horsepower is required to lift a given weight a given distance, a certain amount of mindpower or brainpower is required to produce specific thought processes. Thus he explains that C2's claim that "Anything else that caused minds would have to have causal powers at least equivalent to those of the brain,"

> is a bit like saying that if my petrol engine drives my car at seventy-five miles an hour, then any diesel engine that was capable of doing that would have to have a power output at least equivalent to that of my petrol engine. (Searle 1984, pp. 40–41)

Just as a 100-horsepower diesel engine and a 100-horsepower electric motor are causally equivalent with respect to power capacities, perhaps chimp brains and porpoise brains (e.g.) are causally equivalent with respect to intellectual capacities and, consequently (other things being equal) with respect to their mental performances or endowments. The idea is that there is a mental force (of *consciousness*, Searle would have it) just as there are physical forces of gravitation and electricity; and some brains (e.g., Aristotle’s and Einstein’s) have, so to speak, megawatt capacities, compared to the microwatt capacities of frogs’ brains. This picture is familiar – someone thought lacking in intelligence is called “a dim bulb” – but how are we to understand it?[10]

One way to read “has brainpower” would be as tantamount to “thinks” in the weak, generic sense: “has (some) mental properties.” This alternative follows the question begging path of “equal causal powers” weakly construed. Whether things have the requisite brainpower to allow them to display some specific mental abilities (e.g., whether robots with computerized visual processing really see) will not be independent of (or answerable antecedently to answering) whether they actually display the ability or whether they merely “behave as if they were but. . . are not in fact” (Searle 1989b, p. 197). A more sophisticated variant of this approach – going Aristotelian – would insist there’s a hierarchy of mental powers such that higher powers (e.g., mathematical calculation) presuppose lower ones (e.g., sense perception); so, nothing can calculate that can’t also perceive. This is the route Searle seems to take when he says,

> From an evolutionary point of view, just as there is an order of priority in the development of other biological processes, so there is an order and priority in the development of Intentional phenomena. In this development language and meaning, at least in the sense in which humans have language and meaning, comes very late. Many species other than humans have sensory perception and intentional action, and several species, certainly the primates, have beliefs, desires and intentions, but very few species, perhaps only humans, have the peculiar but also biologically based form of Intentionality we associate with language and meaning. (Searle 1983, p. 160)

This seems an advance over the simplistic approach which made a system’s actual manifestation of a specific mental property the measure of whether it had sufficient brainpower to have that very property: now we take the system’s failure to manifest “lower” mental faculties as evidence of its lacking brainpower sufficient for it to be truly possessed of “higher” capacities (e.g., for mathematical calculation) it might give the appearance of having. Still, this Aristotelian line, on closer consideration, can be seen to partake of the same circularity as the simplistic view. The circle is just larger. It is at least as plausible to regard pocket calculators’ apparent abilities to calculate as evidencing that calculation doesn’t presuppose such "lower" mental capacities as sense perception as it is to regard calculators' seeming lacks of such "lower" faculties as signaling that (appearances to the contrary) calculators don't really calculate.[11]

Suppose, then, instead of going Aristotelian, we try being thoroughly Cartesian, taking the hypothesized general mental force to be a force *of consciousness*: identify different levels of brainpower with having varying degrees of private, introspectable, conscious experience. Certainly, much of what Searle has said in these connections, from his insistence on "the first person point of view" (Searle 1980c, p. 421) to talk of "ontological subjectivity" (Searle 1989b), suggests he is all too willing to appeal to consciousness in this thoroughly Cartesian manner. . . and all too unwilling to face up to the well-known difficulties with such appeals.[12] He curtly dismisses other minds objections, for instance, as attempts to "feign anesthesia" (Searle 1980a, p. 21).[13] Similarly, he dismisses Yorick Wilks's attempt to remind him of Wittgenstein's discussion "to the effect that understanding and similar states cannot *consist in* a feeling or experience" (Wilks 1982, p. 344) as "just irrelevant to my views" (Searle 1982, p. 348) – without explanation. I thought this *was* his view! How else are we to understand "the mind consists of qualia, so to speak, right down to the ground" (Searle 1992, p. 20)? How else to understand the claim that mental phenomena "are conscious experiences" (Searle 1992, p. 63), each "a concrete conscious event" (Searle 1992, p. 225)? That aside – if this is Searle's view – the preceding point still applies. If we credit Searle's intuition that calculators and their ilk haven't a shred of conscious awareness this will still be equivocal in its implications between "Calculators don't calculate" and "Calculation doesn't require consciousness."

On the other hand. . . if as Searle insists, there's "no problem" on his view "about how I know that other people have cognitive states" (Searle 1980a, p. 421); if we can understand his insistence that "I, of course, do not claim that understanding [*sic*] is the name of a feeling or experience" (Searle 1982, p. 348) to be denying that understanding *is* a feeling or experience that the word "understanding" names; then, perhaps (despite appearances to the contrary just noted), we should credit Searle's explicit disavowal of "any. . . Cartesian paraphernalia" (Searle 1987, p.146) and not take him to be positing degrees of "brainpower" identifiable with experienced levels of consciousness after all. Since it's largely through his disavowal of dualism in favor of a "monist interactionist" (Searle 1980b, p. 454) or "biological naturalist" (1992, p. 1) view that "mental phenomena. . . are both caused by the operations of the brain and realized in the structure of the brain" (Searle 983, p. ix) that Searle seeks to sidestep full-bloodless Cartesianism, let's consider whether we mightn't (along these lines) indirectly measure "brainpower" by appealing to such operations and structures in brains as produce it. Though we may be barred from directly measuring the consciousness levels produced by brains, we might, at least, infer the levels of consciousness brains and other systems can produce from some measurable (or at least publicly detectable) structural properties or causal features of brains; much as we calculate the horsepower capacities of internal combustion engines from their displacements and compression ratios.

One trouble with the foregoing proposal, of course, is that we don’t really know in any detail what the relevant structural properties of brains are. We don’t know what features of brains are crucial to determining the mental capacities of brains as displacement size and compression ratio are crucial to determining the power capacities of internal combustion engines. Besides certain vague suggestions that the crucial thought-producing features of brains are chemical and not computational, that “intentionality is. . . causally dependent on the specific biochemistry of its origins” (Searle 1980a, p. 424), Searle has little to say concerning what the crucial mental power producing features of brains are; though he’s adamant enough about what they aren’t. The really fundamental objection to the proposal being considered, however, is not just that Searle has nothing better to propose than the computational hypothesis he derides; it’s this: even the discovery that the crucial causal features of brains *were* chemical would not disprove, nor go far toward disconfirming, AI proper. Suppose future psychologists establish that the crucial thought producing features of brains *really are* chemical. Suppose they discover the electrical switching capacities of brains are as irrelevant to their mental power outputs as the electromagnetic properties (most) internal combustion engines have (as accidental side effects of being made of steel) are irrelevant to their horsepower outputs. It is entirely consistent with such findings and hardly less plausible, given such (if such were found out), that *computers* produce mindpower by *different* means than brains; much as electric motors, despite not having displacements or compression ratios, produce horsepower. The very electrical switching capacities or computational properties Searle conjectures to be irrelevant to the mental power outputs of brains might *still* be crucial determinants of the mental capacities of computers, just as the very electromagnetic properties that are irrelevant to the horsepower capacities of internal combustion engines are crucial determinants of the horsepower capacities of electric motors. From this it is an easy extrapolation to see that *whatever* the crucial determinants of the mental powers *of brains* turn out to be – contrary to the line of argument we are canvassing – AI cannot be shown to be “scientifically. . . out of the question” by showing that computers lack these determinants. Suppose we understood the physiological basis of brains’ production of mental powers as well as we understand how internal combustion engines produce horsepower. Suppose also that no computer has the features which are the crucial determinants of the mental powers of brains. We still cannot conclude, “Therefore, no computer has mental powers” without arguing fallaciously (in effect, denying the antecedent). We would be arguing, in effect: “Of brains, only those with chemical power X produce any mindpower”; “No computer has chemical power X”; therefore, “No computer produces any mindpower.” Again, this is no more valid than to argue “Of internal combustion engines, only those having displacements of (at least) one cubic millimeter produce any horsepower”; “No electric motor has a displacement of(even) one cubic millimeter”; “So, no electric motor produces any horsepower." It's no more vaid than to argue "Of triangles, only those which are equilateral are equiangular"; "Some rectangles are not equilateral"; therefore, "Some rectangles are not equiangular" (Sharvy 1985, p. 126).

To summarize, allowing Searle the charity of the dubious (though familiar) idea of a unitary mental force (brainpower) comparable to the unitary conception of physical force (horsepower) to explore the possibility of staking out a sense of "equal causal powers" intermediate between "equal with respect to causing the specific effect at issue" (which made the supplemental argument beg the question) and "equal in all respects" (which made it invalid) shows this to be a charity that the proponent of AI proper can well afford. The intermediate sense of "equal causal powers" it allows – the sense of "equal brainpower" (*whatever* this turns out to be) – leads the argument into invalidity as surely as the strong rendering of the phrase "equal causal powers" as "equal in all respects." It seems *anything* stronger than the weak (question begging) interpretation of C2 leads down this fallacious path: the question begging interpretation of C2 as a truism, valid on its own head, is the only interpretation of C2 entailed by A4. Inferring C2 from A4 is not going to be valid if "equal" in C2 is understood to include *any* property besides the property of causing the specific mental effect at issue. The only way of interpreting C2 as a valid consequence of A4 yields no interesting (non-question-begging) argument against Actual AI in conjunction with facts about the comparative physical endowments of humans and computers. The Chinese room argument, then, is very far from having any logical force against Actual AI or much inductive force against Possible AI (if - AAI is supposed to disconfirm PAI) at all.

4. Strong AI and Weak AI

Though far from yielding an argument with any logical force against the claim that computers *do think* or significant evidential weight against the thought that they *can*, Searle's original and some subsequent presentations of the Chinese room strongly suggest a sophism which, more than anything, I think, explains the considerable *rhetorical* effectiveness of the Chinese room argument against AI proper. The sophism involves dubbing Essential AI – what Searle elsewhere *un*ambiguously terms "Turing machine functionalism" (see, Searle 1980c) – "strong AI" while, virtually in the same breath, contrasting "strong AI" to the thesis that computers merely simulate the mental abilities they seem to manifest, a thesis Searle dubs "weak AI".[14] Again the reader is tempted to pass invalidly from what is explicitly argued for (- EAI, Searle's C1) to denials of proper claims of AI by means of an implied argument whose invalidity is masked by equivocation on a crucial phrase; here, via something like the following implied "dilemma":

Strong AI or Weak AI
Not Strong AI (by the Chinese room argument).
∴ Weak AI

If Searle's argument succeeds, as it purports to do, in refuting strong AI, then it would seem Searle's argument, via this "dilemma," proves weak AI; and weak AI (they merely simulate) is directly contrary to claims of Actual AI (they really think); and hence, by inductive extrapolation, disconfirmatory of Possible AI. But note – weak AI is contrary to *Actual* AI, not Essential; but Searle's "derivation", if it disproves anything, disproves *Essential* AI (Searle's C1), not Actual. The true form of the "dilemma" Searle's equivocal deployment of "strong AI" suggests, consequently, turns out to be something like the plainly invalid:

AAI or Weak AI
-EAI
∴ Weak AI

Searle's celebrated "refutation of strong AI" insofar as it purports to target AI *proper*, is silly *at best*.

5. Searle's "Brutally Simple" Refutation of Functionalism

Searle sometimes seems to back off from claims to have "refuted" or "demonstrated the falsity" of AI *proper* by the Chinese room argument – allowing, e.g., "I have not tried to prove a computer cannot think" (Searle 1990a, p.27). On this tack, Searle emphasizes the anti*functionalist* thrust of the following "brutally simple" crux (Searle 1989a, p. 703) of the larger argument:

(A1) Programs are syntactical.
(A2) Minds have semantics.
(A3) Syntax by itself is neither sufficient for nor constitutive of semantics.
∴ (CI) Programs by themselves are not [sufficient for] minds.

I dispute the soundness of this "brutally simple" argument (BSA) for antifunctionalist conclusion C1. Simply construed (as a nonmodal, first-order argument), BSA is simply invalid. Adopting the following dictionary for the predicates of Searle's argument

P	:=	is a Program
F	:=	is formal (syntactical)
S	:=	has semantics
M	:=	is a mind

I propose the following most plausible simple reconstruction of BSA:[15]

(A1) $(x)(Px \rightarrow Fx)$
(A2) $(x)(Mx \rightarrow Sx)$
(A3) $-(x)(Fx \rightarrow Sx)$
∴ (C1) $-(x)(Px \rightarrow Mx)$

This is admirably simple, but fallacious. A3 – being equivalent to $(\exists x)(Fx \,\&\, -Sx)$ – asserts that some formal things lack semantics. This is surely true: uninterpreted calculi presented in logic classes lack semantics. Yet it might still be the case, since there are formalisms which are not Programs (or even programs), that all the formalisms that *are* Programs also have semantics and are minds. The falsity of the conclusion is consistent with the truth of the premises. The argument, thus construed, is invalid. Nor, I submit, is there another plausible nonmodal first-order approximation of Searle's argument. In particular, despite Searlean pronouncements like "we should not attribute intentionality [to something] if we knew it had a formal program" and "as soon as we knew the [intelligent seeming] behavior was the result of a formal program we should abandon the assumption of intentionality" (Searle 1980a, p. 421), it would clearly be infelicitous to strengthen C1 to assert that programming *precludes* semantics $((x)(Px \rightarrow -Sx))$, strengthening A3, likewise, to assert that syntax precludes semantics $((x)(Fx \rightarrow -Sx))$. Though the resultant reconstruction would be valid it has no chance of being sound since $((x)(Fx \rightarrow -Sx))$ is blatantly false (sentences on this page, e.g., have both syntax and semantics). No real advantage in terms of interpretative charity accrues to this envisaged strengthening to offset the very substantial failure of fit between Searle's stated axiom (A3) and conclusion (C1) and these would-be strong reconstructions.

No doubt, something is lost in our straightforward (nonmodal, first-order) translation. Searle offers, by way of clarification of C1, for instance, that by it he intends the modal claim "It is possible that (program and not mind)" or, equivalently, "It is not the case that (necessarily(program implies mind))" (Searle 1989a, pp. 702–3). Additionally, "Syntax *by itself* is neither sufficient for nor constitutive of semantics" and "Programs *by themselves* are not minds" (Searle 1989a, p. 703: my emphases) suggest second-order quantification. Following these suggestions, I propose the following most plausible complication of Searle's argument:

(A1^{C}) $F(P)$

Being a Program is a formal (syntactic) property.

(A2^{C}) $\Box(x)(Mx \rightarrow Sx)$

Minds necessarily have semantics.

(A3^{C}) $(\Phi)(F(\Phi) \rightarrow -\Box(x)(\Phi x \rightarrow Sx))$

Formal properties don't necessarily suffice for semantics.

$\therefore$ (C1^{C}) $-\Box(x)(Px \rightarrow Mx)$

Programs don't necessarily suffice for mind.

Thus complicated, the argument is valid, but unsound.[16] It's unsound due to the falsity of AI^{C}.

To see the falsity of A1^{C}, consider that only *running* Programs or Program *executions* are candidate thinkings. E.g., according to Newell's formulation of the functionalist hypothesis (which Searle himself cites) "the essence of the mental is the *operation* [my emphasis] of a physical symbol system." No one supposes that *inert* (nonexecuting) instantiations of Programs (e.g., on diskettes), "by themselves," think or suffice for thought.[17] The Program instantiations in question in the preceding argument, then, should be understood to be just *dynamic* ones. "Program" in the context of this argument connotes execution: P in the formulation above should be understood to be the property of being a Program *run* or *Process*. Now, though every formal or syntactic difference is physical, not every physical difference is formal or syntactic. For instance, the sentence tokens "Ed ate" and "*Ed ate*", though physically differently shaped, are of the same syntatic form: this is what makes them tokens of the same sentence. Similarly, every instantiation of a given program is syntactically identical with every other: this is what makes the spatial sequence of stored instructions and the temporal sequence of operations tokenings or instantiations of one and the same program. It follows that *the difference between inert instantiation and dynamic instantiation is nonsyntactic: P*, the property of being a *Process* at issue, is not a formal or syntactic property but, necessarily (essentially), includes a nonsyntactic element of dynamism besides, contrary to $A1^C$. Note that given Functionalism's identification of thinking with Program execution and the essential dynamism of execution, Searle's denigration of "the robot reply" as tantamount to surrender of the functionalist position (Searle 1980a, p. 420) is unwarranted. The "robot reply" supplements the physical symbol system hypothesis with a causal account of *symbolism* or *reference*. Dynamism being something causal-temporal, appeal to *time* and *causal context* of operations as determinants of semantic contents is a natural extension of the computational hypothesis. It should be further noted, however, that once time is acknowledged to be of the essence, the following High Church Argument for AI proper goes by the board:

(1) Thinking is a species of computation. (Computationalism)
(2) Universal Turing Machines can compute any computable function. (Turing–Church thesis)
(3) Digital computers are (modulo performance limitations) Universal Turing Machines.
∴ (4) Digital computers can think. (Possible AI)

It goes by the board because Turing's thesis is that "*considerations of speed apart*" Turing machines are "universal instruments" capable of computing any computable function (Turing 1950, p. 411: my emphasis).[18] (I will return to this.)

Perhaps there is some alternative construal of Searle's arguments the preceding arguments overlook. Against this possibility, I issue a counterchallenge to Searle or anyone who would maintain that anything like the central thesis (Cl) follows from some such "axioms" (as A1–A3). Give us a precise specification of this conclusion and true "axioms" (invoking modal, second-order quantificational, or whatever complications) from which anything like antifunctionalist conclusion C1 actually follows. If I am right about the nullity of prospects for such specification (or merely pending such), the Chinese room argument is far from a "touchstone" of philosophical and cognitive scientific inquiry into the foundations of AI – more like the Blarney stone. If I'm right, the question that remains is just the historical and sociological one. How have so many been snowed so much by so little?

6. Sociology of Cognitive Science and Sophism

I begin my sociological and historical remarks with some personal history concerning the contribution of Searle's Chinese room argument to the formation of *my own* beliefs about AI. I think not only that there *can* and *someday will be* artificial intelligence; I think *there is already*. Computers, even lowly pocket calculators, really have mental properties – calculating that 7+5 is 12, detecting keypresses, recognizing commands, trying to initialize their printers – answering to the mental predications their intelligent seeming deeds inspire us to make of them. I think this because the evidence of AI (such doings as inspire such predications) is not bad and the arguments against are atrocious. Exhibit A: Searle's Chinese room argument. I call this view "naive AI": it holds that the seemingly intelligent doings of computers inspiring our predications of "calculation, " "recognition," "detection," etc. of them constitute *prima facie warrant* for attributions of calculation, recognition, etc. to them and hence, given the negligibility of the arguments against, typified by Searle's, *actual warrant*. *Naive AI* bases acceptance of Actual AI on the wealth of empirical evidence for and the dearth of credible theoretical reasons against. Possible AI, of course, follows from Actual AI directly: what is possible.[19] I'll return to this amidst the sociology below: cognitive scientists' rejection of naive AI being crucial to the explanation of cognitive scientific credulity in the face of Searle's argument.

How AI's many detractors have been taken in by Searle's "infamous argument" is perhaps easy to explain: take a certain will to believe, a feeling, perhaps, that "machines thinking would be too dreadful" (Turing 1950, p. 444), together with the proverbial human inclination not to look a gift horse in the mouth, plus a modicum of logical *naîveté*; stir an *ignoratio* together with a false dichotomy and a double measure of equivocation in a separate article; fold the latter mixture into the former; and *voilà!* It's harder to explain is how this argument has impressed even some among the philosophical and cognitive scientific cognoscenti in principle favorably disposed to AI. Both factors invoked to explain credulity above – logical *naîveté* and animus against AI – appear to be lacking in the case of the cognoscenti. What, then, explains *their* credulity? The trouble is that, according to my diagnosis, the Chinese room argument is such *bald* sophistry that it remains a mystery how *anyone* with *any* logical acumen, given *any* sympathy for AI, could be *at all* favorably impressed. Yet many have. How so?

Turing, not so long ago, noted that the feeling that machines thinking would be too dreadful "is likely to be quite strong in intellectual people, since they value the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power" (Turing 1950, p. 444). It's notable, today, that cognitive scientific cognoscenti favorably disposed to AI *in principle* are, typically, so disposed *only* in principle. *In fact*, they're opposed. Disdaining the vulgar empiricism of naive AI, eager to rush in where Turing himself refused to tread, as would be definers of "thought" (or would-be reference-fixing discoverers of the nature of thinking), the artificial intelligentsia Searle's argument shook up so considerably have tended to base their allegiance to the proposition that computers *can* think, but *don't*, on something like the high church argument sketched above. Searle's conclusion C1 being the denial of the high church argument's Computationalist premise, the Chinese room argument poses a definite threat to *high church* belief in AI based on some such argument.

Still the question remains. . . if the "brutally simple" part of Searle's argument is as lame as my purview makes it seem, how could anyone with an ounce of logical acumen feel threatened? Well, for one thing, it's really unclear just what Searle's argument *is*: whether modalities and second-order quantification are supposed to be involved (and which and where and how) is left by Searle as an exercise for the reader. The unsoundness of the argument is less apparent than my purview made it seem because the actual structure of the argument intended is less than apparent. To add to this confusion, Searle's famous example of the man shuffling Chinese symbols (following a Chinese natural language understanding program in the form of instructions written in English) yet purportedly "not understanding a word of Chinese" serves not so well to shore up A3, as (sometimes) advertised, but rather better serves to fill and thus obscure the logical gap that scuttled our simplified "brutally simple" argument. The gap, recall, is that the formalism(s) supposed to lack semantics (by way of satisfying A3) might not be Programs. But the formalism Searle in the room is supposed to instantiate, yet still lack semantics, is already a Program! This fills the logical gap in BSA, not nicely, but too well. Given the "experimental" result (ER) that *Program* doesn't suffice for semantics, Cl follows *validly* given just A2 in addition, via what might be called the Chinese room experimental argument (CREA):

(A2) (x)(Mx Sx)

Minds have semantics.

(ER) -(x)(Px $\rightarrow$ Sx)

Programs don't suffice for semantics.

∴ (C1) - (x)(Px $\rightarrow$ Mx)

Programs don't suffice for minds.

So much for shoring up A3. It's odd, to say the least, to propose a more complicated, seemingly invalid, argument (BSA) – as Searle does – in place of this simple, clearly valid one. On the other hand, when the *gedankenexperiment* is challenged we hear a different tune: we hear the *experiment* "rests on" A3, "the simple logical truth that syntax is not the same as, nor is it by itself sufficient for semantics," not vice versa (Searle 1992, p. 200). Some think Searle mounts a powerful argument against AI, I surmise, because they're unduly impressed by the *example* (a failure of intuition, not logic) and insufficiently impressed by the way the Chinese room argument and experiment, together, give out the illusion that both are gainfully employed. The experiment obscures the logical gap in the derivation, while a premise of the demonstration, A3, masks the tendentiousness of the experiment. Each takes in the other's laundry.

7. Indeterminacy Warmed Over

But shouldn't we give Searle his experimental result? And given this result doesn't Searle's wonted antifunctionalist result follow (via CREA) as just indicated? Even more straightforwardly, skipping the detour through semantics and appealing directly to intuitions about understanding, shouldn't we be impressed by "the following 'bare bones' version of the Chinese Room Argument"?[20]

(1) If instantiating a right program R is a sufficient condition for understanding Chinese, then anything that instantiates R understands Chinese.
(2) While in the Chinese Room [a] Searle himself instantiates R but [b] fails to understand Chinese.
∴ (3) Instantiating a right program R is *not* a sufficient condition for understanding Chinese.

We should not. Both conjuncts of (2) are eminently disputable.

Contrary to [a], to hand trace a program as complex as R probably *inherently* (i.e., necessarily and essentially) exceeds human causal capacities.[21] In the original paper-shuffling version of the scenario the performance will be too slow for the subject to respond passably in real conversational or even real epistolary time as required by the Turing test. If the program is supposed to be committed to memory and the operations performed "in one's head" as in Searle's revised scenario (offered in response to the "Systems reply"), if such a feat of memorization *were* nomologically possible, still, the gain in speed therefrom would be offset by loss of accuracy. Now, the probability that, due to missteps, the program will crash or will output garbage approaches certainty.

Contrary to [b], if a human being *could* step through R fast enough on paper or accurately enough "in their head" to respond passably in Chinese in the manner envisaged, it's unclear (1) what the conscious upshot of such superhuman paper shuffling or memorization would be and (2) why the absence of any consciousness of understanding on the part of the agent (if that were the upshot) would be decisive. Concerning the first point, in the memorization scenario, I suspect, if I could perform such a superhuman feat of memorization and mental computation I *would* become cognizant of the meanings of the symbols. More crucially, apropos the second point, even supposing one could respond passably in Chinese by the envisaged method without coming to have any shred of *consciousness* of the meanings of the Chinese symbols, it still does not follow that one fails, thereby, to understand. Perhaps one understands *unconsciously*. In the usual case, when someone doesn't understand a word of Chinese, this is apparent both from the "first-person point of view" of the agent and the "third-person perspective" of the querents. The envisaged scenario is designedly abnormal in just this regard: third-person and first-person evidence of understanding drastically diverge. To credit one's introspective sense of not understanding in the face of overwhelming evidence to the contrary tenders overriding epistemic privileges to first-person reports. This makes the crucial inference from *seeming to oneself not to understand* to *really not understanding* objectionably theory dependent. Functionalism does not so privilege the first-person. Nor does Behaviorism. The point of experiments is to adjudicate between competing hypotheses. Here the troublesome result for Functionalism and Behaviorism only follows if something like Searle's Cartesian identification of thought with private experiencing (the thesis of "ontological subjectivity" (Searle 1989b, p. 194)) with its epistemic corollary of privileged access is already (question-beggingly) assumed. Conflicting "intuitions" about the Chinese room and like scenarios confirm this. Privileging the first person fatally biases the thought experiment.[22]

The detour through semantics taken by CREA promises a two-fold advantage over the "bare bones" version. First, the bare bones result would pertain just to *understanding*: Searle intends *his* result to generalize to *all intentional mental states*. Since all intentional mental states have semantics, the detour through semantics, if it succeeds, allows us to draw the more general conclusion. Secondly (the point that will concern us), the detour through semantics allows Searle to respond to the accusation that the experiment relies, essentially, on question-begging Cartesian "intuitions" about privileged access by insisting,

> The point of the argument is not that somehow or other we have an 'intuition' that I don't understand Chinese, that I find myself *inclined to say* that I don't understand Chinese but, who knows, perhaps I really do. That is not the point. The point of the story is to remind us of a conceptual truth that we knew all along; namely, that there is a distinction between manipulating the syntactical elements of languages and actually understanding the language at the semantic level. What is lost in the AI *simulation of* cognitive behavior is the distinction between syntax and semantics. (Searle 1988, p. 214)

Such insufficiency of *syntax* for semantics has been argued for variously and persuasively; most famously by Quine, Putnam, and Wittgenstein (seconded by Kripke).[23] Since *Processes* are *not purely* syntactic, however, the "conceptual truth" Searle invokes is hardly decisive. In *practice,* there is no *more* doubt about the "cherry" and "tree" entries in the cherry farmer's spreadsheet referring to cherries and trees (rather than natural numbers, cats and mats, undetached tree parts or cherry stages, etc.) than there is about "cherry" and "tree" in the farmer's conversation; or, for that matter, the farmer's cogitation.[24] Conversely, in *theory* there is no *less* doubt about the farmer's representations than about the spreadsheet's. Reference, whether computational, conversational, or cogitative, being equally scrutable in practice and vexed in theory, the "conceptual truth" Searle invokes impugns the aboutness of computation no more or less than the aboutness of cogitation and conversation.

If Searle or anyone had an adequate account of what *does* suffice for reference *plus* an argument to show that such further conditions as cogitation or conversation do meet (such as adequately determine *their* reference) are not met or could not be met by computers, this would at least evidence the falsity of AI proper. Of course this would still not *refute* AI proper or *demonstrate the falsity* of it unless these sufficient conditions were also shown to be necessary. Hypotheses such as Searle's, invoking consciousness, in effect, look to *qualia* to be the "metaphysical glue" (Putnam 1983, p.18) that sticks thought to things; looking to qualia to be necessary and sufficient for semantics. This is famously tempting, and a notorious nonstarter. Making the syntactic "marks" *experiences* distinguished by their *quale*tative "shapes", for instance, does nothing to blunt the force of the various indeterminacy arguments: *qualia* (alone, or in conjunction with computation) underdetermine reference also.[25] The several *live* hypotheses currently claiming to delineate sufficient conditions for reference all look to *causal* relations between sign and signified: all, however, propose conditions that *could be met* by computers. Some such proposals put conditions on reference that extant computers, arguably, *do not yet meet*; though no such proposal is at all well established.[26] Finally, if such causal relations as promising available hypotheses take to be implicated in reference are apt to be *socially* as well as perceptually and behaviorally mediated – if (at least some) reference involves a social "division of linguistic labor" as the reflections of Putnam 1975 and Burge 1979, e.g., suggest – then, what has all along made the farmer's *accountant's* calculations be about the debts, assets, income, expenses, etc., of the farm, seems now to make the farmer's *spreadsheet's* calculations be about the debts, assets, income, expenses, etc., of the farm. Program plays much the same role in the linguistic/arithmetic division of labor as accountant.[27] This, I think, is what mainly inspires us to characterize the computer's performance and the accountant's in like intentional terms. Until further theoretical notice, it also *warrants* so speaking, general troubles about referential determinacy and renewed dithyrambic outbursts about consciousness not withstanding.

Troubles about semantic determinacy are ill brought out by the Chinese room example anyhow – being all mixed up, therein, with dubious intuitions about consciousness and emotions about computers. Searle's "religious diatribe against AI" (Hofstadter 1980, p. 433) has indeed "commanded more respect than it deserves" (Rey 1986, p. 169). *Much* more. The derivation is unsound; the *gedankenexperiment* is fatally biased; and it's even a misleading reminder.

8. Notes

*This essay is based on my dissertation (Hauser 1993a). I am deeply indebted to Herbert Hendry, my thesis director, for his comments, guidance, criticism and advice throughout. Thanks are also due to Richard Hall, Joseph Hanna, Mark Risjord, and Winston Wilkinson: the members of my committee. I have also profited from discussions with Barbara Abbott and the comments of Gregory Sheridan and the Philosophy faculty at Western Michigan University where I presented an earlier version of this paper (May 20, 1994). Thanks too to two anonymous reviewers at *Minds and Machines* for many helpful criticisms and suggestions. I welcome your comments. My website: *http://members.aol.com/lshauser.*

[1] The assertion that computers *can* think being the weakest possible *assertion* of artificial intelligence – anything less than this (as with so-called "weak AI") actually being *denial* thereof – it is really inapt, I think, to call this view "strong". For this reason, the work on which this paper is based (Hauser 1993a) designates the claim that computers *do* think, "Strong AI proper." This usage struck one reviewer as odd "since [in] the sense usually given that term. . . 'strong' AI is the claim that suitably programmed computers can (will) think." My present terminology bows in this reviewer's direction. Nevertheless it is part of the burden of this paper to show that *Searle's* use of the expression "strong AI" *equivocates tendentiously* between the claim that computers "can be literally said to *understand* and have other cognitive states" (Searle 1980a, p. 417) and the claim computation is "the essence of the mental"; between AI proper and Computationalism.

There is ample evidence that "strong AI" remains, in its present applications, infected with the ambiguities Searle originally imbued it with. The reviewer's characterization identifies Strong AI with AI proper: AI is possible. Compare: "strong AI [is] roughly the view that cognition is computation" (Moor 1988, p. 39): identifying it with Computationalism. A query made to the Usenet newsgroup "Artificial Intelligence Discussions" (comp.ai) asking "What do people understand by the phrase "strong AI" (Hauser 1996), strikingly, received two replies. Marvin Minsky (1996): "My impression. . . is that Strong AI is when a machine is "really intelligent": identifying "strong AI" with AI proper. Joshua Singer (e-mail reply): "Searle identifies 'strong AI' with Functionalism" (which Singer takes to be its accepted sense): Essential AI.

[2] Call the right sort of computation "Computation" (with a capital 'C'). Technically expressed, the essentialist claim of EAI is that *in all possible worlds* (under any *conceivable* circumstances) all thinking is Computation (the *metaphysical necessity* of computation for thought) or all Computing is thinking (the *metaphysical sufficiency* of Computation for thought) or both (the *metaphysical equivalence* of thought and Computation). Claims of metaphysical necessity (and sufficiency) are very strong; what's metaphysically necessary obtains *whatever* the laws of nature might be. The stronger the *necessity,* of course, the weaker the corresponding type of *possibility*. Mere metaphysical possibility, consequently, is very weak: what's

nomologically impossible (i.e., inconsistent with the actual laws of nature) can yet be metaphysically possible. (This is why it's no objection to Searle's Chinese room experiment understood as a counterexample to the claim that computation *metaphysically suffices* for thought to complain about the nomological impossibility of the scenario.) PAI, to be of interest, needs to assert *at least* the nomological possibility of computers thinking. It is even better construed, as I construe here (in the text), more strongly yet, as an assertion of genuine or *practical possibility*. *Practical* possibilities are answerable to actually obtaining or obtainable *circumstances* as well as actual natural laws. (See note 5, below, for related discussion.)

[3] Regarding "weak AI", Minsky (1996) writes, "My impression. . . is that weak AI is when a machine can do anything except be 'really intelligent'." Minsky also suggests that "AI researchers" disdain this "strong AI," "weak AI" terminology . . . as well they should! In addition to the *bamboozlement* that the expression "strong AI" sows by equivocating between AI proper and Computationalism, this terminology *lowballs* the would-be AI advocate by styling *no* AI "AI" (albeit "weak") and the very *weakest* assertion of AI (the mere possibility thereof) "strong". It's ironic that so many would-be AI advocates (in what Minsky 1996 terms "non-technical discussions") adopt the crucial term of sophistical art on which the main argument of their chief detractor depends to name the view they would defend.

[4]Roughly, Programs are partial Turing test passing programs insofar as Behaviorism is at issue, and they're partial Turing test passing programs that do it the right way, by implementing "the right programs" (whatever they are) insofar as Functionalism is at issue. By "partial Turing tests" I mean more or less isolated behavioral tests of specific mental attributes (e.g., understanding Chinese) or abilities (e.g., to extract square roots) rather than the full "Imitation game" test of passing for human in *every* (conversationally) discernible mental respect proposed by Turing 1950. What motivates my concern with partial Turing tests (and my unconcern – at least for present purposes – with the total "Imitation game" test) is well expressed by Dretske: "We don't, after all, deny someone the capacity to love because they can't do differential calculus. Why deny the computer the ability to solve problems or understand stories because it doesn't feel love, experience nausea, or suffer indigestion?" (Dretske 1985, p. 24). In a related vein, Turing himself notes his own proposal "may perhaps be criticized on the ground that the odds are weighted too heavily against the machine" since, were "the man to try and be the machine he would clearly make a very poor showing" being "given away at once," e.g., "by his slowness and inaccuracy in arithmetic" (Turing 1950, p. 435). Also, we credit animals with mental abilities despite their inabilities to pass anything like a full Turing test.

[5] The *ignoratio elenchi* fallacy "is committed when an argument purporting to establish a particular conclusion is instead directed to proving a different conclusion" (Copi 1986, p.103). It resembles the "bait and switch" sales technique: one thing (the bait) is advertised cheap to bring you into the store where they try to sell you something else more expensive (the switch). In this connection,

one reviewer for *Minds and Machines* complains that I am unfair "to characterize Searle as believing that he is rejecting PAI on the grounds that he is shooting down EAI" since "Searle claims at points that only machines can think and he explicitly allows for PAI in various passages. (It just can't be done with formal programs.)" Since I do "acknowledge this side of Searle," (in sections 3 and 5, below) this reviewer goes on to wonder if my characterization of Searle's argument as an *ignoratio* is not "misleading." It would be, if this were the only side of Searle to be acknowledged. But there *is* another side. "My discussion," Searle begins, "will be directed at. . . specifically the claim that the appropriately programmed computer literally has cognitive states" (1980a, p. 417). He continues, "I will argue that, in the literal sense the programmed computer [running Schank and Abelson's story understanding program SAM] understands. . . exactly nothing" (Searle 1980a, p. 419); says, "the same would apply to. . . any Turing machine simulation of human mental phenomena" (Searle 1980a, p. 417); etc. Absent explicit retraction of such claims – in fact, given his continuing reiteration of them (see section 3, below) – the seeming qualifications to which the reviewer alludes are more obfuscatory than clarificatory. Even apparent acknowledgments of the possibility of AI – as when Searle says, "indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use" (1980a, p. 422) and "for all we know it might be possible to produce a thinking machine out of different materials [than flesh] altogether – say out of silicon chips or vacuum tubes" (1990a, p. 26) are not so concessive as they seem. There is no concession that this might come about by any *foreseeable* means and no allowance here of any *actual* but only of some bare logical or metaphysical possibility that is probably "scientifically. . . out of the question" (Searle 1990a, p. 31).

I speak here of "computers" and even "extant computers" not just of "machines": the Ur AI issue is whether *these* machines (extant computers) and others of *this sort* (their foreseeable descendants) really think, as they give appearances of doing. Since Searle denies that we are machines *of this sort* – except in the Pickwickian sense in which, supposedly, "the wall behind my back is right now implementing the WordStar program because there is some pattern of molecule movements that is isomorphic with the formal structure of WordStar" (Searle 1992, pp. 208–209) – Searle's disarming physicalistic pieties about us being "*machines* of a special biological sort," again, are hardly an endorsement of PAI. Rather they trivialize PAI, the better to dismiss it: the better to make all seem to hinge on "the right question to ask" (1980a, p. 422) – "But could something think, understand, and so on *solely* in virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?" (1980a, p. 422). The better to make all seem to hinge on the question of *Essential* AI.

Occasionally the *ignoratio* is almost in the open. . . as when Searle writes, "When. . . someone feels comfortable with the idea that a computer would suddenly and miraculously have mental states just in virtue of running a certain sort of computer program the underlying assumptions that make this view seem plausible are seldom stated explicitly" (1992, p. 9); which suggests the only thing supporting the idea the computers might (soon if not already) be genuinely possessed of mental states is allegiance to Computationalism. Rather, I submit, it's just the apparent achievements and prospects of computers that make *Computionalism* seem at all plausible. The locution by which the bait and switch characteristic of the *ignoratio* fallacy gets facilitated here (as it is in other contexts facilitated by equivocal deployment of "strong AI") is "just." If you take "You can *X* just by *Y*ing" for the claim that *Y*ing is *logically* or *metaphysically* sufficient for *X*ing then, *of course*, if computation is logically or metaphysically insufficient for thought you can't – in the sense of "just" that connotes logical or metaphysical sufficiency – think *just* by computing. That's the *bait*, not EAI (which is all the Chinese room argument and scenario, if they show anything, might really show). Now the *switch*. A less stringent, more natural sense of "just" – connoting *actual* or *practical* sufficiency – suggests itself almost irresistibly, here. "You can kill someone just by beheading them" would naturally be taken to be *true*; yet beheading does *not* suffice *logically or metaphysically*. Still it *actually suffices* quite amply. So may computation *actually suffice* (even already have sufficed) for thought; though EAI be false.

I don't say *Searle* believes shooting down EAI warrants rejection of PAI; perhaps he doesn't. (Of this I don't presume to judge.) I do say his formulations invite this inference, and *certainly* many of his readers succumb to the temptation of it.

[6] The axioms and conclusions here are exactly as stated by Searle 1990a with one exception. A2, as Searle 1990a states it, reads, "*Human minds have mental contents (semantics)*" (Searle 1990a, p. 27). Since the point at issue is whether computers have minds *at all* the "human" here is a red herring. Granted: computers aren't human; so, their minds (if they have them) aren't *human*. So what? Cats aren't human either; though cats (who want to be fed, see birds, etc.) – being possessed of mental attributes – certainly have minds and are thinking things; *res cogitans*; what's at issue. Also note, there is no "human" in Searle's conclusions – just "minds". In several other presentations (Searle 1984, p. 39; 1988, p. 232; 1989a, p. 703) Searle himself speaks simply of "minds", stating the axiom much as I do here.

[7] I myself doubt this usual understanding. According to my own understanding of the point of the parable, the thought experiment rests on A3 not *vice versa*; or rather it's sometimes one, sometimes the other. (See, e.g., Searle 1992, p. 200; 1988, p. 214. This is discussed below.) Since this does not affect the *validity* of the derivation at all and can only bear negatively on the warrantness of the premises, insofar as my focus here is on the *argument* rather than the *experiment*, this complication can be ignored for now. It will concern us when we consider the experiment on its own terms, in the concluding section.

[8] Since the derivations of C3 and C4 depend on C1 and C2 as intermediate results

my explicit criticism of the soundness of Searle's cases for C1 and C2 will be implied criticism of his cases for C3 and C4.

[9] When Kenneth MacQueen (1990, pp.193–194) accuses Searle of this same "logical error" as Carleton's analysis and mine here suggest, Searle insists he intends no inference "from the fact that certain conditions are sufficient to produce a phenomenon" to the fact "that those conditions are necessary" (Searle 1990b, p. 164) and insists be means "causal powers (at least) equivalent to those of brains" in what amounts to the weak question-begging sense. His failure to acknowledge that it *is* question-begging and seeming still to suggest that even thus weakly understood C2 argues somehow against AI proper, however, clouds this "clarification". I also note, in this connection, that being sufficient to produce a phenomenon is not a matter of degree. Something's sufficient (hence equivalent on this understanding of "equivalent") or it's not. Why say "(at least)" here, then? Why except to lead us into the temptation of this "unintended" inference! Note, also, how readily the conclusion, here, lends itself to a reading contrary to the *possibility* of or *prospects* for computers causing minds; how precariously the inhospitable induction from nonactuality to impossibility impends.

[10] The underlying thought is similar to Descartes' portrayal of thought (like Newtonian force) as "a universal instrument" and even suggestive of Descartes' allusions to this mental force as the "light of nature." "Consciousness," Searle writes,"is an on/off switch: You are either conscious or not. Though once conscious, the system functions like a rheostat, and there can be an indefinite range of different degrees of consciousness, ranging from the drowsiness just before one falls asleep to the full blown complete alertness of the obsessive" (Searle, 1990c, p. 635).

[11] Hauser 1993b argues it's *more* plausible.

[12] Familiar philosophical difficulties attending views of this sort include other minds problems about how I know whether anything else is conscious, or (if they are) what *their* private conscious experiences are like; doubts (of Watsonian vintage) about the public scientific utility of private introspected "facts"; troubles (of a Wittgenstinean stripe) about the explanatory relevance of conscious experience even where it might seem germane in one's own case, e.g., for explaining what it is to know what the word "red" or the word "pain" means; problems about the possible existence of unconscious mental states such as most research programs in modern psychology (from Freud to Marr and Chomsky) have posited; and notorious longstanding troubles about mind-body interaction. Hauser 1993a (Chapt.6) shows how Searle's "as if dualism" (see also Hauser 1993d) inherits all these difficulties, often in spades.

[13] Searle's more recent (1992, pp.71–78) attempt finally to address other minds problems inherent in such consciousness-centered views as he espouses merely reiterates (and in downplaying the evidential import of behavior, even weakens) the traditional analogical argument.

[14] See the opening salvos of both Searle 1980a and Searle 1990a. The cognitivistic consolation that is supposed to sweeten the bitter metaphysical pill of "weak AI"

is, of course, irrelevant to the argument here

[15] It is essential, here, to understand “P” to abbreviate “Program” (with a capital P) – referring to just the intelligent acting or Turing test passing programs (insofar as Behaviorism is at issue) or just those whose so acting is accomplished in “the right way” (insofar as Functionalism is at issue). Functionalism does not generally hold that any old program at all suffices. Also recall that “minds,” here, abbreviates “intentional mental states.”

[16] To see the validity. . . instantiate Φ to P in $A3^{C}$, yielding $F(P) \rightarrow -\Box(x)(Px \rightarrow Sx)$: this, in conjunction with $A1^{C}$, gives us $-\Box(x)(Px \rightarrow Sx)$ by *modus ponens*. $-\Box(x)(Px \rightarrow Sx)$ means $-(x)(Px \rightarrow Sx)$ is true at some possible world *w*; meaning, in turn, that $-(Pa \rightarrow Sa)$ is true for some individual *a* at *w*. Now suppose, *per impossible*, $\Box(x)(Px \rightarrow Mx)$; meaning $(x)(Px \rightarrow Mx)$ is true at all worlds, hence at *w*. Instantiating $(x)(Px \rightarrow Mx)$ to *a*, then, gives us $(Pa \rightarrow Ma)$ at *w*. Similarly for $A2^{C}$: $\Box(x)(Mx \rightarrow Sx)$ means $(x)(Mx \rightarrow Sx)$ is true at all worlds, including *w*; then instantiating $(x)(Mx \rightarrow Sx)$ to *a* gives us $(Ma \rightarrow Sa)$ at *w*. Finally, by *hypothetical syllogism*, the two preceding results – $(Pa \rightarrow Ma)$ and $(Ma \rightarrow Sa)$ – entail $(Pa \rightarrow Sa)$ at *w*; contradicting $-(Pa \rightarrow Sa)$, which was our earlier result. Hence, by *reductio ad absurdum*, $-\Box(x)(Px \rightarrow Mx)$.

[17] As Moor puts it: ‘Even if the formal structure is instantiated, e.g. as a stack of punch cards on a filing cabinet, it is not a mind or by itself sufficient for having a mind” (Moor 1988, p. 42).

[18] This point “that only running programs . . . what are technically known as ‘processes”’ are candidate thinkings has previously been stressed by Rapaport (1988). Rapaport presses the point in arguing against A3, not against A1 – or, more precisely, $A1^{C}$ as I do here. (Thanks to an anonymous reviewer for reminding me of this.) In fact, the importance of this point was probably first impressed upon me by Rapaport’s speculation “that when we move from *dynamic* systems – physical devices executing programs – to *static* ones – *texts* of programs, reference books, etc. – we’ve crossed a boundary” (Rapaport 1993, p.19).

[19] Chapter 4 of Hauser 1993a elaborates, defends, and names this position “naive AI.” See also Hauser 1993b; 1993c; 1993d; 1994a; 1994b.

[20] Gregory Sheridan (personal communication).

[21] Moor 1988 argues similarly. Moor remarks, “[O]nce the thought experiment is made coherent we realize that we could not be the person in the room even in principle. Searle-in-the-room couldn’t be Searle or any other human being” (Moor 1988, p. 40). (Thanks to an anonymous referee for alerting me to this work of Moor.)

[22] I note, here, that other researchers report results contrary to Searle in connection with similar “blind trace” experiments. During the Second World War, Wrens (Women Royal Engineers) “blindly” deciphered German naval communications following programs of Turing’s devising until machines (called “bombes”) replaced the Wrens. Like Searle in the room “the Wrens did their appointed tasks without knowing what any of it was for” but rather than conclude (with Searle) that

neither Wrens nor bombes were really deciphering, Turing conjectured both were doing so and, in so doing, doing something intellectual unawares (Hodges 1983, p. 211). Similarly, Newell and Simon, hoping to embed human deductive strategies in their General Problem Solver program, collected running commentaries from human subjects doing deductions in the blind (Newell & Simon 1963). This procedure, like Turing's conjecture (again, contrary to the "intuitions" commended by Searle), takes "blind deducing" to be real deducing.

[23]For the Wittgenstein-Kripke argument see Kripke 1982 and Wittgenstein 1958 (sections201ff), especially. Putnam's "model-theoretic argument" is set forth in Putnam 1981. Putnam's (1975) "twin earth" argument, is also germane. Quine 1960, chapt. 2 (and elsewhere) argues for the indeterminacy of sense and inscrutability of reference. The granddaddy of all these indeterminacy arguments is the "paradoxical" Löwenheim–Skolem theorem that "every formal system expressed in the first order functional calculus has a denumerable model"; so every such formalism, regardless of what it is *intuitively* about, also "possesses a model on the integers." "The 'paradox' thus concerns the inadequacy of formalism to its supposed informally conceived object" (Myhill 1951, pp. 43-44). Such radical indeterminacy of translation and inscrutability of reference as alleged by Quine and Putnam amount, in Putnam's words, to "the Skolemization of absolutely everything" (Putnam 1983, p. 15).

[24] I take it nothing turns here on the whether there is a principled distinction between the so-called "derived intentionality" of "conventional signs" (e.g., conversation) and the "intrinsic intentionality" of "organic signs" (i.e., cogitations): the standard Gricean or Neo-Lockean hope is to provide an account of the semantic determinacy of "primordial signs" of "mentalese"' that explains the semantic determinacy of "conventional signs" of natural languages given (in addition) an appropriate account of the conventional derivation of the latter from the former. Though this is not a hope I share, I note here that even were such a two stage account forthcoming, the *derived* nature of "conventional representation" does not automatically give such representation a different *ontological* status : it would not follow from the meaning of computations being derived from the *intrinsic* meanings of cogitations (if the Gricean program pans out) that the meanings of computations are *inferior to* or *differ in nature from* the meanings of cogitations. In human history, "primordial fires" (due to lightning strikes, spontaneous combustion, etc.) undoubtedly are *etiologically* prior to "derived fires" (lit from already burning ones): still, "primordial" and "derived" fires are of the same *nature*.

[25] Searle admits, the "gap in my account is. . . that I do not explain the details of the relation between intentionality and consciousness" (Searle 1991, p. 181). This gap scuttles his portrayal of Quinean indeterminacy of translation and inscrutability of reference as a "reductio ad absurdum" *specifically* of "extreme linguistic behaviorism" (Searle 1987, p. 126) and Computationalism. The problem of intentionality is to specify nonsemantic determinants thereof, and the trouble about indeterminacy is not just "that alternative reference schemes are consistent with all the *public*

empirical data" (Searle 1987, p. 139) but that alternative reference schemes seem consistent with all conceivable data, it seems, *except* such "data" as, e.g., that 'rabbit' means rabbit. But, besides being a cheat (*assuming* reference under the pretext of *explaining* it), there is nothing inherently *private* or consciousness-involving about this datum: the resolution of indeterminacy here is due to bringing *intentionality* directly to bear (assuming it as a primitive) and has nothing to do with bringing *qualia* to bear. The Kripke–Wittgenstein arguments Searle evades (see his 1982 "reply" to Wilks's attempt to remind him of Wittgenstein, cited above) transpose nicely into the Quinean key: suppose the radical translator has direct telepathic access to all the nonintentional contents of the native speaker's *mind* – the very *qualia* produced by the native speaker's rabbit encounters. Note that rabbit encounters are *qualetatively* identical with rabbit-stage encounters and undetached-rabbit-part encounters: alternative reference schemes remain consistent with all *this* data also. 'Gavagai' might *still* mean undetached-rabbit-parts or rabbit-stage for all we know. Likewise, in Putnam's famous "Twin Earth" experiment, Oscar and his twin are supposed to be *qualetatively* as well as physically identical: nevertheless "water" means different things to Oscar's "twin" and Oscar.

[26] Here, I have in mind proposals of Jerry Fodor (see, e.g., Fodor 1990, chapter 4), Ruth Millikan (Millikan 1984), and Fred Dretske (Dretske 1988). Dretske, in particular, explicitly contends his own proposals deny extant computers – though they might allow some future learning machines – the status of thinking things.

[27] As displaced accountants can testify!

References

Boden, M. A. (1990), 'Escaping from the Chinese Room', in Margaret Boden, ed., *The Philosophy of Artificial Intelligence*, New York: Oxford University Press, pp. 89–104. Originally appearad as Chapter 8 of Boden, *Computer Models of the Mind*, Cambridge University Press: Cambridge (1988).

Burge, T. (1979), 'Individualism and the mental', in P. French, T. Uehling, and H. Wettstein, eds., *Studies in Metaphysics: Midwest Studies in Philosophy, vol. 4*, Minneapolis: University of Minnesota Press, pp. 73–121.

Carleton, L. (1984), 'Programs, Language Understanding, and Searle', *Synthese* 59, pp. 219–233.

Copi, I. (1986), *Introduction to Logic* (7th edition), New York: Macmillan Publishing Company.

Churchland, P. and Smith Churchland, P. (1990), 'Could a Machine Think?', *Scientific American* 262, pp. 32–39.

Dennett, D. (1980), 'The Milk of Human Intentionality', *Behavioral and Brain Sciences* 3, pp. 425–430.

Dretske, F. (1985), 'Machines and the Mental', *Proceedings and Addresses of the American Philosophical Association* 59, pp. 23–33.

Dretske, F. (1988), *Explaining Behavior: Reasons in a World of Causes*, Cambridge, MA: MIT Press.

Fisher, J. A. (1988), 'The Wrong Stuff: Chinese Rooms and the Nature of Undertanding', *Philosophical Investigations* 11, pp. 279–299.

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

Harnad, S. (199l), 'Other Bodies, Other Minds: a Machine Incarnation of an Old Philosophical Problem', *Minds and Machines* 1, pp. 5–25.

Hauser, L. (1993a), *Searle's Chinese Box: The Chinese Room Argument and Artificial Intelligence*, East Lansing, Michigan: Michigan State University (Doctoral Dissertation).

Hauser, L. (1993b), 'Why Isn't my Pocket Calculator a Thinking Thing?', *Minds and Machines* 3, pp. 3–10.

Hauser, L. (1993c), 'The Sense of "Thinking"', *Minds and Machines* 3, pp. 21–29.

Hauser, L. (1993d), 'Reaping the Whirlwind: Reply to Harnad's "Other Bodies, Other Minds"', *Minds and Machines* 3, pp. 219–238.

Hauser, L. (1994a), 'Acting, Intending, and Artificial Intelligence', *Behavior and Philosophy* 22, pp. 22–28.

Hauser, L. (1994b.) 'Propositional *Actitudes*: Reply to Gunderson', *Behavior and Philosophy* 22, pp. 35–40.

Hauser, L. (1996), "Strong AI" "Weak AI"', posting to the Usenet Newsgroup *Artificial Intelligence Discussions* (comp.ai), 10 Feb.1996.

Hayes, P. J. (1982), 'Introduction', in P. J. Hayes and M. M. Lucas, eds., *Proceedings of the Cognitive Curricula Conference, vol. 2*, Rochester, NY : University of Rochester.

Hodges A., (1983), *Alan Turing: the Enigma*, New York: Simon and Schuster.

Hofstadter, D. (1980), 'Reductionism and Religion', *Behavioral and Brain Sciences* 3, pp. 433–434.

Kripe, S. (1982), *Wittgenstein, Rules, and Private Language*, Cambridge, MA: Harvard University Press.

Kripke (1982), Cambridge, MA: Harvard University Press. *Wittgenstein, Rules, and Private Languages*.

MacQueen, K. G. (1990), 'Not a Trivial Consequence' *Behavioral and Brain Sciences* 13, pp. 193–194.

Millikan, R. (1984), *Language, Thought and Other Biological Categories*, Cambridge, MA: MIT Press.

Minsky, M. (1996), 'RE: "Strong AI" "Weak AI"', posting to the Usenet Newsgroup *Artificial Intelligence Discussions* (comp.ai), 11 Feb. 1996.

Moor, J. H. (1988), 'The Pseudorealization Fallacy and the Chinese Room Argument', in J. H. Fetzer, ed., *Aspects of Artificial Intelligence*, Kluwer Academic Publishers, pp. 35–53.

Myhill, J. (1951), 'On the Ontological Significance of the Löwenheim–Skolem Theorem', in M. White, ed., *Academic Freedom, Logic, and Religion*, Philadelphia: University of Pennsylvania Press.

Newell, A. (1979), 'Physical Symbol Systems', *Lecture at the La Jolla Conference on Cognitive Science*.

Newell, A. and Simon, H. A. (1963), 'GPS, a Program That Simulates Human Thought', in E. Feigenbaum and J. Feldman, eds., *Computers and Thought*, New York: McGraw-Hill, pp. 279–293.

Putnam, H. (1975), 'The Meaning of "Meaning"', in K. Gunderson, ed., *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science, vol. 7*, Minneapolis: University of Minnesota Press.

Putnam, H. (1981), *Reason, Truth and History*, Cambridge: Cambridge University Press.

Putnam, H. (1983), *Philosophical Papers vol. 3: Realism and Reason*, Cambridge: Cambridge University Press.

Quine, W. O. (1960), *Word and Object*, Cambridge, MA: MIT Press.

Rapaport, W. J. (1988), 'Syntactic Semantics: Foundations of Computational Natural Language Understanding', in J. Fetzer, ed., *Aspects of Artificial Intelligence*, Dordrecht, Netherlands: Kluwer, pp. 81–131.

Rapaport, W. J. (1993), 'Because Mere Calculating isn't Thinking?', *Minds and Machines* 3, pp. 11–20.

Rey, G. (1986), 'Searle's "Chinese Room"', *Philosophical Studies* 50, pp. 169–185.

Searle, J. R. (1980a), 'Minds, Brains, and Programs', *Behavioral ond Brain Sciences* 3, pp. 417–424.

Searle, J. R. (1980b), 'Intrinsic Intentionality', *Behavioral and Brain Sciences* 3, 450–456.

Searle, J. R. (1980c), 'Analytic Philosophy and Mental Phenomena', in *Midwest Studies in Philosophy, vol. 5*, Minneapolis: University of Minnesota Press, pp. 405–423.

Searle, J. R. (1982), 'The Chinese Room Revisited', *Behavioral and Brain Sciences* 5, pp. 345–348.

Searle, J. R. (1983), *Intentionality: an Essay in the Philosophy of Mind*, New York: Cambridge University Press.

Searle, J. R. (1984), *Minds, Brains, and Science*, Cambridge: Harvard University Press.

Searle, J. R. (1987), 'Indeterminacy, Empiricism, and the First Person', *Journal of Philosophy* LXXXIV, pp. 123–146.

Searle, J. R. (1988), 'Minds and Brains Without Programs', in C. Blakemore and S. Greenfield, eds., *Mindwaves*, Oxford: Basil Blackwell, pp. 209–233.

Searle, J. R. (1989a), 'Reply to Jacquette', *Philosophy and Phenomenological Research* XLIX, pp. 701–708.

Searle, J. R. (1989b), 'Consciousness, Unconsciousness, and Intentionality', *Philosophical Topics* XVII, pp. 193–209.

Searle, J. R. (1990a) 'Is the Brain's Mind a Computer Program?', *Scientific American* 262, pp. 26–31.

Searle J. R. (1990b), 'The Causal Powers of the Brain', *Behavioral and Brain Sciences* 13, p. 164.

Searle, J. R. (1990c), 'Who is Computing with the Brain?', *Behavioral and Brain Sciences* 13, pp. 632–640.

Searle, J., R. (1991). 'Perception and the Satisfactions of Intentionality', in E. Lepore and R. Van Gulick, eds., *John Searle and His Critics*, Cambridge, MA: Basil Blackwell, pp. 181–192.

Searle, J. R (1992), *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Searle, J. R., J. McCarthy, H. Dreyfus, M. Minsky, and S. Papert (1984), 'Has Artificial Intelligence Research llluminated Human Thinking?', *Annals of the New York City Academy of Arts and Sciences* 426, pp. 138-160.

Sharvy, R (1985), 'It Ain't the Meat it's the Motion', *Inquiry* 26, pp. 125–134.

Turing, A. (1950), 'Computing Machinery and Intelligence', *Mind* LIX, pp. 433–460.

Weiss, T. (1990), 'Closing the Chinese Room', *Ratio* (New Series) III, pp. 165–181.

Wilks, Y. (1982), 'Searle's Straw Man', *Behavioral and Brain Sciences* 5, pp. 344–345.

Wittgenstein, L. (1958), *Philosophical Investigations*, trans. G. E. M. Anscombe, Oxford: Basil Blackwell.