THE CHINESE ROOM ARGUMENT: DEAD BUT NOT YET BURIED

Robert I. Damper
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK.

Abstract

This article is intended as an accompaniment to Anthony Freeman's review of Preston and Bishop's collection, *Views into the Chinese Room*, reflecting on some pertinent outstanding questions about the Chinese room argument (CRA). Although there is general agreement in the artificial intelligence (AI) community that the CRA is somehow wrong, debate continues on exactly why and how it is wrong. Is there a *killer* counter-argument and, if so, what is it? One remarkable fact is that the CRA is prototypically a thought experiment, yet it has been very little discussed from the perspective of thought experiments in general. Here, I argue that the CRA fails as a thought experiment because it commits the fallacy of undersupposing, i.e., it leaves too many details to be filled in by the audience. Since different commentators will often fill in details differently, leading to different opinions of what constitutes a decisive counter, the result is 21-plus years of inconclusive debate.

There can be few contributions to the literature of artificial intelligence (AI), cognitive science and the philosophy of mind to rival Searle's 1980 paper introducing the Chinese room argument (CRA) for the debate and controversy which it has engendered. It is certainly remarkable that this short paper continues to be argued over after all these years. The present article is intended as an accompaniment to Anthony Freeman's review of *Views into the Chinese Room*, edited by Preston and Bishop (2002) to mark the 21st anniversary of the CRA's first appearance.

The longevity of the debate naturally prompts the question: Is there anything new to say after all this time? Actually, I think there is (hence this article), but let us defer an immediate verdict on this and pose two other questions:

1. *What is the current status of the CRA?* That is, is it decisive against AI or, on the other hand, is it logically flawed or just plain wrong? And if it is decisive, what particular conception of AI does it demolish? Are there conceptions of AI left untouched?

2. *Is the CRA* useful *in contributing to the way that we think about AI, cognitive science, and philosophy of mind?* Many doubt this, e.g., according to **Harnad**[1] (p. 295), Pat Hayes has called it "false—and silly". But the volume of debate it has engendered surely means that it cannot be so easily dismissed: It is incumbent on its detractors to provide a 'killer' argument against it. Does such a killer rebuttal yet exist?

In the remainder of this article, I will attempt (a little discursively) to answer these questions. My purpose will be to show that the CRA is not obviously false and silly, yet it is flawed in a particular way, and hence it is not decisive against its target. Debate surrounding the CRA has been somewhat useful, in forcing its detractors to sharpen up their counter-arguments and to reflect on what concepts like artificial intelligence and machine consciousness really mean and how we might recognise them. But the particular way in which the CRA is flawed also leads to sterile and indeterminate debate. In this sense, it is dead but not yet buried.

There can be few readers of this journal who are unaware of Searle's original (1980) argument and the more or less standard rebuttals of it. Basically, the CRA directly attacks Searle's conception of 'strong AI'—the notion that "the appropriately programmed computer really *is* a mind". He contends that an AI program designed to answer questions posed in Chinese, and capable of passing the Turing test (Turing 1950), could not be said to

[1]Since I will be citing individual chapters in the Preston and Bishop book quite extensively, I will refer to these by giving the chapter author's name in bold font without year of publication, since this is implicitly 2002.

'understand' since Searle—understanding no Chinese—could hand-simulate the execution of the AI program but he would still understand Chinese not at all.

Searle's original formulation of the CRA pictured him ensconced in a sealed room, presumably to emphasise that he knew nothing of the purpose (whose purpose?) of the interaction in which he was unwittingly engaged. From outside the room, he would be passed cards on which were written strange (to him) 'squiggles and squoggles' which were actually questions in Chinese, posed by Chinese-speaking interrogators. Consulting a manual compiled by the AI engineer (i.e., the AI program) telling him explicitly how to process the 'squiggles and squoggles', he would then—quite unknowingly—supply answers to the posed questions in a way sufficient to pass the Turing test, i.e., such that the interrogators could not tell that answers were being provided by Searle hand-simulating an AI program, rather than by another Chinese-speaking person. This formulation with its room and supplied manual naturally invites the popular attack on the CRA known as the "systems reply" (e.g., Wilensky 1980) which, along with its many variants, holds that the 'intelligence' of the Chinese room resides in the complete system of which Searle is only a part, not in Searle alone. His favourite answer to this, picturesquely known the "outdoor version" of the CRA, is that he could simply (!) dispense with the room, any contents such as pencil and paper for performing intermediate calculations, etc., and memorise the instruction manual—so that there would be *nothing but Searle* to the system—yet (he claims) he would still not understand. Further, Searle points out, the systems reply simply assumes—without supporting argument—that intelligence (or understanding) arises at the system level.

In spite of what Searle obviously sees as his demolishing counter-arguments, many (perhaps most) AI people take the systems reply to be decisive against the CRA[2]. Such 'Systematists' mostly agree with Hayes that the CRA is false (and silly) and consider further

[2]In particular, judging from class discussions of the CRA with my students over the years, the idea that such intelligence as exists resides in the manual compiled by the knowledgeable AI engineer is especially persuasive.

debate to be futile. Indeed, so entrenched is the systems reply that many commentators feel compelled to put intellectual distance between it and their own rejoinders to the CRA.

For instance, **Copeland** (pp. 110–111) is at pains to emphasise that his "logical reply" is different from the systems reply as follows. The systems reply begs the question by its assumption (without argument) that the system as a whole must understand Chinese whereas the logical reply is "a point about entailment *[which]* involves no claim about the truth—or falsity—of the statement that the Room can understand Chinese". **Copeland** takes Searle's proposition that the symbol manipulation performed by Clerk (**Copeland**'s name for the human in the Chinese room) does not enable him to understand Chinese, and reasons that this does not entail the quite different proposition that this same symbol manipulation does not enable the room as a whole to understand. According to **Copeland**: "One might as well claim that the statement 'The organization of which Clerk is a part has no taxable assets in Japan' follows logically from the statement 'Clerk has no taxable assets in Japan'." This argument is clearly fallacious; an organisation of the kind which can have taxable assets is created in company law to act as proxy for an individual—a situation quite disanalogous to the Chinese room. But leaving this aside, the distinction that **Copeland** seeks to draw does not seem to get us very far. For if we interpret the systems reply as holding that intelligence *could* reside in the system as a whole, then this subsumes the logical reply. **Copeland**'s wish to make the two different appears to be based on a too-narrow interpretation of the many and various threads that together make up the systems reply.

Harnad is one of the few who sides with Searle on the inadequacy of the systems reply and its variants[3], which he describes as "ad hoc … dogged … would-be rebuttals" (p. 301; see also Harnad 1989). He takes the outdoor version of the CRA to be "decisive" and accuses Systematists of "resorting to the even more ad hoc counter-argument that even inside Searle

[3]Of course, **Harnad** opposes Searle implacably on what he takes to be the core of the CRA.

there would be a system, consisting of a different configuration of parts of Searle and that that system would indeed be understanding" (pp. 301–302). He writes that "ad hoc speculations of this order ... show only that the CRA is not a proof", but I suspect that this is an anti-Searle proposition on which **Harnad** and the Systematists would heartily agree!

In spite of **Harnad**'s reservations, however, the so-called "part-whole fallacy" (**Haugeland**'s term, see p. 380) has been enthusiastically embraced by many as an effective rejoinder to Searle. As we might expect from his saga of Clerk's lack of taxable assets in Japan, **Copeland** has been prominent in championing this line, both in the Preston and Bishop collection (pp. 112–113) and in his earlier work (e.g., Copeland 1993). The "fallacy" is supposed to be this: The (claimed) fact that Searle as a *whole* does not understand Chinese does not entail that there is no *part* of Searle that understands Chinese. In his 1993 book, Copeland develops this counter-argument via a thought experiment in which "fanatical AI researchers" highjack part of their victim's brain (p. 129) which they then use (unbeknown to the poor victim) to compute solutions to tensor equations! Copeland apparently believes that this wildly imaginative and totally implausible scenario counters "Searle's principle that if you can't do X, then no part of you can do X". One can easily endorse **Harnad**'s low opinion of these inventive conjectures, which he dubs "heroics" (p. 302).

Yet the counter does have some virtue, if only to remind us of the complex, non-binary nature of *understanding*. Exactly how does understanding relate to the conscious process of memorisation which underpins the outdoor CRA? This point was taken up by Abelson (1980) in his commentary on the 1980 BBS article, and the next few personal remarks owe their origins partly to him.

When I was young, I learned multiplication tables by rote (small beer compared to the prodigious feats of memorisation required by the outdoor CRA). Now to what extent did I 'understand' multiplication as a school child? As Abelson writes: "At what point does

a person graduate from 'merely' manipulating rules to 'really' understanding?" (p. 424). According to the outdoor CRA, never, since memorisation of the steps required to answer questions in Chinese is deemed insufficient for understanding. But the more I used these memorised facts of multiplication for problem solving, the more I felt that I did understand. Admittedly, there is a difference of kind between the situation of the child at school, in which a teacher encourages understanding by active engagement with the problem at hand, and that of Searle-in-the-room (and, we must surely infer, outdoor-Searle too) who, by the premises of the thought experiment, is isolated from the interaction with the interrogator and knows nothing of its question-answer nature. But is it really so clear and obvious that outdoor-Searle does not understand Chinese?

These considerations serve to emphasise that the sort of "understanding" that might or might not be going on in the Chinese room is not the all-or-nothing phenomenon that Searle repeatedly assumes it to be. As a sophisticated philosopher, Searle obviously knows well that "there are many different degrees of understanding" (Searle 1980, p. 418), yet he still sees fit to assert that there are *clear cut cases* in which "understanding" literally does or does not apply, and the Chinese room is one of these. Well, is it? And if so clear cut, why has controversy raged so over the years?

Thus far, we have explored at some length the systems reply to the CRA and Searle's response to it. We could, of course, continue in the same vein to look at the so-called robot and brain simulator replies, also the subject of Searle's refutation in his original article. Like the systems reply, the robot and brain simulator replies are thoroughly worked over in the literature, as is the newer connectionist response dubbed "the Chinese Gym" (e.g., **Copeland**, p. 116) yet, I contend, the debate has still failed to reach any definitive conclusion in terms of identifying the clinching argument that will settle the issue, once and for all.

There are, however, those who believe that Searle has responded (often at length) to

counter-arguments, like the systems reply, where he feels relatively secure but has ignored those where he is more vulnerable—and just maybe one of these is our elusive *decisive* counter. Back in 1991, Dennett certainly believed so, referring to "the definitive refutation, still never adequately responded to Searle" (footnote 2, p. 436) which he attributed to Hofstadter (1980). Dennett cites *The Mind's I* (Hofstadter and Dennett 1981) as his source of the definitive refutation but this is actually a fleshing-out of their original BBS commentaries. Referring to the outdoor CRA, Dennett (1980) writes that it "... suggests either that there are two people, one of whom understands Chinese, inhabiting one body, or that one English-speaking person has, in effect, been engulfed within another person ... who understands Chinese" (p. 429). Hofstadter (1980) accuses Searle of inviting the reader "to participate in a great fallacy", namely that "he is inviting you to identify with a non-human which he lightly passes off as a human" (p. 434), since to operate as it does, passing the Turing test, it must simulate with unimaginable speed and power, way beyond that attainable by any human. In *The Mind's I*, Dennett conjoins and develops these two arguments (pp. 373–382).

So has Searle really ignored this "definitive refutation"? In *The Mystery of Consciousness* (Searle 1997), there appears a sometimes vitriolic exchange between Dennett and Searle on this point (pp. 115–131). Searle (p. 126) is adamant that Dennett's claim of disregard is false, detailing all the occasions on which he has replied. But to reply is not the same as to address the fundamental criticism. As a neutral, my reading is that Searle has definitely replied, yet he has not obviously addressed the main criticism(s) that Dennett has in mind. I do, however, have sympathy with Searle since it is not outstandingly clear what Dennett believes the killer rebuttal to be. Is it that outdoor-Searle really does understand Chinese? Or is it that one *part* of the schizophrenic character engulfed within another understands Chinese? Or is it (*á la* Hofstadter) that no human could pull off the cognitive feat required of outdoor-Searle? Until Dennett is clearer on this, I do not feel he can yet claim a definitive, clinching refutation.

Subsequent to his original formulation in 1980, Searle has somewhat modified or extended the argument (e.g., 1984, 1990, 1997). Thus, he writes in the new collection:

> "The Chinese Room Argument, in short, rests on two absolutely fundamental logical truths … First, syntax is not semantics … and secondly, simulation is not duplication." (p. 52)

One might reasonably ask if Searle is hereby innocently and usefully elucidating and amplifying the CRA with the purpose of avoiding ambiguity, or is he subtly shifting the argument to make a moving target for its detractors? Undoubtedly, Searle himself believes he is doing the former, yet I am not convinced.

Considering his first point, many commentators have argued strongly that the semantics of computation is not so simply divorced from syntax, as for example in **Haugeland**'s contribution[4]. He criticises Searle for putting up a straw man version of strong AI, in which syntax is not sufficient for semantics (p. 382 et seq.). No serious AI practitioner would agree with this, says **Haugeland**. For a computer program *really* to be a program, it must be concretely implementable and semantically interpretable, as well as being describable syntactically. By "concretely implementable", **Haugeland** means it must have the 'right causal powers', a phrase of Searle's that has been widely criticised as lacking definition in his original article, or subsequently. So: "The only point of contention is what the right causal powers are … serious AI is nothing other than a theoretical proposal as to the genus of the requisite causal powers" (p. 388).

[4]The interested reader should also examine the recent book of Baum (2004) which argues extensively, from the perspective of a fairly proselytising brand of computationalism, for a tight coupling between syntax and semantics (both in computation and in human thought, as these are strongly related for Baum). In human thought, this tight coupling is brought about by the joint action of evolution and Occam's razor, an idea which seems to owe something to Ernst Mach who stressed "the biological necessity of conforming thought to environment" (Sorenson 1992, p. 51).

Similarly, many commentators have taken up the cudgels on the "simulation is not duplication" issue. Probably the best answer is that of Copeland (1993, p. 182), who points out that there are two distinctly different kinds of simulation: those that duplicate the phenomena of interest and those that don't. For the former, simulation can be duplication. Consider, for example, a computerised automobile engine management system that replaces an earlier mechanical system for the same purpose. The new system works by sensing engine conditions, simulating the old mechanical system on its internal computer, and adjusting fuel mixtures, timings etc. accordingly via output effectors.

Now is this simulation not a duplication? Certainly, it has the right causal properties, but only if we include the input sensors and output effectors within the definition of the 'simulation' or 'computation'. As **Haugeland** points out, "… within the narrow system (the computer itself), the necessary causal interactions … have to be mediated by special facilities (called 'transducers')" (p. 388). This, of course, is the very essence of the robot reply. And as with computers so with minds. Just as we have to include the transducers in the computerised system to tie semantics to syntax, so "intelligence requires a body", to quote McFarland and Bösser (1993, p. 271). In this case, Searle's frequent insistencies that computation is observer-relative disappear. We do not need the car driver or the designer of the engine management system to 'interpret' the outputs and impose the desired semantics; the larger system does this for itself. So exactly what is computation and can it validly include transducers (cf. **Haugeland**) or not?

Despite attempts like those of Harnad (1994) and Copeland (1996), I think it is fair to say that computation has never been absolutely and precisely defined in terms on which all could agree[5]. It seems that much of the debate on "semantics and syntax" and "syntax and

[5]For example, Pylyshyn (1984, p. 69) writes: "… despite some 50 years of study (starting with Turing's famous paper on computability), there is still no consensus on just what are the essential elements of computing". See Smith (2002) for a more recent statement of the same view.

physics" over the years boils down to disagreement over what computation is and isn't. And even if we adopt a strict definition that excludes transducers (such as "computation is the operation of a universal Turing machine"), **Harnad** points to an interpretation that preserves a role for computation when he writes: "... the CRA shows that cognition cannot be *all* just computational, it certainly doesn't show that it cannot be computation *at all* ... Searle seems to have drawn stronger conclusions than the CRA warranted" (p. 303).

Returning to the earlier question, are these issues about semantics and simulation-versus-duplication central to the CRA or a diversion from the real argument? In *The Mystery of Consciousness*, Searle writes of the CRA: "This is such a simple and decisive argument, that I am embarrassed to have to repeat it (p. 11)". But perhaps the fact that he needs to repeat it at different times and in different terms is a reflection of an inherent under-specification of the original formulation. The CRA certainly seems deficient in not spelling out more clearly exactly what it would entail to be Searle simulating the Chinese-understanding AI program. But of course, he does not know! The feat is so far beyond human capabilities that no one knows. The question "What is it like to be outdoor-Searle?" is at least as challenging to answer as the famous philosophical question "What is it like to be a bat?" discussed by Nagel (1974) who writes:

> "So if extrapolation from our own case is involved in the idea of what it is like to be a bat, the extrapolation must be incompletable. We cannot form more than a schematic conception of what it is like." (p. 436)

And as with the bat, so it is with outdoor-Searle.

In his introduction to *Views into the Chinese Room*, **Preston** (pp. 24–25) attempts to dismiss certain "misunderstandings ... which should be quashed from the start". One of these is that "Searle's scenario is unrealistic ...", going on to say that it is *in principle* irrelevant that the human simulator would have to work at unimaginable speed or might be

unable to memorise the programs in question[6]. But the point surely is that Searle claims to *know* what it would be like to be the human simulator and, further, claims that all readers know it too, by virtue of being human. Yet he cannot know this. It is too far beyond our experience. **Preston** also considers briefly the fact that the CRA is (rather obviously) a thought experiment. He sees nothing remarkable in this, as a thought experiment merely reflects on what would follow in some counterfactual situation and "in this respect it does not differ from Einstein's request for us to imagine what we would observe if, *per impossible*, we were riding on the front of a beam of light". But are thought experiments in physics, where we have a sound body of extant theory to guide us, the same sort of enterprise as thought experiments in the philosophy of mind and consciousness, where no such theoretical underpinnings exist? I aim to take up this question before concluding.

Considering thought experiments in general, Brown (1991) writes: "… there is very little literature on the subject of thought experiments" (p. x), a situation which has changed but little in recent years (see Sorenson 1992, Bunzl 1996, Arthur 1999 and Gendler 2000 for particular contributions). Brown credits Wilkes (1988) with allowing that thought experiments are useful in physical science but not in the philosophy of mind (pp. 28–31). The problem is that the latter "take us too far from reality". Thought experiments work by evoking intuitions, with which discussants are invited to agree. Although these can be useful in many cases, these intuitions can also be misleading. Wilkes's concern is that imaginary cases (in the domain of personal identity) which are wildly implausible and/or lack sufficient background definition can evoke unreliable if not erroneous intuitions.

Brown argues mildly against this in the case of the CRA, saying that "… there is enough background information to legitimize (in principle) … Searle's Chinese room thought experiment". However, he gives no arguments to support this position. I believe the

[6]I can't see Hofstadter agreeing. See also French (2000) and Brooks (2002) to name just two from a host of respected commentators who have taken this "misunderstanding" seriously.

present article offers plenty of reasons for thinking that Searle signally fails to give enough background. A contrary view to Brown's is that of Cole (1984), who says of the CRA:

1. It is not clear (in spite of Searle's denials) that the human in the Chinese room does not understand.

2. There is an important disanalogy between the machine simulation of human performance and the human simulation of machine performance.

He argues that "a fallacy of composition is at work here" and is likely to occur "whenever one takes the *perspective* of the subsystem or constituent" (p. 432). Cole has subsequently (1991) refined this to his multi-personality reply to the CRA. That is, a mind realised by running a computer program as Searle envisages would be a new entity, logically distinct from the person or computer executing the instructions. This is strongly reminiscent of the arguments of Dennett and Hofstadter referred to above.

Interestingly, Hofstadter and Dennett are quite sanguine about the chances of thought experiments contributing to cognitive science and the philosophy of mind. In *The Mind's I*, they write:

> "These are not directly empirical questions but rather conceptual ones, which we may be able to answer with the help of thought experiments." (p. 8)

I am far less optimistic, preferring to side with Wilkes. My position on this has been shaped by the realisation that even in physics, Einstein's intuitions were sometimes famously wrong[7] while philosophy abounds with wildly implausible scenarios masquerading as serious arguments. Not only is the CRA one of these, but rejoinders like those of **Copeland** mentioned earlier (fanatical AI researchers highjacking innocent victims' brains to perform tensor calculus) are no better.

[7]... as in the celebrated case of the EPR thought experiment (Einstein, Podolsky, and Rosen 1935).

In what is probably the definitive text on the subject, Sorenson (1992) details in Chapter 2 common scepticisms about thought experiments whereas in Chapter 10 he outlines some of the fallacies to which thought experiments are prone, and also attempts to answer the common scepticisms (introducing what he calls 'antifallacies'). In the remainder of this article, I will discuss one antifallacy and one fallacy which seem especially relevant to the CRA.

According to Sorenson (1992), an antifallacy is a good inference rule that looks like a bad one. He specifically deals with objections to a thought experiment on the grounds that it is "too unrealistic" or "too bizarre" under the name the Far Out Antifallacy (pp. 277–284). The "far out" objection has surfaced during this article, and is also at the heart of many AI scientists' rejections of the CRA (e.g., Brooks 2002). So is it really an antifallacy, in and of itself? Sorenson certainly believes so; pointing out the popularity of this objection, he calls it the "master antifallacy ... the rich man's version". He avers that a demonstration that the supposition of a thought experiment suffers from the 'right' kind of impossibility constitutes a legitimate and successful attack, but it is not easy to see precisely what he means by this. He writes: " 'Impossibility' has to be relativized to the proper background constraints" (p. 278) but I confess that is too cryptic for me. He seems to have in mind that there are different kinds of impossibility (logical, physical, practical, ...) and only logical impossibility is the right kind[8]. Well, if so, there are certainly many who believe that Searle describes a logical impossibility. Although this may seem to give Searle an easy out ("There, I *told* you strong AI was impossible!"), it is only Searle's conception of strong AI

[8]As a slight aside, this has always been my objection to the famous (but, I believe, vacuous) Twin Earth thought experiment of Putnam (1975). It is logically impossible for water $= H_2O$ in one earth and water $=$ XYZ in its twin to be different yet indistinguishable. It is also physically impossible, since the chemical composition of a compound dictates its observable physical properties by universal laws that hold everywhere.

that is refuted. And, as we have seen (e.g., **Harnad**, **Haugeland**), many believe this to be a straw man.

In what Sorenson calls the fallacy of *undersupposing* (pp. 258–259), the designer of the thought experiment fails to be specific enough, with the consequence that the audience:

> "… unwittingly read in extraneous details. If their creativity leads them to supply diverging details, they become embroiled in a dispute or seduced into a consensus that is merely verbal."

This seems to me to be a very fair characterisation of the CRA debate over the last 21-plus years! So what specific details has Searle left out? A reasonable answer to this question is *everything*! Searle prides himself on what he sees as the conciseness of his CRA, yet it is concise just because it remains silent on the internal workings of the AI program, its underlying assumptions, how it handles world knowledge in such a way as to cope with the frame problem, how it is able to answer context-dependent questions (like "what was the question that I asked just before the last one?"), and so on. And Searle cannot supply these details because he has simply no idea how to construct an AI program capable of passing the Turing test; no one does.

Of course, Searle could well say that it is not his job to supply such details but the AI community's, since they are the ones making the claims about machine understanding that he wishes to refute. The original CRA took Schank and Abelson (1977) as an especially clear example of the best contemporary work in AI. But it is abundantly clear that no Schank-Abelson-style program based on scripts stood a hope of passing the Turing test, then or in the future. So it would be disingenuous to equate a script-based program with the sort of thing that Searle has in mind for the Chinese room. The necessity stands for Searle to provide a better specification, or admit that he cannot. Otherwise, the debate will continue interminably; and we will be forever discussing beliefs about what it means to be Searle the

Simulator, rather than facts about minds and machines.

References

Abelson, R. P. (1980). Searle's argument is just a set of Chinese symbols. *Behavioral and Brain Sciences 3*(3), 424–425. (Peer commentary on Searle, 1980).

Arthur, R. (1999). On thought experiments as *a priori* science. *International Studies in the Philosophy of Science 13*(3), 215–229.

Baum, E. B. (2004). *What is Thought?* Cambridge, MA: Bradford Books/MIT Press.

Brooks, R. A. (2002). *Robot: The Future of Flesh and Machines*. London, UK: Penguin.

Brown, J. R. (1991). *The Laboratory of the Mind: Thought Experiments in the Natural Sciences* (1993 paperback ed.). London and New York: Routledge.

Bunzl, M. (1996). The logic of thought experiments. *Synthese 106*(2), 227–240.

Cole, D. (1984). Thought and thought experiments. *Philosophical Studies 45*, 431–444.

Cole, D. (1991). Artificial intelligence and personal identity. *Synthese 88*, 399–417.

Copeland, B. J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford, UK: Blackwell.

Copeland, B. J. (1996). What is computation? *Synthese 108*(3), 335–359.

Copeland, B. J. (2002). The Chinese room from a logical point of view. See Preston and Bishop (2002), pp. 109–122.

Dennett, D. (1980). The milk of human intentionality. *Behavioral and Brain Sciences 3*(3), 428–430. (Peer commentary on Searle, 1980).

Dennett, D. C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown and Company.

Einstein, A., B. Podolsky, and N. Rosen (1935). Can quantum mechanical description be considered complete? *Physics Review 47*, 777–780.

French, R. M. (2000). The Chinese room: Just say "no"! In *Proceedings of 22nd Annual Cognitive Science Society Conference*, Philadelphia, PA, pp. 657–662.

Gendler, T. S. (2000). *Thought Experiment: On the Powers and Limits of Imaginary Cases*. New York, NY: Garland Press.

Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence 1*(1), 5–25.

Harnad, S. (1994). What is computation (and is cognition that) – Preface. *Minds and Machines 4*(4), 377–378.

Harnad, S. (2002). Minds, machines and Searle 2: What's wrong and right about the Chinese room argument. See Preston and Bishop (2002), pp. 294–307.

Haugeland, J. (2002). Syntax, semantics, physics. See Preston and Bishop (2002), pp. 379–392.

Hofstadter, D. (1980). Reductionism and religion. *Behavioral and Brain Sciences 3*(3), 433–434. (Peer commentary on Searle, 1980).

Hofstadter, D. R. and D. C. Dennett (1981). *The Mind's I: Fantasies and Reflections on Self and Soul*. Brighton, UK: Harvester Press.

McFarland, D. and T. Bösser (1993). *Intelligent Behavior in Animals and Robots*. Cambridge, MA: Bradford Books/MIT Press.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review 83*(4), 435–450.

Preston, J. (2002). Introduction. See Preston and Bishop (2002), pp. 1–50.

Preston, J. and M. Bishop (Eds.) (2002). *Views into the Chinese Room: Essays on Searle and Artificial Intelligence*. Oxford, UK: Clarendon Press.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (Ed.), *Language, Mind and Knowledge*, pp. 131–193. Minneapolis, MN: University of Minnesota Press.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: Bradford Books/MIT Press.

Schank, R. C. and R. P. Abelson (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Searle, J. (1984). *Minds, Brains and Science: The 1984 Reith Lectures*. London, UK: Penguin.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417–457. (Including peer commentary).

Searle, J. R. (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association 64*, 21–37.

Searle, J. R. (1997). *The Mystery of Consciousness*. London, UK: Granta.

Smith, B. C. (2002). The foundations of computing. In M. Scheutz (Ed.), *Computationalism: New Directions*, pp. 23–58. Cambridge, MA: Bradford Books/MIT Press.

Sorenson, R. A. (1992). *Thought Experiments*. New York, NY: Oxford University Press.

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

Wilensky, R. (1980). Computers, cognition and philosophy. *Behavioral and Brain Sciences 3*(3), 449–450. (Peer commentary on Searle, 1980).

Wilkes, K. V. (1988). *Real People: Personal Identity without Thought Experiments*. Oxford, UK: Clarendon.