The Logic of Searle's Chinese Room Argument*

R. I. Damper
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK.
email `rid@ecs.soton.ac.uk`

Abstract

John Searle's Chinese room argument (CRA) is a celebrated thought experiment designed to refute the hypothesis, popular among artificial intelligence scientists and philosophers of mind, that "the appropriately programmed computer really is a mind". Since its publication in 1980, the CRA has evoked an enormous amount of debate about its implications for machine intelligence, the functionalist philosophy of mind, theories of consciousness, etc. Although the general consensus among commentators is that the CRA is flawed, and not withstanding the popularity of the systems reply in some quarters, there is remarkably little agreement on exactly how and why it is flawed. A newcomer to the controversy could be forgiven for thinking that the bewildering collection of diverse replies to Searle betrays a tendency to unprincipled, ad hoc argumentation and, thereby, a weakness in the opposition's case. In this paper, treating the CRA as a prototypical example of a 'destructive' thought experiment, I attempt to set it in a logical framework (due to Sorensen), which allows us to systematise and classify the various objections. Since thought experiments are always posed in narrative form, formal logic by itself cannot fully capture the controversy. On the contrary, much also hinges on how one translates between the informal everyday language in which the CRA was initially framed and formal logic and, in particular, on the specific conception(s) of possibility that one reads into the logical formalism.

*Based on a paper presented at International Congress on *Thought Experiments Rethought*, Centre for Logic and Philosophy of Science, Ghent University, Belgium, 24–25 September 2004.

Keywords: Chinese room argument, modal logic, philosophy of mind, strong AI, thought experiments

1 Background

The Chinese room argument (CRA) is a celebrated thought experiment due to John Searle. First formalised in 1980 in a target article in *Behavioral and Brain Sciences*, it was designed to show the futility of the search for 'strong' artificial intelligence (AI). The CRA has stirred up an enormous amount of debate and controversy among AI scientists and engineers, philosophers of mind and cognitive scientists. Gomila (1991, footnote 9, p. 88) describes the literature on the CRA as "nearly infinite", and the editor at the time (Stevan Harnad) has since described it as "BBS's most influential target article ... as well as something of a classic in cognitive science" (Harnad, 2002, p. 295). Nor does the controversy show signs of abating, as evidenced by the more recent appearance of *Views into the Chinese Room*, a collection of essays on the CRA edited by Preston and Bishop (2002), and *John Searle*, a collection inspired by Searle's philosophy and edited by Smith (2003) in which the CRA figures prominently (see especially chapter ten of Moural 2003).

On the face of it, it might appear strange that the debate continues since the general consensus of virtually all commentators is that the CRA is flawed[1]. Why continue to argue if (almost) all agree on this point? The answer is that controversy still rages over exactly how and why the CRA is deficient. Certainly, many commentators *claim* to have (or to know of) the definitive counter. For instance, Dennett refers to "the definitive refutation, still never adequately responded to by Searle" (Dennett, 1991, footnote 2, p. 436). As Harnad writes, the CRA has not only "challenged the computational view of mind", it has also "inspired in many respondents the conviction that they have come up with decisive, knock-down counterarguments" (Harnad, 1989, p. 5). Yet the reality is that the debate continues, often with the focus of attention switched to the adequacy of the counterarguments. Is it then even possible to reach a satisfactory conclusion? Perhaps it is just the case that the CRA raises questions that we are not yet in a position to answer, or is somehow ill-formed. Although some AI researchers have given up on the CRA as fruitless, "false — and silly"[2], such dismissiveness is no replacement for rational appraisal of its scientific status.

In this paper, I explore the extent to which recastings of the CRA from its original informal (natural language) version into a more formal structure of modal logic might allow us to assess its status and, thereby, to systematise and classify the various counters to it. This, I contend, lends much-needed clarity to the debate. Given that the CRA is widely-recognised as a prototypical thought experiment, it seems worth analysing it as such, especially with respect to its logical form. Sorensen (1992) provides a useful framework for this in terms of modal logics, and I will utilise his work extensively.

Although extremely well known, it is necessary at the outset to give a brief outline of the argument for completeness. Searle envisages a situation in which he is hidden in a room and is presented questions in Chinese written on an 'input' card, posted in to his room by unseen enquirers. Searle knows no Chinese; indeed, he is quite unaware of the enterprise in which he is engaged and is ignorant of the fact that the strange marks on the cards represent questions framed in Chinese. He consults a manual telling him (in English) precisely what equally strange marks to write on an 'output' card, which he posts back to the outside world. By virtue of the 'machine intelligence' embodied in the manual (which is actually a formalisation of the steps in an AI program), these marks on the output card constitute an answer to any input question. To a Chinese speaker external to the room, by virtue of its question answering ability, the system passes the Turing test for machine intelligence (Turing, 1950), yet the system implemented by Searle-in-the-room is entirely without understanding simply because Searle understands nothing.

Searle concludes that an AI program could give the impression of intelligence to an external observer, but have no understanding. This is contrary to the tenets of strong AI — essentially that computational states are functionally equivalent to mental states — as exemplified in the (then contemporary) work of Schank and Abelson (1977), McCarthy (1979), Newell (1980) and others.

2 Standard replies to the CRA

Like the CRA itself, the standard replies to it are well known but are very briefly outlined here for completeness. In his 1980 BBS article, Searle identifies (and answers) the following objections:

- **Systems reply:** intelligence resides in the total system not just in Searle himself, who is merely a component.
- **Robot reply:** if we replace the disembodied AI program by a robot with sensors, effectors, etc., this is 'intelligent' in just the way that a human is.
- **Brain simulator reply:** instead of the AI program, let us simulate the actual neurons of a Chinese person answering questions in Chinese. This would be an 'intelligent' computer program.
- **Combination reply:** some combination of all of the above counters adds up to a refutation of the CRA.

Searle identifies two other replies in his 1980 article, although they have received less attention subsequently:

- **Other minds reply:** how do we know that Searle understands English when he claims to? Only because humans naturally ascribe intelligence ('intentionality') to other humans. So if we ascribe intentionality to Searle, we must do the same for the computer program. (This is, of course, the essential reasoning behind the Turing test anyway.)
- **Many mansions reply:** there are many possible kinds of 'computer' and 'computation'. In the future, we may have yet unimagined forms of 'computer' that *would* display artificial intelligence.

The other minds reply is said by Searle to "miss the point" because AI and cognitive science must "presuppose the reality and knowability of the mental" (pp. 421–2), while the many mansions reply implicitly accepts that strong AI is *not* just symbol manipulation.

2.1 Systems reply

In various guises, this is easily the most popular counter against the CRA. It can be succinctly stated thus: *Because a part of the system (i.e. Searle) does not understand Chinese, this does not mean that the complete system does not understand Chinese.*

Searle claims to have the decisive rebuttal which he calls the 'outdoor' CRA. He argues that he simply (!) 'internalizes' everything, committing the entire manual to memory, and then proceeds as before. There is then, he argues, nothing but the human in the system who *still* does not understand Chinese. How does he know this? Because the human in question is Searle himself, and he knows that he does not understand Chinese. The popular rejoinder by proponents of AI to this manoeuvre is to point out that it begs the question by assuming the truth of the CRA. Searle responds by arguing that the systems reply itself begs the question in the first place by assuming without argument that the system understands. And in any case, says Searle, he must be the ultimate arbiter of what he does and doesn't know and understand. He insists that he does not understand Chinese, and memorising a "meaningless" (to him) manual of instructions cannot alter this "fact".

2.2 Robot reply

The nub of this counter is to assert that Searle makes an error in the CRA by viewing intelligence — exemplified by question-answering alone — as 'disembodied' and thereby disconnected from the wide spectrum of interlocking abilities that together conspire to produce a cognitive agent (see Lycan 1980 and Russow 1984 for early commentaries along these lines). Adding sensors, effectors, etc. produces a causal link with the world and, so some of Searle's opponents say, such a link is essential to intelligence. The notion that (to cite McFarland and Bösser 1993, p. 271) "intelligence requires a body" has been enormously influential in AI post-1980 — see for instance Clark (1987), Brooks (1999, 2002) and Pfeifer and Scheirer (1999). For his part, Searle believes that the addition of sensors, effectors, etc. changes nothing. For him, the 'computer inside the robot' continues to manipulate uninterpreted symbols and still does not understand Chinese. To the AI proponent, as before in the case of the systems reply, this rejoinder begs the question by assuming the success of the CRA. But according to Searle, the boot is on the other foot. As before, the robot reply begs the question by assuming that the robot understands!

2.3 Brain simulator reply

According to this counter to the CRA, we suppose that each brain cell of a Chinese person engaged in understanding Chinese is replaced by a faithful computer simulation of that neuron's action. That, so the counter goes, *would* be an AI program that understood Chinese, since it is operating on precisely the same principles as a human brain engaged in question answering in Chinese. But according to Searle, this implicitly abandons the physical symbol system (PSS) hypothesis (Newell, 1980) which is the cornerstone of the functionalist conception of intelligence and mind. It abandons it because the PSS hypothesis is based on the axiom that one doesn't *have* to know how the brain works to replicate its function. Searle's opponents object that the point of the brain simulation is that this is a computer program that he surely *has* to admit is intelligent. However, Searle states clearly at the outset his belief that machine intelligence is indeed possible, albeit on the somewhat shaky grounds that the brain is itself a 'machine'. Hence, he is unimpressed with this rejoinder. In his view, the CRA still applies. The simulation simulates the wrong thing about the brain — its formal symbol manipulation properties — and misses the important thing, namely its causal properties. Unfortunately, however, he is vague in the extreme about what exactly "its causal properties" might be (Sloman and Croucher, 1980; Haugeland, 2002; Damper, 2004), at least in his 1980 BBS article. In Searle (1983), he does attempt to develop his ideas of a biologically-based causation of intensional states in some detail. But see Jacquette (1989) for a criticism of this approach. As Jacquette writes:

> "What Searle does say about the causal powers of the mind is uninformative and disappointing. The difficulty is not that Searle prematurely owes his audience a scientific explanation of the exact causal mechanisms of the brain resulting in the production of intentionality, but that there are conceptual difficulties about the causal model he presents . . . " (p. 613).

2.4 Combination reply

This is the notion that the systems reply, robot reply, etc. by themselves may be insufficient but that they somehow *add up* to a convincing refutation of the CRA. I have never been able to take this seriously as a putative counter-argument. Why would anyone think that the logical union of two or more invalid arguments could ever be valid? I mention it here only for completeness; it will not be considered further.

3 Thought experiments in general

Having briefly outlined the (mostly) well-known background to the CRA, I propose to consider the following question: Can we gain insight into the CRA by considering it as a specific instance of the general class of thought experiments? Some years ago, Brown (1991, p. x) wrote: "... there is very little literature on the subject of thought experiments. This lamentable state of affairs is about to change radically". The radical change that Brown foresaw was the publication of the volume of collected works edited by Horowitz and Massey (1991). Since 1991, we have seen other valuable additions to the literature by Gomila (1991), Sorensen (1992), Bunzl (1996), Norton (1996), Häggqvist (1996), Arthur (1999), Gendler (2000), Peijnenburg and Atkinson (2003), Souder (2003) and others, to the extent that we now have a small but increasingly coherent body of prior art on the topic.

It is notoriously difficult to define a thought experiment unequivocally, because in effect thought experiments merely form a loose class of vaguely similar arguments[3]. A useful view is offered by Reiss (2002, p. 18) who writes: "... in thought experiments, we experiment on *us*, not on nature", but this obviously falls short of a definition. Arguably, we recognise a thought experiment by its possession, more or less, of certain stereotypical features. Sorensen (1992) provides just such a listing of stereotypical features. He also offers some logical structures for common forms of thought experiment, and some *fallacies* and *antifallacies* often encountered in reasoning about thought experiments. I will use Sorensen's structure and insights extensively in what follows.

3.1 Thought experiments in philosophy of mind

The view is often expressed that thought experiments can be useful in physical sciences, but are much less so in the humanities. Those who express this view frequently cite Wilkes (1988, p. 2), who writes of thought experiments that "[in] philosophy ... and in particular in the domain of the philosophy of mind, they can be ... both problematic and positively misleading". Since, as Reiss says, they involve experimenting "on *us*, not on nature", thought experiments work by evoking intuitions. Famously, Dennett (1980) has referred to the CRA as 'an intuition pump' but Searle maintains that his CRA is nothing to do with intuition but rather hinges on the *fact* that he does not understand Chinese. However, few commentators seem to agree with him. For instance, Gomila (1991) writes of thought experiments that "their contribution consists of the explicit formulation of our intuitions by resort to boundary situations" (p. 87), and Wakefield (2003) states "the CRA as presented by Searle is an unsound argument that relies on a question-begging appeal to intuition" (p. 285). The further the imagined scenario is from everyday experience, and the less well defined it is, the less reliable are these intuitions.

Again, quoting Reiss (2002):

> "Thought experiments do not have a life of their own ... because of the following dilemma. We either have experience with a situation that is relevant to the thought experimental situation or we don't. If we do, then our thought experiment just tells us what we know from concrete experiments (or observations). But if *[we]* don't, then the thought experimental result amounts to mere guesswork." (p. 12)

Brown (1991, p. 31) is "inclined to think that there is enough background information to legitimise (in principle) ... Searle's Chinese room thought experiment" and thereby to stop it being mere guesswork — but is there? It has only ever been sketched out briefly and in informal language by Searle. Given this, how can one fill in any necessary "background information" without an appeal to prior experience? The problem with many (perhaps most) thought experiments in AI, cognitive science and philosophy of mind is that that *no one* can possibly have experience with the envisaged situation because it is too far-fetched and/or fanciful. According to Peijnenburg and Atkinson (2003,

p. 305), thought experimenters "like Jackson, Searle and Putnam do not eschew the most bizarre accounts of zombies, swapped brains … and famous violinists who are plugged into another body". As Damper (2006) writes: "thought experiments can be harmful".

3.2 Is the CRA a typical thought experiment?

Sorensen (1992, p. 208) lists stereotypical features of a thought experiment as follows:

- **Autonomy**: a stereotypical thought experiment is remote from implementation (or one might as well execute the *actual* experiment);
- **Mental imagery**: it evokes rich, vivid internal portrayal of the imagined scenario;
- **Bizarreness**: the scenario is more or less fanciful, making it obvious that execution is not intended.

By this characterisation, the CRA must be considered very typical. The envisaged scenario in which Searle manually simulates (in real time!) a program for Chinese natural language understanding, capable of passing the Turing test, is pretty well as remote from implementation with current capabilities in AI as one could conceive. Throughout its history, it has also evoked a rich array of mental imagery, as the many colorful accounts of the CRA in introductory texts on AI, popular science books and television science programs attest. As for bizarreness, Sorensen presumably has in mind something other than that actual implementation "is more or less fanciful", or this feature would become identical to autonomy. But even if we could implement the Chinese-understanding program automatically on some computing machine, so that autonomy is no longer an issue, hand-implementation of this program, as required for the outdoor CRA, surely remains a pretty bizarre proposition.

3.3 Logical structure of thought experiments

Of course, much of the difficulty in defining thought experiments unequivocally stems from the fact that they exhibit a great variety of type and form. Can we therefore say anything useful about thought experiments in general and, in particular, about their logical structure? Brown (1991, pp. 33) draws a distinction between constructive and destructive thought experiments. A destructive thought experiment is, as the name suggests, "an argument directed against a theory" (p. 34). Clearly, the CRA is of this type — the theory[4] is strong AI — and there seems no reason why we cannot say useful things about the logical structure of destructive thought experiments, especially typical ones.

Indeed, Peijnenburg and Atkinson (2003, p. 306) offer the following simple (modus tollens) structure of a destructive thought experiment in terms of propositional logic:

$$\begin{array}{l} (T \wedge E) \Rightarrow S \\ \neg S \\ \hline E \vdash \neg T \end{array} \tag{1}$$

Here, T is a theory, E is a thought experiment and S is some situation which "everyone knows" is not the case. In the Chinese room argument, T is strong AI and E is the CRA itself. The premise is that T and E together imply S, namely that Searle understands Chinese. But, so the argument goes, "everyone knows" that Searle does not understand Chinese, $\neg S$. Hence, $\neg T$ and strong AI is false.

In their footnote 1, Peijnenburg and Atkinson note the suggestion that a modal formulation might be preferable to this simple, propositional one but reject it as unnecessary for their purposes. In the remainder of this paper, I aim to demonstrate that logical schema (1) is too "coarse-grained" and a modal rendition is indeed more insightful. Quoting *The Stanford Encyclopedia of Philosophy* (`http://plato.stanford.edu/entries/logic-modal/`, last accessed 29 November 2005): "A modal is an expression (like 'necessarily' or 'possibly') that is used to qualify the truth of a judgement"[5]. Conventionally, the modalities 'it is necessary that' and 'it is possible that' are symbolised as $\Box$ and $\Diamond$, respectively. In line with everyday intuitions about necessity and possibility, these modalities are not independent but are related via the definition:

$$\Diamond A \stackrel{\text{def}}{=} \neg \Box \neg A \tag{2}$$

for statement A (Gabbay, 1998, p. 218). They also satisfy a de-Morgan-like relationship whereby:

$$\neg\Diamond A \equiv \Box\neg A \qquad \text{(negating (2))} \tag{3}$$

$$\neg\Box A \equiv \Diamond\neg A \qquad \text{(replacing } A \text{ by } \neg A \text{ in (2))} \tag{4}$$

from which we quickly obtain the alternative form of (2):

$$\Box A \stackrel{\text{def}}{=} \neg\Diamond\neg A$$

As well as these definitions, we also need axioms to extend propositional logic to a workable modal logic. The most commonly-used (alethic) modal logic is S5 (C. I. Lewis 1918) in which the added axioms are: $\Box A \Rightarrow A$, $\Box A \Rightarrow \Box\Box A$, $\Diamond A \Rightarrow \Box\Diamond A$. The first of these asserts that what is necessary is the case. The second asserts that any string of boxes can be replaced by a single box, and similarly for diamonds. That is, repeated application of modal operators is unnecessary; to say that A is necessarily necessary is redundant. Finally, any string of boxes/diamonds in S5 is equivalent to the last modal operator in the string. Alternatively, we can replace this by $A \Rightarrow \Box\Diamond A$; that is, whatever is the case is necessarily possible.

Now the CRA is centrally concerned with the necessity that *any* implementation of a Chinese-understanding program understands Chinese, and with the possibility that a human (Searle-in-the-room) can hand-implement this program. Hence, contrary to Peijnenburg and Atkinson's view, a formulation in modal logic for this (and for other thought experiments with similar structure) seems at least worthwhile if not mandatory. Again, quoting *The Stanford Encyclopedia of Philosophy*, "… modal logic is particularly valuable in the formal analysis of philosophical argument, where expressions from the modal family are both common and confusing"[6].

Sorensen (1992, Chapter 6) gives two typical structures for the class of thought experiment that he calls "alethic refuters". By use of the term 'refuters', he is clearly focusing on destructive thought experiments that refute some theory. By 'alethic', he refers to the modalities *necessary* and *possible* as in S5. The two kinds of structure are "necessity refuters" and "possibility refuters", with the former much more common or typical than the latter. We will, therefore, deal with the necessity refuter first.

Its structure is (Sorensen, 1992, pp. 135–6):

1.	S :	Modal source statement
2.	$S \Rightarrow \Box I$:	Modal extractor
3.	$(I \wedge C) \mathrel{\Box\!\!\rightarrow} W$:	Counterfactual
4.	$\neg \Diamond W$:	Absurdity
5.	$\Diamond C$:	Content possibility

As a note of caution, because I wish to retain the notation used by Sorensen, the symbols here have obviously different meanings to those in (1) above; for example, S here is the 'theory' denoted T by Peijnenburg and Atkinson. It is the form of the modal extractor, Statement 2, in terms of the necessity of implication I that accounts for the description "necessity refuter".

In the above, $A \mathrel{\Box\!\!\rightarrow} B$ denotes the subjunctive conditional — famously discussed by David Lewis (1973) in relation to 'possible world semantics' — interpreted as 'if A *were* the case, then B would be the case'. By his use of the description "Counterfactual", Sorensen seems implicitly to accept the view that the subjunctive conditional is necessarily counterfactual, but this has been widely questioned (e.g. Bennett 2003). Indeed, as Bennett makes clear, it has to be said that the semantic interpretation of subjunctive conditionals, and determining a precise relation to the indicative conditional (or material implication), $A \Rightarrow B$, is the subject of enormous debate. Since it would take us too far from the CRA, and because I am not qualified to do so anyway, I avoid any serious discussion of this debate here.

Sorensen asserts it to be obvious that Statements 1 to 5 are mutually contradictory, but it seems to me that the difficulty of interpreting the subjunctive conditional complicates this simple, direct inference. A formal confirmation of contradiction would require that we made explicit the logical schema, rules of inference, means of determining truth values, etc. appropriate to $A \mathrel{\Box\!\!\rightarrow} B$, which is not straightforward (because its semantics is controversial). That said, I believe that few commentators would question Sorensen's assertion on a pragmatic reading. Accepting the contradictory nature of the statements at face value, one cannot hold all five simultaneously. So (ruling out conjunctions) there are at most five distinct and consistent responses to the thought experiment. The thought experimenter's intent will be to refute the source statement, yet up to four alternative counters are available that retain the validity of source statement. Sorensen (1992, pp. 136–52) labels the set of five possible responses to this alethic refuter as:

1. Bad source statement.
2. Misconnection.
3. Erroneous counterfactual.
4. Pseudo-absurdity.
5. Impossibility theorem.

4 The CRA as a necessity refuter

In the particular case of the Chinese room argument, Statements 1 to 5 are interpreted as follows:

1. S : The modal source statement is the ‘theory’ of strong AI, namely that executing an AI program is necessarily constitutive of understanding.
2. $S \Rightarrow \Box I$: The modal source statement S implies the logical necessity that any implementation of the Chinese-understanding program understands Chinese, in accordance with the tenets of strong AI.
3. $(I \wedge C) \mathrel{\Box\!\!\rightarrow} W$: If Searle were to hand-implement any such program then he would understand Chinese — a ‘weird’ consequence, W.
4. $\neg \Diamond W$: Searle does not understand Chinese (this is an ‘absurdity’).
5. $\Diamond C$: It is possible that Searle can hand-implement the Chinese understanding program.

As we have seen for the general case, since Statements 1 to 5 are mutually contradictory, at least one of these must be false. The purpose of alethic refuters is to refute the source statement; hence, Searle presents the CRA as disproving Statement 1: Executing an AI program is *not* necessarily constitutive of understanding.

4.1 Countering the CRA

The majority of counters over the years have attempted to preserve the modal source statement S. That is, they aim to uphold some concept of machine intelligence, not necessarily identical to Searle's conception of strong AI but retaining at least some aspects of a computational 'theory' of mind. Any counter of this type is therefore (implicitly) a refutation of one or more of Statements 2 to 5, perhaps in terms of the translation between informal natural language and formal logical statements. The framework provided by Sorensen thus allows us to systematise and classify the apparently many and various objections to the CRA. In spite of their wide-ranging appearance, there should in fact be no more than five categories of counter! One of these (see Section 4.2 below) will in effect challenge the thought experimenter's form of the source statement.

The popular systems reply and what I take to be its variants (such as Hofstadter 1980; Weiss 1990; Copeland 1993; 2002a) are in the 'erroneous counterfactual' category[7]. In effect, the systems reply holds that Searle's view, which equates himself with the AI program in the counterfactual 3, is too narrow. It is not Searle who must necessarily understand Chinese if strong AI is true, but the wider system of which he is a part. And this, declare the proponents of AI, is indeed the case.

Some commentators (e.g. Abelson 1980; Cole 1984; 1991; Damper 2004) attack the absurdity claim 4, arguing that it is not at all clear that Searle would not understand Chinese in the envisaged scenario. As Cole (1984, p. 431) writes "it is not clear, despite Searle's denials, that his imagined simulation of a machine would not produce understanding". This is only Searle's intuition; it is not shared by all. And as outlined above, the central part played by intuition is an Achilles' heel for thought experiments. It is noteworthy that Searle did not identify this counter in his 1980 article and so it has not been glorified with a universally-recognised name in the literature, but Sorensen's 'pseudo-absurdity' seems as good as anything. It is possible that Searle did not christen this counter back in 1980 because he genuinely had not met it, although one suspects (from the strength of his insistence over the years that he really does not understand Chinese and he alone is the ultimate authority on what he does and does not understand) that he would have great difficulty recognising its existence even now.

Still others attack the content possibility 5, for example French (2000a), Brooks (2002) and Damper (2004). This appeal to what Sorensen calls the 'impossibility theorem' amounts to an argument about what it is and is not possible for Searle to do. Can he really hand-simulate the Chinese-understanding program, even in principle, and remain the same old John Searle that he believes and recognises himself to be? In an early and influential counter, Hofstadter and Dennett (1981) wrote:

> "We think Searle has committed a serious and fundamental misrepresentation by giving the impression that it makes any sense to think that a human being could do this. By buying this image, the reader is unwittingly sucked into an impossibly unrealistic concept of the relation between intelligence and symbol manipulation." (p. 373)

In fact, they present this as an introductory comment preparatory to outlining what amounts to a form of the systems reply, arguing (p. 375) that "nearly all of the understanding must lie in the billions of symbols ..." in the AI program. Nonetheless, their remarks about the feasibility of Searle's imagined scenario remain apposite. In similar vein, French (2000a, p. 657) attacks 'the tacit assumption ... that somehow there could be such a Room'. While this too can be seen as an appeal to the 'impossibility theorem', French actually intends to *support* Searle's contention that symbol manipulation is not constitutive of intelligence, not via the CRA but by "showing that the Chinese Room itself would be an impossibility". That is, Searle cannot hand-implement a Chinese-understanding program capable of passing the Turing test because such a program makes no sense. No machine executing such a program could "pass the Turing test unless it had experienced the world as we humans have". In effect, French is attacking the concept of 'understanding' which pervades the CRA as something which can be operationalised via the Turing test. As such, he is rejecting the whole basis on which the necessity refuter form of argument is translated into natural language.

Many commentators argue (or appear to) that content possibility, Statement 5, is irrelevant to the CRA. For instance, Preston (2002) writes:

> "... the fact that the person in the room could not handwork the programs fast or reliably enough does not matter ... The Chinese Room is a thought experiment, an investigation into what would

> follow if something thoroughly counterfactual were to be the case ... In such scenarios, one is allowed to imagine what would happen if some contingent and variable limitation ... were idealized." (p. 25)

But (provided we accept it as having some value) the logical framework developed here indicates that the argument actually turns on the content possibility, $\Diamond C$. By no means are we at liberty to say that it "does not matter". Preston is confusing impossibilities that do not matter as far as the content of the thought experiment is concerned with those that do matter, and it is one virtue of the formalism presented in this paper to make this clear (see Section 9 later for further discussion).

Thus far we have considered challenges to Statements 3, 4 and 5 as alternatives to the refutation of the source Statement 1, in the form of 'erroneous counterfactual', 'pseudo-absurdity' and 'impossibility theorem' manoeuvres. What about the 'misconnection' counter? Has this ever been deployed against the CRA? I am not aware that it has, nor is it easy to see how it might be. The modal extractor $S \Rightarrow \Box I$ draws the relevant modal (necessary) implication from the source statement — in the case of the CRA, that the theory of strong AI implies the necessity of machine understanding. Häggqvist (1996, pp. 100–2) argues persuasively that the modal extractor is *not* open to challenge, since the necessity implication should properly be seen as part and parcel of the modal source statement[8].

4.2 Challenging the modal source statement

A minority of commentators happily accept that the CRA is decisive against its target of strong AI. Interpreting Searle's modal source statement to reflect his conception of strong AI, they implicitly agree with the argument's logical form and with Statements 2 to 5. But, they say, Statement 1 is a straw man conception of AI (e.g. Wilks 1982). It embodies a misunderstanding or misrepresentation of what AI is really about. The robot reply in effect takes this stance. The disembodied program as described by Searle would not understand, because intelligence and understanding require causal connection ('grounding') with the outside world. The brain simulator reply and the many mansions reply are also in this category. The former advances a different conception (to symbolic, rule-based processing of the kind underlying the PSS hypothesis) of the internal workings of an AI program whereas the latter poses the possibility of new and more powerful forms of computer and computation in the future, perhaps like the 'Super-Turing' or 'Hypercomputation' ideas that one finds in the work of Siegelmann (1999) and Copeland (2002b).

Harnad (1989, 2002) has consistently argued over many years that the CRA really is an attack on a straw man conception of AI that no informed person would seriously hold. He takes the CRA seriously but believes it *should* have been deployed against a conception of AI that many informed people do believe in, but is (he says) false. This false conception, argues Harnad, is *computationalism*: "the hypothesis that cognition is the computation of functions" (Dietrich, 1990, p. 135)[9]. Against this target, he says, the CRA is decisive[10]. On the one hand, the virtue of a counter such as Harnad's is that it forces us to think hard about what we mean by 'artificial intelligence'. On the other hand, on the basis of painful past experience, many will doubtless think that attempts to define AI are sterile, witness the decades of debate on the Turing test that attempts to sidestep the need for a definition (e.g. Moor 1976; French 1990, 2000b; Copeland 2000; Saygin *et al.* 2000). To press home the point, recall that Newell (1973) wrote "Sciences are not defined, they are recognized" (p. 1), whereas according to Wilensky (1983, p. xii): "Artificial intelligence is a field renowned for its lack of consensus on fundamental issues", and Häggqvist (1996, p. 17) writes of "a theoretical pluralism characteristic of newly cultivated intellectual fields".

5 The CRA as a possibility refuter

According to Sorensen (1992, pp. 153), a minority of thought experiments has a slightly different form and can be described as possibility refuters. Since Searle framed the CRA in informal, natural language rather than using the language of logic, his argument could be said not to have a precise logical form[11]. Its translation to logical form is a matter of interpretation. Thus, it is interesting to see if we can interpret it as a possibility refuter, which has the following structure:

1.	S :	
2′.	$S \Rightarrow \Diamond I$:	Possibility extractor
3.	$(I \wedge C) \mathrel{\Box\!\!\rightarrow} W$:	
4.	$\neg \Diamond W$:	
5′.	$\Diamond I \Rightarrow \Diamond (I \wedge C)$:	Content copossibility

where the changes from the necessity refuter are the possibility extractor and content copossibility, marked with primes.

In the specific context of the CRA, Statement 2′ (the possibility extractor) would assert that strong AI implies that it is *possible* for a Chinese-understanding computer program actually to understand Chinese. This is much weaker than the necessity extractor (Statement 2 of the necessity refuter) and it is also apparently far weaker than Searle intended. It is perhaps a good description of the position of those adversaries of Searle who favor the brain simulator or many mansions replies; what Searle says is necessary (strong AI means that he would understand Chinese when hand-implementing the right program) is not necessary but merely possible in some cases: $\neg \Box I \equiv \Diamond \neg I \not\equiv \neg \Diamond I$. Searle likewise admits the possibility of machine intelligence in some form, provided this form has the right causal powers (whatever they are), but the difference is that this is not drawn as an implication from strong AI.

Statement 5′ (content copossibility) would assert that if some Chinese-understanding program exists, then it is possible to hand-simulate it. Presumably Searle would be happy with Statement 5′ as a part of the CRA, since he apparently feels he can (in principle) hand-implement *any* program. If there is a problem with viewing the CRA as a possibility refuter, it lies with the possibility extractor (Statement 2′) rather than the content copossibility (Statement 5′). It does seem that the necessity refuter is the better characterisation of the CRA.

6 Searle rewrites the CRA

Searle has periodically rephrased the CRA, purportedly with the goal of clarification. For example, he has quite recently written (Searle, 2002, p. 52) that it "… rests on two absolutely fundamental logical truths", namely:

1. Syntax is not semantics.

2. Simulation is not duplication.

But these are surely more assertions than 'truths' — things that might (perhaps) follow should the CRA be valid. They have themselves led to a widening of the debate (e.g. Rapaport 1986; Anderson 1987; Ben-Yami 1993; Haugeland 2002; Wakefield 2003) and in a direction which sometimes seems to me to take us away from the nub of the CRA. So, rather than representing a clarification, this looks more like a different (if strongly related) and new argument. Indeed, Melnyk (1996) has even christened this "Searle's abstract argument" indicating that he too thinks it is different to the CRA proper, as I do.

Searle himself has apparently vacillated on whether this is a separate argument or not. In the original BBS article, he mentions only briefly and obliquely the issues of syntax and semantics (p. 422). In 1984, he seemingly still thought that these formed part of the CRA proper, see Searle (1984, pp. 32–3). By 1997, they had become separate arguments, "... one about Strong AI, the other about the existence of consciousness" (Searle, 1997, p. 129). By 2002, Searle has reverted to portraying questions of syntax, semantics and simulation as central to the CRA.

In my opinion, this 'rewriting' of the CRA is unhelpful, and so will not be considered further[12]. The main reason for mentioning it here is to warn the reader that it is a different argument, so as to keep our focus clearly on the CRA itself.

7 Fallacies and antifallacies

Thus far, we have examined the logical structure of thought experiments in general and of the CRA in particular. There is, however, an obvious issue about the translation between the informal natural language in which thought experiments are invariably framed and some tighter, more formal logical schema such as that developed in this paper. It can of course be the case that many arguments about thought experiments are in effect disagreements about this translation.

Sorensen (1992, Chapter 10) outlines a variety of fallacies and antifallacies commonly encountered in reasoning about thought experiments. I take these to be of interest and importance in the present context because they bear on the very translation just mentioned: between the informal (natural language) version of a thought experiment and a more formal logical schema. According to Sorensen, a *fallacy* is a bad rule of inference that looks good (a false positive), whereas an *antifallacy* is a good rule that looks bad (a false negative), and these can be used to separate bad thought experiments from good ones.

An especially common fallacy (i.e. hallmark of a bad thought experiment) — says Sorensen — is missupposition, which comes in two flavours:

- **Oversupposing**: In this case, the thought experimenter assumes too much and so "may inadvertently trivialize the very problem the thought experiment was intended to solve" (p. 257).
- **Undersupposing**: Here, "the designer of the thought experiment fails to be specific enough *[and]* the usual flaw is indecisiveness" (p. 258).

Stated in this way, I have to admit to a difficulty in separating these two putative 'flavours' of missupposition, since I take assuming too much and failing to be specific enough to be synonymous. Sorensen, however, seems to have in mind with oversupposing "the godlike power of stipulation" of the thought experimenter as the source of the trivialisation. Anyway, let us take what I think to be the simpler case first and see if and how it applies to the CRA.

7.1 The fallacy of undersupposing

As Sorensen writes:

> "When the designer of the thought experiment fails to be specific enough ... either the audience recognises the shortfall and complains about insufficient data, or they unwittingly read in extraneous details. If [they] supply diverging details, they become embroiled in a dispute or seduced into a consensus that is merely verbal" (pp. 258–9).

This seems to me to be a pretty accurate characterisation of more than two decades of inconclusive debate on the CRA.

Searle occasionally remarks on the simplicity and 'conciseness' of the CRA[13]. Yet this very conciseness is in my view no more than a symptom of undersupposing. In what respect(s) does Searle fail "to be specific enough"? His statement of the CRA is under-specified 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. In short, what is involved in specifying a 'Chinese understanding' program capable of passing the Turing test?

Searle, of course, believes that the CRA is decisive against any form of strong AI, machine functionalism, call it what you will. His argument is impervious to mere details, so supplying such details is none of his business — that's a job for the proponents of AI who not only think it makes sense to have the implementation of a natural language understanding computer program on their agenda, but that this program would literally have a mind. As he writes of the task of implementing such a program (Searle, 1980, p. 453): "I am not even sure it can be done". But apparently, he *is* sure that if it can be done, then he can hand-implement it (content possibility, $\diamond C$) and it will have the weird consequence W that he would understand Chinese when "everyone knows" he does not. This does seem to be a remarkably incautious position for anyone to hold, to say the very least.

7.2 The fallacy of oversupposing

If this is to be a distinctly different missupposition to undersupposing, then we should perhaps interpret Sorensen as intending that oversupposing trivialises the problem through some stipulation that assumes too much. He gives as an example John Locke's tale of the prince and the cobbler in which each wakes up one morning remembering the past associated with the other's body. However, as "you can only remember what you really did ... the thought experiment begs the question in favor of the psychological criterion of personal identity" (Sorensen, 1992, p. 257).

So does Searle err in this way and beg the question by oversupposing? Well, perhaps the outdoor CRA commits this fallacy, by stipulating that he internalises the Chinese-understanding program by rote learning, without understanding.

7.3 The far out antifallacy

Sorensen, in contradiction of the often-cited view of Wilkes (1988), holds that 'bizarreness' is a stereotypical feature of thought experiments, and does not constitute reason to reject the argument. To do so is a mistake that he dubs the "far out antifallacy". Without 'bizarreness', we could execute the real experiment and have no need of the thought experiment. Yet many sophisticated commentators (e.g. French 2000a; Brooks 2002) reject the CRA on the grounds that the imagined scenario is "ludicrous". It is not only Wilkes who apparently opposes Sorensen on this point, who for his part seems undaunted calling this the "master antifallacy ... the rich man's version" (p. 277), because of the frequency with which it appears in the literature.

So does the bizarreness of the CRA count against it or is this an antifallacy? I have to say that I side more with Wilkes and less with Sorensen on this point. The CRA, as we have seen, hinges on the content possibility, $\diamond C$. If the content of the thought experiment is so bizarre as to call into question its possibility, then this amounts to an attack on Statement 5 of section 3.3, which can be decisive. Although Sorensen agrees that an attack on content possibility is decisive, he adds that:

> "... people tend to equivocate by latching on to the wrong kind of impossibility ... An attack on a thought experiment that shows the supposition to be logically impossible is sure to be successful. But the choice of a weaker impossibility courts the danger of too weak a response." (p. 278)

So we confront a problem in seeking to use modal logic, hinted at in note 5. The problem is that logicians and philosophers recognise *different kinds* of possibility. So what are these different kinds and which one should we be using in thinking about the Chinese room?

8 What is possible and what isn't?

I have argued that imposing a formal logical structure on the Chinese room thought experiment is helpful if not essential if we are to assess its implications correctly, and understand the various counters that have been made to it over the years. Further, a natural formulation for this destructive thought experiment is in terms of (alethic) modal logic, using the modalities *necessary* and *possible*.

As Bunzl (1996, p. 228) writes: “questions of possibility and impossibility enter at the ground floor” where thought experiments are concerned. As the question of what exactly we mean by ‘possibility’ is so central, we can postpone its treatment no longer.

Logicians and philosophers recognise a bewilderingly large number of different kinds of possibility including logical, technical, metaphysical, physical, real, epistemic and so on (e.g. Hacking 1967; 1975; Seddon 1972; DeRose 1991 among others). Unfortunately, there is (to say the least!) a lack of consensus on how many different kinds there might be and/or exactly how they are different. However, by my reading of the literature, there is a measure of agreement on the nature and status of logical possibility and physical possibility, and these are probably also reasonably good delimiters of the range of qualification covered by the modality.

Logical possibility is described by Hacking (1975, p. 324) as “the logician’s favorite”. According to Wilkes (1988):

> “... something is logically possible if it is not ruled out by the laws of logic. So although it is not logically possible that $2+2=5$, it is logically possible that gold does not have atomic number 79, that water is not H_2O, that whales are fish ...” (p. 17)

Logically possibility is often tied to conceivability; that is, that which is conceivable without contradiction is logically possible (see Yablo 1993; Gendler and Hawthorne 2002) and to Leibniz’s notion of ‘possible worlds’[14]. A famous philosophical thought experiment that turns on the existence of a possible world in which water is not H_2O but XYZ is Putnam’s twin earth (Putnam, 1975), but the obvious problem with this is that we simply have no idea what else would have to be different between twin earth and our own for water to be XYZ. These differences might easily invalidate the thought experiment.

Physical possibility is what might occur in some world with the same physical laws as this one. Thus, it is physically possible that Bill Clinton was never the President of the USA, since this contingency is not decreed by the laws of nature. Disregarding isotopes (cf. heavy water), it is physically impossible that water could have a different chemical formula from H_2O. This is, I think, close to Lewis’s possible worlds semantics, where we are concerned with what would have been the case if the antecedent of a subjunctive conditional were true in some world other than this one, but all else, and *a fortiori* the laws of nature, remains the same. Obviously, physical impossibility is much 'weaker' (i.e. more restrictive in what it counts as possible) than logical possibility.

As we have seen, Sorensen's view is that an attack on content possibility in terms of logical possibility is decisive against a destructive thought experiment. But does logical possibility make good sense in the context of the CRA? For the purposes of the argument, it is necessary to postulate a world in which (counterfactually, according to Searle) strong AI is possible and also that it is possible for Searle to hand-implement the AI program. Regarding the former, it seems to me that Searle must have had in mind something much closer to physical than to logical possibility, since he never attacks strong AI on the basis that it is logically impossible but rather that a machine "defined as an instantiation of a computer program" (p. 422) is made of the "wrong kind of stuff" (p. 423) to have mental states; that is, it doesn't have the right "causal powers". Clearly, this is a point of physics, not logic. Since the two kinds of possibility must coexist in the same possible world for the CRA to have coherence, we are led to the conclusion that we also must interpret the content possibility (that Searle can hand-implement the AI program) in terms closer to physical than to logical possibility. According to Sorensen, this may be too weak. But surely, it is only too weak if refutation in terms of (something close to) physical possibility leaves open a 'gap' between physical and logical impossibility in which the thought experiment can survive unscathed. If such a gap exists in the case of the CRA, it cannot be very wide since Searle's argument seems not actually to be about anything very close to logical possibility. His attack is on AI scientists trying to construct artificial intentionality in *this* world.

We have concentrated thus far on logical and physical possibility, but perhaps this dichotomy is too coarse. Brooks (1994, p. 78) has argued that: "A notion of possibility intermediate between physical and logical possibility might be much more useful" in reasoning about thought experiments. In fact, Brooks was primarily concerned with thought experiments addressing personal identity, but perhaps this intermediate notion — which he calls *natural* possibility — could be useful in thinking about other philosophical thought experiments. A naturally possible world operates according to natural laws, but these are not necessarily the same as in our world (Brooks, 1994, p. 79). This is portrayed by Brooks as giving philosophers rein "to stretch their imaginations" but at the same time "to improve their intuitions of naturalness" (p. 82). The difficulty with this proposal as I see it (at least for potential application to the CRA) is that *whenever* we stray from physical possibility, we are at the mercy of unknown and unknowable differences between the possible world and our own, which render intuitions evoked by the thought experiment unreliable. We do not have to go as far as logical possibility for this to be a concern. To understand the CRA properly, I suggest, we must stick with physical possibility.

9 Summary and Conclusions

To summarise, the Chinese room argument (CRA) is a prototypical destructive thought experiment in the philosophy of mind. Searle claims that his argument is simple and obvious, but I believe the analysis presented in this paper shows that this is far from the case. Over the years, the CRA has been debated at great length with the majority opinion being that it is flawed; yet there is (notwithstanding the popularity of the systems reply in some quarters) little consensus on exactly how and why it is flawed. It has not been the purpose of this article to settle the issue one way or the other, either for/against the CRA or to pinpoint, in Harnad's words, "the decisive knock-down counterarguments". Rather, I have attempted a formal analysis of the CRA based on the logical structure provided by Sorensen. So what exactly does the codification in terms of modal logic bring to the debate, or is it just 'mathematizing the obvious'?[15] I maintain that there are essentially two beneficial outcomes.

First and foremost, it allows us to systematise and classify counterarguments to the CRA in a way that is much less transparent in absence of the analytical framework. We have seen that if we wish to refute the CRA (and assuming acceptance of the framework presented here) there is only a limited number of counters available if the commentator wishes to maintain the viability of some functionalist conception of intelligence and mind: We can attack one or more of Statements 2 to 5 of Sorensen's necessity refuter (Section 4). I share with Häggqvist, however, the view that Statement 2 is not really open to attack because the modal source statement is only meaningful to the extent that it is connected as antecedent to some (modal) consequent. An example drawn from the text should help make the point. We have encountered the statement of Preston (2002) to the effect that Searle's inability to hand-simulate the AI scientist's Chinese-understanding program "does not matter" because thought experiments are inherently concerned with "thoroughly counterfactual" situations. But the counterfactual of interest in the CRA is the conjunction $(I \wedge C)$ of the proposition that any implementation of such a program understands Chinese and the content. So Preston is misled into considering the *wrong* term (C in place of $(I \wedge C)$) as counterfactual, and the modal framework of this paper has the virtue of exposing the mistake.

Second, since thought experiments have a universally narrative character, phrased in informal everyday language, there will always be issues surrounding the translation from the thought experimenter's 'story' (the CRA in this case) into formal modal logic. One way this can be manifest is for a commentator to maintain the validity of Searle's argument, but hold that the conception of strong AI attacked is somehow flawed, as does Harnad (2002). This amounts to asserting mistranslation from Searle's narrative description of strong AI to the modal source statement S. Another aspect of the translation from everyday language to modal logic that invites scrutiny is the interpretation of the modal qualifier 'possible' as it impacts on content possibility. Logicians and philosophers entertain a large number of different kinds of possibility designed to licence exercise of the imagination without straying too far from the 'commonsense' notion of what is (physically) possible. I have argued that the CRA is actually founded on physical possibility, since Searle is centrally concerned with "causal powers" in *this* world.

Acknowledgement

I am indebted to David Atkinson, Alan Bundy, Martin Bunzl, Jack Copeland, Robert French, Sören Häggqvist, Stevan Harnad, Kieron O'Hara, Jeanne Peijnenburg, Adam Prügel-Bennett, Nigel Shadbolt and Aaron Sloman for critical comments on this paper, which helped me to improve it in clarity and presentation. These acknowledgements should not be taken to imply endorsement of the content.

Notes

[1] See Maloney (1987) for a noteworthy exception.

[2] ... a description attributed by Harnad (2002, p. 295) to Pat Hayes.

[3] The idea that thought experiments (in science, at least) are, in effect, *all* just arguments in another guise has been championed by Norton (1996). It is not universally accepted in the field.

[4] Of course, the term 'theory' is not intended to have the same force here, where we are dealing with the nascent science of the mental, as it would have in the mature natural sciences.

[5] We will postpone until Section 8 the semantic interpretation of 'possibility'. Since this modality is qualitative, we should expect that this semantic interpretation is not going to be entirely straightforward, to say the least.

[6] Actually, 'controversial in their interpretation' might be a better description than 'confusing'.

[7] I consider them variants because they are based on the so-called "part–whole fallacy" (see Haugeland 2002, p. 380) which I take to be a recasting of the systems reply. Note, however, that Copeland (2002a) is at pains to put distance between his 'logical reply' and the standard systems reply as follows. Whereas the systems reply begs the question by its assumption (without argument) that the system as a whole must understand Chinese, 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' (pp. 110–1).

[8] Häggqvist (personal communication) tells me that he intends this to be an argument *against* Sorensen's formalism, rather than something that supports the latter framework. That is, it is offered as something more than a minor correction.

[9] See also Scheutz (2002) for a recent survey.

[10] I do not myself see the wide gulf between computationalism and strong AI that Harnad obviously does, nor indeed between any of the many and various flavours of functionalism-cum-symbolism, but this is beside the present point.

[11] This, of course, is no accident. As Sorensen (1998) writes: "Thought experimenters never brutely stipulate that $\Diamond C$. Instead, they craft scenarios that are intended to *compel* assent." (p. 114) In writing this, he may well have had Searle in mind!

[12] ... except that I cannot resist quoting Anderson (1987) to the effect that "Searle proposed no ordinary simulation. From the outside, the simulation and the real thing were to be identical" (p. 390). This use by Searle of words sometimes to have a general, everyday meaning (as in "simulation is not duplication") but at other times (as in the CRA itself) to have a particular and special force is a typical pitfall of narrative thought experiments.

[13] In Searle (1997, p. 11), he writes: "This is such a simple and decisive argument that I am embarrassed to have to repeat it." Note, however, that his repetition is in terms of the abstract argument, not the CRA proper.

[14] ... enshrined in his famous fundamental theorem of optimism: "everything is for the best in this best of all possible worlds".

[15] I am grateful to Robert French for encouraging me to confront this question explicitly.

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).

Anderson, D. (1987). Is the Chinese room the real thing? *Philosophy*, **62**(3), 389–393.

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

Ben-Yami, H. (1993). A note on the Chinese room. *Synthese*, **95**(2), 169–172.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford University Press, New York, NY.

Brooks, D. H. M. (1994). The method of thought experiment. *Metaphilosophy*, **25**(1), 71–83.

Brooks, R. A. (1999). *Cambrian Intelligence*. Bradford Books/MIT Press, Cambridge, MA.

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

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

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

Clark, A. (1987). Being there: Why implementation matters to cognitive science. *Artificial Intelligence Review*, **1**(4), 231–244.

Cole, D. (1984). Thought and thought experiments. *Philosophical Studies*, **45**(3), 431–444.

Cole, D. (1991). Artificial intelligence and personal identity. *Synthese*, **88**(3), 399–417.

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

Copeland, B. J. (2000). The Turing test. *Minds and Machines*, **10**(4), 519–539.

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

Copeland, B. J. (2002b). Hypercomputation. *Minds and Machines*, **12**(4), 461–502.

Damper, R. I. (2004). The Chinese room argument: Dead but not yet buried. *Journal of Consciousness Studies*, **11**(5–6), 159–169.

Damper, R. I. (2006). Thought experiments can be harmful. *The Pantaneto Forum,*, **Issue 26**. `http://www.pantaneto.co.uk/`.

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

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

DeRose, K. (1991). Epistemic possibilities. *Philosophical Review*, **100**(4), 581–605.

Dietrich, E. (1990). Computationalism. *Social Epistemology*, **4**(2), 135–154.

French, R. M. (1990). Subcognition and the limits of the Turing test. *Mind*, **99**(393), 53–65.

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

French, R. M. (2000b). The Turing test: The first 50 years. *Trends in Cognitive Science*, **4**(3), 115–122.

Gabbay, D. (1998). *Elementary Logics: A Procedural Perspective*. Prentice Hall Europe, Hemel Hempstead, UK.

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

Gendler, T. S. and Hawthorne, J., editors (2002). *Conceivability and Possibility*. Clarendon Press, Oxford, UK.

Gomila, A. (1991). What is a thought experiment? *Metaphilosophy*, **22**(1–2), 84–92.

Hacking, I. (1967). Possibility. *Philosophical Review*, **76**(2), 143–168.

Hacking, I. (1975). All kinds of possibility. *Philosophical Review*, **84**(3), 321–337.

Häggqvist, S. (1996). *Thought Experiments in Philosophy*. Almqvist & Wiksell, Stockholm, Sweden.

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

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

Haugeland, J. (2002). Syntax, semantics, physics. In Preston and Bishop (2002), pages 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 Dennett, D. C. (1981). *The Mind's I: Fantasies and Reflections on Self and Soul*. Harvester Press, Brighton, UK.

Horowitz, T. and Massey, G., editors (1991). *Thought Experiments in Science and Philosophy*. Rowman and Littlefield, Lanham, MD.

Jacquette, D. (1989). Adventures in the Chinese room. *Philosophy and Phenomenology*, **49**(4), 605–623.

Lewis, C. I. (1918). *A Survey of Symbolic Logic*. University of California Press, Berkeley, CA.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge, MA.

Lycan, W. (1980). The functionalist reply (Ohio State). *Behavioral and Brain Sciences*, **3**(3), 434–435. (Peer commentary on Searle, 1980).

Maloney, J. C. (1987). The right stuff. *Synthese*, **70**(3), 349–372.

McCarthy, J. (1979). Ascribing mental qualities to machines. In M. Ringle, editor, *Philosophical Perspectives in Artificial Intelligence*, pages 161–195. Humanities Press, Atlantic Highlands, NJ.

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

Melnyk, A. (1996). Searle's abstract argument against strong AI. *Synthese*, **108**(3), 391–419.

Moor, J. H. (1976). An analysis of the Turing test. *Philosophical Studies*, **30**(4), 249–257.

Moural, J. (2003). The Chinese room argument. In Smith (2003), pages 214–260.

Newell, A. (1973). Artificial intelligence and the concept of mind. In R. C. Shank and K. M. Colby, editors, *Computer Models of Thought and Language*, pages 1–60. Freeman, San Francisco, CA.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, **4**(2), 135–183.

Norton, J. (1996). Are thought experiments just what you always thought? *Canadian Journal of Philosophy*, **26**(3), 333–366.

Peijnenburg, J. and Atkinson, D. (2003). When are thought experiments poor ones? *Journal for General Philosophy of Science*, **34**(2), 305–322.

Pfeifer, R. and Scheirer, C. (1999). *Understanding Intelligence*. MIT Press, Cambridge, MA.

Preston, J. (2002). Introduction. In Preston and Bishop (2002), pages 1–50.

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

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

Rapaport, W. J. (1986). Searle's experiments with thought. *Philosophy of Science*, **53**(2), 271–279.

Reiss, J. (2002). Causal inference in the abstract or seven myths about thought experiments. Technical Report CTR 03/02, Centre for Philosophy of Natural and Social Science, London School of Economics, London, UK.

Russow, L.-M. (1984). Unlocking the Chinese room. *Nature and System*, **6**, 221–227.

Saygin, A. P., Cicekli, I., and Akman, A. (2000). Turing test: 50 years later. *Minds and Machines*, **10**(4), 463–518.

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

Scheutz, M., editor (2002). *Computationalism: New Directions*. Bradford Books/MIT Press, Cambridge, MA.

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

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

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

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

Searle, J. R. (2002). Twenty one years in the Chinese room. In Preston and Bishop (2002), pages 51–59.

Seddon, G. (1972). Logical possibility. *Mind*, **81**(324), 481–494.

Siegelmann, H. T. (1999). *Neural Networks and Analog Computation: Beyond the Turing Limit*. Birkhäuser, Boston, MA.

Sloman, A. and Croucher, M. (1980). How to turn an information processor into an understander. *Behavioral and Brain Sciences*, **3**(3), 447–448. (Peer commentary on Searle, 1980).

Smith, B., editor (2003). *John Searle*. Cambridge University Press, Cambridge, UK.

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

Sorensen, R. A. (1998). Review of Sören Häggqvist's "Thought experiments in philosophy". *Theoria*, **64**(1), 108–118.

Souder, L. (2003). What are we to think about thought experiments? *Argumentation*, **17**(2), 203–217.

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

Wakefield, J. C. (2003). The Chinese room argument reconsidered: Essentialism, indeterminacy and strong AI. *Minds and Machines*, **13**(2), 285–319.

Weiss, T. (1990). Closing the Chinese room. *Ratio*, **3**(2), 165–181.

Wilensky, R. (1983). *Planning and Understanding: A Computational Approach to Human Reasoning*. Addison-Wesley, Reading, MA.

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

Wilks, Y. (1982). Searle's straw men. *Behavioral and Brain Sciences*, **5**(2), 343–344. (Continuing peer commentary on Searle, 1980).

Yablo, S. (1993). Is conceivability a guide to possibility? *Philosophy and Metaphysical Research*, **53**(1), 1–42.