Stanford Encyclopedia of Philosophy

The Chinese Room Argument

First published Fri Mar 19, 2004; substantive revision Wed Apr 9, 2014

The argument and thought-experiment now generally known as the Chinese Room Argument was first published in a paper in 1980 by American philosopher John Searle (1932-). It has become one of the best-known arguments in recent philosophy. Searle imagines himself alone in a room following a computer program for responding to Chinese characters slipped under the door. Searle understands nothing of Chinese, and yet, by following the program for manipulating symbols and numerals just as a computer does, he produces appropriate strings of Chinese characters that fool those outside into thinking there is a Chinese speaker in the room. The narrow conclusion of the argument is that programming a digital computer may make it appear to understand language but does not produce real understanding. Hence the "Turing Test" is inadequate. Searle argues that the thought experiment underscores the fact that computers merely use syntactic rules to manipulate symbol strings, but have no understanding of meaning or semantics. The broader conclusion of the argument is that the theory that human minds are computer-like computational or information processing systems is refuted. Instead minds must result from biological processes; computers can at best simulate these biological processes. Thus the argument has large implications for semantics, philosophy of language and mind, theories of consciousness, computer science and cognitive science generally. As a result, there have been many critical replies to the argument.

---

1. Overview

Work in Artificial Intelligence (AI) has produced computer programs that can beat the world chess champion and defeat the best human players on the television quiz show *Jeopardy*. AI has also produced programs with which one can converse in natural language, including Apple's *Siri*. Our experience shows that playing chess or *Jeopardy*, and carrying on a conversation, are activities that require understanding and intelligence. Does computer prowess at challenging games and conversation then show that computers can understand and be intelligent? Will further development result in digital computers that fully match or even exceed human intelligence? Alan Turing (1950), one of the pioneer theoreticians of computing, believed the answer to these questions was "yes". Turing proposed what is now known as "The Turing Test": if a computer can pass for human in online chat, we should grant that it is intelligent. By the late 1970s some AI researchers claimed that computers already understood at least some natural language. In 1980 U.C. Berkeley philosopher John Searle introduced a short and widely-discussed argument intended to show conclusively that it is impossible for digital computers to understand language or think.

Searle argues that a good way to test a theory of mind, say a theory that holds that understanding can be created by doing such and such, is to imagine what it would be like to do what the theory says would create understanding. Searle (1999) summarized the Chinese Room argument concisely:

> Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese.

Searle goes on to say, "The point of the argument is this: if the man in the room does not understand Chinese on the basis of implementing the appropriate program for understanding Chinese then neither does any other digital computer solely on that basis because no computer, qua computer, has anything the man does not have."

Thirty years later Searle 2010 describes the conclusion in terms of consciousness and intentionality:

> I demonstrated years ago with the so-called Chinese Room Argument that the implementation of the computer program is not by itself sufficient for consciousness or intentionality (Searle 1980). Computation is defined purely formally or syntactically, whereas minds have actual mental or semantic contents, and we cannot get from syntactical to the semantic just by having the syntactical operations and nothing else. To put this point slightly more technically, the notion "same implemented program" defines an equivalence class that is specified independently of any specific physical realization. But such a specification necessarily leaves out the biologically specific powers of the brain to cause cognitive processes. A system, me, for example, would not acquire an understanding of Chinese just by going through the steps of a computer program that simulated the

> behavior of a Chinese speaker (p.17).

Searle's shift from machine understanding to consciousness and intentionality is not directly supported by the original 1980 argument. However the re-description of the conclusion indicates the close connection between understanding and consciousness in Searle's accounts of meaning and intentionality. Those who don't accept Searle's linking account might hold that running a program can create understanding without necessarily creating consciousness, and a robot might have creature consciousness without necessarily understanding natural language.

Thus Searle develops the broader implications of his argument. It aims to refute the functionalist approach to understanding minds, the approach that holds that mental states are defined by their causal roles, not by the stuff (neurons, transistors) that plays those roles. The argument counts especially against that form of functionalism known as the Computational Theory of Mind that treats minds as information processing systems. As a result of its scope, as well as Searle's clear and forceful writing style, the Chinese Room argument has probably been the most widely discussed philosophical argument in cognitive science to appear since the Turing Test. By 1991 computer scientist Pat Hayes had defined Cognitive Science as the ongoing research project of refuting Searle's argument. Cognitive psychologist Steven Pinker (1997) pointed out that by the mid-1990s well over 100 articles had been published on Searle's thought experiment—and that discussion of it was so pervasive on the Internet that Pinker found it a compelling reason to remove his name from all Internet discussion lists.

This interest has not subsided, and the range of connections with the argument has broadened. A search on Google Scholar for “Searle Chinese Room” limited to the period from 2010 through early 2014 produced over 750 results, including papers making connections between the argument and topics ranging from embodied cognition to theater to talk psychotherapy to postmodern views of truth and “our post-human future” – as well as discussions of group or collective minds and discussions of the role of intuitions in philosophy. This wide-range of discussion and implications is a tribute to the argument's simple clarity and centrality.

2. Historical Background

2.1 Leibniz’ Mill

Searle's argument has three important antecedents. The first of these is an argument set out by the philosopher and mathematician Gottfried Leibniz (1646–1716). This argument, often known as “Leibniz’ Mill”, appears as section 17 of Leibniz’ *Monadology*. Like Searle's argument, Leibniz’ argument takes the form of a thought experiment. Leibniz asks us to imagine a physical system, a machine, that behaves in such a way that it supposedly thinks and has experiences (“perception”).

> 17. Moreover, it must be confessed that perception and that which depends upon it are inexplicable on mechanical grounds, that is to say, by means of figures and motions. And supposing there were a machine, so constructed as to think, feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. Thus it is in a simple substance, and not in a compound or in a machine, that perception must be sought for. [Robert Latta translation]

Notice that Leibniz's strategy here is to contrast the overt behavior of the machine, which might appear to be the

product of conscious thought, with the way the machine operates internally. He points out that these internal mechanical operations are just parts moving from point to point, hence there is nothing that is conscious or that can explain thinking, feeling or perceiving. For Leibniz physical states are not sufficient for, nor constitutive of, mental states.

2.2 Turing's Paper Machine

A second antecedent to the Chinese Room argument is the idea of a paper machine, a computer implemented by a human. This idea is found in the work of Alan Turing, for example in "Intelligent Machinery" (1948). Turing writes there that he wrote a program for a "paper machine" to play chess. A paper machine is a kind of program, a series of simple steps like a computer program, but written in natural language (e.g., English), and followed by a human. The human operator of the paper chess-playing machine need not (otherwise) know how to play chess. All the operator does is follow the instructions for generating moves on the chess board. In fact, the operator need not even know that he or she is involved in playing chess—the input and output strings, such as "N–QB7" need mean nothing to the operator of the paper machine.

Turing was optimistic that computers themselves would soon be able to exhibit apparently intelligent behavior, answering questions posed in English and carrying on conversations. Turing (1950) proposed what is now known as the Turing Test: if a computer could pass for human in on-line chat, it should be counted as intelligent. By the late 1970s, as computers became faster and less expensive, some in the burgeoning AI community claimed that their programs could understand English sentences, using a database of background information. The work of one of these, Yale researcher Roger Schank (Schank & Abelson 1977) came to the attention of John Searle (Searle's U.C. Berkeley colleague Hubert Dreyfus was an earlier critic of the claims made by AI researchers). Schank developed a technique called "conceptual representation" that used "scripts" to represent conceptual relations (a form of Conceptual Role Semantics). Searle's argument was originally presented as a response to the claim that AI programs such as Schank's literally understand the sentences that they respond to.

2.3 The Chinese Nation

A third more immediate antecedent to the Chinese Room argument emerged in early discussion of functionalist theories of minds and cognition. Functionalists hold that mental states are defined by the causal role they play in a system (just as a door stop is defined by what it does, not by what it is made out of). Critics of functionalism were quick to turn its proclaimed virtue of multiple realizability against it. In contrast with type-type identity theory, functionalism allowed beings with different physiology to have the same types of mental states as humans —pains, for example. But it was pointed out that if aliens could realize the functional properties that constituted mental states, then, presumably so could systems even less like human brains. The computational form of functionalism is particularly vulnerable to this maneuver, since a wide variety of systems with simple components are computationally equivalent (see e.g., Maudlin 1989 for a computer built from buckets of water). Critics asked if it was really plausible that these inorganic systems could have mental states or feel pain.

Daniel Dennett (1978) reports that in 1974 Lawrence Davis gave a colloquium at MIT in which he presented one such unorthodox implementation. Dennett summarizes Davis' thought experiment as follows:

> Let a functionalist theory of pain (whatever its details) be instantiated by a system the subassemblies of which are not such things as C-fibers and reticular systems but telephone lines and offices staffed by people. Perhaps it is a giant robot controlled by an army of human beings that inhabit it. When

> the theory's functionally characterized conditions for pain are now met we must say, if the theory is true, that the robot is in pain. That is, real pain, as real as our own, would exist in virtue of the perhaps disinterested and businesslike activities of these bureaucratic teams, executing their proper functions.

In "Troubles with Functionalism", also published in 1978, Ned Block envisions the entire population of China implementing the functions of neurons in the brain. This scenario has subsequently been called "The Chinese Nation" or "The Chinese Gym". We can suppose that every Chinese citizen would be given a call-list of phone numbers, and at a preset time on implementation day, designated "input" citizens would initiate the process by calling those on their call-list. When any citizen's phone rang, he or she would then phone those on his or her list, who would in turn contact yet others. No phone message need be exchanged; all that is required is the pattern of calling. The call-lists would be constructed in such a way that the patterns of calls implemented the same patterns of activation that occur between neurons in someone's brain when that person is in a mental state—pain, for example. The phone calls play the same functional role as neurons causing one another to fire. Block was primarily interested in qualia, and in particular, whether it is plausible to hold that the population of China might collectively be in pain, while no individual member of the population experienced any pain, but the thought experiment applies to any mental states and operations, including understanding language.

Thus Block's precursor thought experiment, as with those of Davis and Dennett, is a system of many humans rather than one. The focus is on consciousness, but to the extent that Searle's argument also involves consciousness, the thought experiment is closely related to Searle's.

3. The Chinese Room Argument

In 1980 John Searle published "Minds, Brains and Programs" in the journal *The Behavioral and Brain Sciences*. In this article, Searle sets out the argument, and then replies to the half-dozen main objections that had been raised during his earlier presentations at various university campuses (see next section). In addition, Searle's article in *BBS* was published along with comments and criticisms by 27 cognitive science researchers. These 27 comments were followed by Searle's replies to his critics.

In the decades following its publication, the Chinese Room argument was the subject of very many discussions. By 1984, Searle presented the Chinese Room argument in a book, *Minds, Brains and Science*. In January 1990, the popular periodical *Scientific American* took the debate to a general scientific audience. Searle included the Chinese Room Argument in his contribution, "Is the Brain's Mind a Computer Program?", and Searle's piece was followed by a responding article, "Could a Machine Think?", written by philosophers Paul and Patricia Churchland. Soon thereafter Searle had a published exchange about the Chinese Room with another leading philosopher, Jerry Fodor (in Rosenthal (ed.) 1991).

The heart of the argument is an imagined human simulation of a computer, similar to Turing's Paper Machine. The human in the Chinese Room follows English instructions for manipulating Chinese symbols, where a computer "follows" a program written in a computing language. The human produces the appearance of understanding Chinese by following the symbol manipulating instructions, but does not thereby come to understand Chinese. Since a computer just does what the human does—manipulate symbols on the basis of their syntax alone—no computer, merely by following a program, comes to genuinely understand Chinese.

This narrow argument, based closely on the Chinese Room scenario, is specifically directed at a position Searle

calls “Strong AI”. Strong AI is the view that suitably programmed computers (or the programs themselves) can understand natural language and actually have other mental capabilities similar to the humans whose behavior they mimic. According to Strong AI, these computers really play chess intelligently, make clever moves, or understand language. By contrast, “weak AI” is the much more modest claim that computers are merely useful in psychology, linguistics, and other areas, in part because they can simulate mental abilities. But weak AI makes no claim that computers actually understand or are intelligent. The Chinese Room argument is not directed at weak AI, nor does it purport to show that no machine can think—Searle says that brains are machines, and brains think. The argument is directed at the view that formal computations on symbols can produce thought.

We might summarize the narrow argument as a *reductio ad absurdum* against Strong AI as follows. Let L be a natural language, and let us say that a “program for L” is a program for conversing fluently in L. A computing system is any system, human or otherwise, that can run a program.

1. If Strong AI is true, then there is a program for Chinese such that if any computing system runs that program, that system thereby comes to understand Chinese.
2. I could run a program for Chinese without thereby coming to understand Chinese.
3. Therefore Strong AI is false.

The second premise is supported by the Chinese Room thought experiment. The conclusion of this narrow argument is that running a program cannot endow the system with language understanding. Searle's wider argument includes the claim that the thought experiment shows more generally that one cannot get semantics (meaning) from syntax (formal symbol manipulation). That and related issues are discussed in the section The Larger Philosophical Issues.

4. Replies to the Chinese Room Argument

Criticisms of the narrow Chinese Room argument against Strong AI have often followed three main lines, which can be distinguished by how much they concede:

(1) Some critics concede that the man in the room doesn't understand Chinese, but hold that nevertheless running the program may create something that understands Chinese. These critics object to the inference from the claim that the *man* in the room does not understand Chinese to the conclusion that *no understanding* has been created. There might be understanding by a larger, or different, entity. This is the strategy of The Systems Reply and the Virtual Mind Reply. These replies hold that the output of the room reflects understanding of Chinese, but the understanding is not that of the room's operator. Thus Searle's claim that he doesn't understand Chinese while running the room is conceded, but his claim that there is no understanding, and that computationalism is false, is denied.

(2) Other critics concede Searle's claim that just running a natural language processing program as described in the CR scenario does not create any understanding, whether by a human or a computer system. But these critics hold that a *variation* on the computer system could understand. The variant might be a computer embedded in a robotic body, having interaction with the physical world via sensors and motors (“The Robot Reply”), or it might be a system that simulated the detailed operation of an entire brain, neuron by neuron (“the Brain Simulator Reply”).

(3) Finally, some critics do not concede even the narrow point against AI. These critics hold that the man in the

original Chinese Room scenario might understand Chinese, despite Searle's denials, or that the scenario is impossible. For example, critics have argued that our intuitions in such cases are unreliable. Other critics have held that it all depends on what one means by "understand"—points discussed in the section on The Intuition Reply. Others (e.g. Sprevak 2007) object to the assumption that any system (e.g. Searle in the room) can run any computer program. And finally some have argued that if it is not reasonable to attribute understanding on the basis of the behavior exhibited by the Chinese Room, then it would not be reasonable to attribute understanding to humans on the basis of similar behavioral evidence (Searle calls this last the "Other Minds Reply"). The objection is that we should be willing to attribute understanding in the Chinese Room on the basis of the overt behavior, just as we do with other humans (and some animals), and as we would do with extra-terrestrial Aliens (or burning bushes or angels) that spoke our language.

In addition to these responses specifically to the Chinese Room scenario and the narrow argument to be discussed here, some critics also independently argue against Searle's larger claim, and hold that one can get semantics (that is, meaning) from syntactic symbol manipulation, including the sort that takes place inside a digital computer, a question discussed in the section below on Syntax and Semantics.

4.1 The Systems Reply

In the original BBS article, Searle identified and discussed several responses to the argument that he had come across in giving the argument in talks at various places. As a result, these early responses have received the most attention in subsequent discussion. What Searle 1980 calls "perhaps the most common reply" is the Systems Reply.

The Systems Reply, which Searle says was originally associated with Yale, concedes that the man in the room does not understand Chinese. But, the reply continues, the man is but a part, a central processing unit (CPU), in a larger system. The larger system includes the huge database, the memory (scratchpads) containing intermediate states, and the instructions—the complete system that is required for answering the Chinese questions. So the Sytems Reply is that while the man running the program does not understand Chinese, the system as a whole does.

Ned Block was one of the first to press the Systems Reply, along with many others including Jack Copeland, Daniel Dennett, Jerry Fodor, John Haugeland, Ray Kurzweil and Georges Rey. Rey (1986) says the person in the room is just the CPU of the system. Kurzweil (2002) says that the human being is just an implementer and of no significance (presumably meaning that the properties of the implementer are not necessarily those of the system). Kurzweil hews to the spirit of the Turing Test and holds that if the system displays the apparent capacity to understand Chinese "it would have to, indeed, understand Chinese"—Searle is contradicting himself in saying in effect, "the machine speaks Chinese but doesn't understand Chinese".

Margaret Boden (1988) raises levels considerations. "Computational psychology does not credit the brain with seeing bean-sprouts or understanding English: intentional states such as these are properties of people, not of brains" (244). "In short, Searle's description of the robot's pseudo-brain (that is, of Searle-in-the-robot) as understanding English involves a category-mistake comparable to treating the brain as the bearer, as opposed to the causal basis, of intelligence". Boden (1988) points out that the room operator is a conscious agent, while the CPU in a computer is not—the Chinese Room scenario asks us to take the perspective of the implementer, and not surprisingly fails to see the larger picture.

Searle's response to the Systems Reply is simple: in principle, the man can internalize the entire system, memorizing all the instructions and the database, and doing all the calculations in his head. He could then leave the room and wander outdoors, perhaps even conversing in Chinese. But he still would have no way to attach "any meaning to the formal symbols". The man would now *be* the entire system, yet he still would not understand Chinese. For example, he would not know the meaning of the Chinese word for hamburger. He still cannot get semantics from syntax. (See below the section on Syntax and Semantics).

In his 2002 paper "The Chinese Room from a Logical Point of View", Jack Copeland considers Searle's response to the Systems Reply and argues that a homunculus inside Searle's head might understand even though the room operator himself does not, just as modules in minds solve tensor equations that enable us to catch cricket balls. Copeland then turns to consider the Chinese Gym, and again appears to endorse the Systems Reply: "…the individual players [do not] understand Chinese. But there is no entailment from this to the claim that the simulation as a whole does not come to understand Chinese. The fallacy involved in moving from part to whole is even more glaring here than in the original version of the Chinese Room Argument". Copeland denies that connectionism implies that a room of people can simulate the brain.

John Haugeland writes (2002) that Searle's response to the Systems Reply is flawed: "…what he now asks is what it would be like if he, in his own mind, were consciously to implement the underlying formal structures and operations that the theory says are sufficient to implement another mind". According to Haugeland, his failure to understand Chinese is irrelevant: he is just the implementer. The larger system implemented would understand—there is a level-of-description fallacy.

Shaffer 2009 examines modal aspects of the logic of the CRA and argues that familiar versions of the System Reply are question-begging. But, Shaffer claims, a modalized version of the System Reply succeeds because there are possible worlds in which understanding is an emergent property of complex syntax manipulation. Nute 2011 is a reply to Shaffer.

Stevan Harnad has defended Searle's argument against Systems Reply critics in two papers. In his 1989 paper, Harnad writes "Searle formulates the problem as follows: Is the mind a computer program? Or, more specifically, if a computer program simulates or imitates activities of ours that seem to require understanding (such as communicating in language), can the program itself be said to understand in so doing?" (Note the specific claim: the issue is taken to be whether the program itself understands.) Harnad concludes: "On the face of it, [the CR argument] looks valid. It certainly works against the most common rejoinder, the 'Systems Reply'…." Harnad appears to follow Searle in linking understanding and states of consciousness: Harnad 2012 (Other Internet Resources) argues that Searle shows that the core problem of conscious "feeling" requires sensory connections to the real world.

4.1.1 The Virtual Mind Reply

The Virtual Mind reply concedes, as does the System Reply, that the operator of the Chinese Room does not understand Chinese merely by running the paper machine. However the Virtual Mind reply holds that what is important is whether understanding is created, not whether the Room operator is the agent that understands. Unlike the Systems Reply, the Virtual Mind reply (VMR) holds that a running system may create new, virtual, entities that are distinct from both the system as a whole, as well as from the sub-systems such as the CPU or operator. In particular, a running system might create a distinct agent that understands Chinese. This virtual agent would be distinct from both the room operator and the entire system. The psychological traits, including linguistic

abilities, of any mind created by artificial intelligence will depend entirely upon the program and the Chinese database, and will not be identical with the psychological traits and abilities of a CPU or the operator of a paper machine, such as Searle in the Chinese Room scenario. According to the VMR the mistake in the Chinese Room Argument is to make the claim of strong AI to be "the computer understands Chinese" or "the System understands Chinese". The claim at issue for AI should simply be whether "the running computer creates understanding of Chinese".

A familiar model of virtual agents are characters in computer or video games, and personal digital assistants, such as Apple's Siri and Microsoft's Cortana. These characters have various abilities and personalities, and the characters are not identical with the system hardware or program that creates them. A single running system might control two distinct agents, or physical robots, simultaneously, one of which converses only in Chinese and one of which can converse only in English, and which otherwise manifest very different personalities, memories, and cognitive abilities. Thus the VM reply asks us to distinguish between minds and their realizing systems.

Minsky (1980) and Sloman and Croucher (1980) suggested a Virtual Mind reply when the Chinese Room argument first appeared. In his widely-read 1989 paper "Computation and Consciousness", Tim Maudlin considers minimal physical systems that might implement a computational system running a program. His discussion revolves around his imaginary Olympia machine, a system of buckets that transfers water, implementing a Turing machine. Maudlin's main target is the computationalists' claim that such a machine could have phenomenal consciousness. However in the course of his discussion, Maudlin considers the Chinese Room argument. Maudlin (citing Minsky, and Sloman and Croucher) points out a Virtual Mind reply that the agent that understands could be distinct from the physical system (414). Thus "Searle has done nothing to discount the possibility of simultaneously existing disjoint mentalities" (414–5).

Perlis (1992), Chalmers (1996) and Block (2002) have apparently endorsed versions of a Virtual Mind reply as well, as has Richard Hanley in *The Metaphysics of Star Trek* (1997). Penrose (2002) is a critic of this strategy, and Stevan Harnad scornfully dismisses such heroic resorts to metaphysics. Harnad defended Searle's position in a "Virtual Symposium on Virtual Minds" (1992) against Patrick Hayes and Don Perlis. Perlis pressed a virtual minds argument derived, he says, from Maudlin. Chalmers (1996) notes that the room operator is just a causal facilitator, a "demon", so that his states of consciousness are irrelevant to the properties of the system as a whole. Like Maudlin, Chalmers raises issues of personal identity—we might regard the Chinese Room as "two mental systems realized within the same physical space. The organization that gives rise to the Chinese experiences is quite distinct from the organization that gives rise to the demon's [= room operator's] experiences"(326).

Cole (1991, 1994) develops the reply and argues as follows: Searle's argument requires that the agent of understanding be the computer itself or, in the Chinese Room parallel, the person in the room. However Searle's failure to understand Chinese in the room does not show that there is no understanding being created. If we flesh out the conversation in the original CR scenario to include questions in Chinese such as "How tall are you?", "Where do you live?", "What did you have for breakfast?", "What is your attitude toward Mao?", and so forth, it immediately becomes clear that the answers in Chinese are not *Searle's* answers. Searle is not the author of the answers, and his beliefs and desires, memories and personality traits are not reflected in the answers and, apart from his industriousness!, are causally inert in producing the answers to the Chinese questions. Hence if there is understanding of Chinese created by running the program, the mind understanding the Chinese would not be the computer, nor, in the Chinese Room, would the person understanding Chinese be the room operator. The person understanding the Chinese would be a distinct person from the room operator, with beliefs and desires bestowed

by the program and its database. Hence Searle's failure to understand Chinese while operating the room does not show that understanding is not being created.

Cole (1991) offers an additional argument that the mind doing the understanding is neither the mind of the room operator nor the system consisting of the operator and the program: running a suitably structured computer program might produce answers submitted in Chinese and also answers to questions submitted in Korean. Yet the Chinese answers might apparently display completely different knowledge and memories, beliefs and desires than the answers to the Korean questions—along with a denial that the Chinese answerer knows any Korean, and vice versa. Thus the behavioral evidence would be that there were two non-identical minds (one understanding Chinese only, and one understanding Korean only). Since these might have mutually exclusive psychological properties, they cannot be identical, and ipso facto, cannot be identical with the mind of the implementer in the room. Analogously, a video game might include a character with one set of cognitive abilities (smart, understands Chinese) as well as another character with an incompatible set (stupid, English monoglot). These inconsistent cognitive traits cannot be traits of the XBOX system that realizes them. The implication seems to be that minds generally are more abstract than the systems that realize them (see Mind and Body in the Larger Philosophical Issues section).

In short, the Virtual Mind argument is that since the evidence that Searle provides that there is no understanding of Chinese was that *he* wouldn't understand Chinese in the room, the Chinese Room Argument cannot refute a differently formulated equally strong AI claim, asserting the possibility of creating understanding using a programmed digital computer. Maudlin (1989) says that Searle has not adequately responded to this criticism.

Others however have replied to the VMR, including Stevan Harnad and mathematical physicist Roger Penrose. Penrose is generally sympathetic to the points Searle raises with the Chinese Room argument, and has argued against the Virtual Mind reply. Penrose does not believe that computational processes can account for consciousness, both on Chinese Room grounds, as well as because of limitations on formal systems revealed by Kurt Gödel's incompleteness proof. (Penrose has two books on mind and consciousness; Chalmers and others have responded to Penrose's appeals to Gödel.) In his 2002 article “Consciousness, Computation, and the Chinese Room” that specifically addresses the Chinese Room argument, Penrose argues that the Chinese Gym variation—with a room expanded to the size of India, with Indians doing the processing—shows it is very implausible to hold there is “some kind of disembodied ‘understanding’ associated with the person's carrying out of that algorithm, and whose presence does not impinge in any way upon his own consciousness” (230–1). Penrose concludes the Chinese Room argument refutes Strong AI. Christian Kaernbach (2005) reports that he subjected the virtual mind theory to an empirical test, with negative results.

4.2 The Robot Reply

The Robot Reply concedes Searle is right about the Chinese Room scenario: it shows that a computer trapped in a computer room cannot understand language, or know what words mean. The Robot reply is responsive to the problem of knowing the meaning of the Chinese word for hamburger—Searle's example of something the room operator would not know. It seems reasonable to hold that we know what a hamburger is because we have seen one, and perhaps even made one, or tasted one, or at least heard people talk about hamburgers and understood what they are by relating them to things we do know by seeing, making, and tasting. Given this is how one might come to know what hamburgers are, the Robot Reply suggests that we put a digital computer in a robot body, with sensors, such as video cameras and microphones, and add effectors, such as wheels to move around with, and arms with which to manipulate things in the world. Such a robot—a computer with a body—could do what

a child does, learn by seeing and doing. The Robot Reply holds that such a digital computer in a robot body, freed from the room, could attach meanings to symbols and actually understand natural language. Margaret Boden, Tim Crane, Daniel Dennett, Jerry Fodor, Stevan Harnad, Hans Moravec and Georges Rey are among those who have endorsed versions of this reply at one time or another. The Robot Reply in effect appeals to "wide content" or "externalist semantics". This can agree with Searle that syntax and internal connections are insufficient for semantics, while holding that suitable causal connections with the world can provide content to the internal symbols.

Searle does not think this reply to the Chinese Room argument is any stronger than the Systems Reply. All the sensors do is provide additional input to the computer—and it will be just syntactic input. We can see this by making a parallel change to the Chinese Room scenario. Suppose the man in the Chinese Room receives, in addition to the Chinese characters slipped under the door, a stream of binary digits that appear, say, on a ticker tape in a corner of the room. The instruction books are augmented to use the numerals from the tape as input, along with the Chinese characters. Unbeknownst to the man in the room, the symbols on the tape are the digitized output of a video camera (and possibly other sensors). Searle argues that additional syntactic inputs will do nothing to allow the man to associate meanings with the Chinese characters. It is just more work for the man in the room.

Jerry Fodor, Hilary Putnam, and David Lewis, were principle architects of the computational theory of mind that Searle's wider argument attacks. In his original 1980 reply to Searle, Fodor allows Searle is certainly right that "instantiating the same program as the brain does is not, in and of itself, sufficient for having those propositional attitudes characteristic of the organism that has the brain." But Fodor holds that Searle is wrong about the robot reply. A computer might have propositional attitudes if it has the right causal connections to the world—but those are not ones mediated by a man sitting in the head of the robot. We don't know what the right causal connections are. Searle commits the fallacy of inferring from "the little man is not the right causal connection" to conclude that no causal linkage would succeed. There is considerable empirical evidence that mental processes involve "manipulation of symbols"; Searle gives us no alternative explanation (this is sometimes called Fodor's "Only Game in Town" argument for computational approaches). Since this time, Fodor has written extensively on what the connections must be between a brain state and the world for the state to have intentional (representational) properties, while most recently emphasizing that computationalism has limits because the computations are intrinsically local and so cannot account for abductive reasoning.

In a later piece, "Yin and Yang in the Chinese Room" (in Rosenthal 1991 pp.524–525), Fodor substantially revises his 1980 view. He distances himself from his earlier version of the robot reply, and holds instead that "instantiation" should be defined in such a way that the symbol must be the proximate cause of the effect—no intervening guys in a room. So Searle in the room is not an instantiation of a Turing Machine, and "Searle's setup does not instantiate the machine that the brain instantiates." He concludes: "…Searle's setup is irrelevant to the claim that strong equivalence to a Chinese speaker's brain is ipso facto sufficient for speaking Chinese." Searle says of Fodor's move, "Of all the zillions of criticisms of the Chinese Room argument, Fodor's is perhaps the most desperate. He claims that precisely because the man in the Chinese room sets out to implement the steps in the computer program, he is not implementing the steps in the computer program. He offers no argument for this extraordinary claim." (in Rosenthal 1991, p. 525)

In a 1986 paper, Georges Rey advocated a combination of the system and robot reply, after noting that the original Turing Test is insufficient as a test of intelligence and understanding, and that the isolated system Searle describes in the room is certainly not functionally equivalent to a real Chinese speaker sensing and acting in the

world. In a 2002 second look, "Searle's Misunderstandings of Functionalism and Strong AI", Rey again defends functionalism against Searle, and in the particular form Rey calls the "computational-representational theory of thought—CRTT". CRTT is not committed to attributing thought to just any system that passes the Turing Test (like the Chinese Room). Nor is it committed to a conversation manual model of understanding natural language. Rather, CRTT is concerned with intentionality, natural and artificial (the representations in the system are semantically evaluable—they are true or false, hence have aboutness). Searle saddles functionalism with the "blackbox" character of behaviorism, but functionalism cares how things are done. Rey sketches "a modest mind"—a CRTT system that has perception, can make deductive and inductive inferences, makes decisions on basis of goals and representations of how the world is, and can process natural language by converting to and from its native representations. To explain the behavior of such a system we would need to use the same attributions needed to explain the behavior of a normal Chinese speaker.

Tim Crane discusses the Chinese Room argument in his 1991 book, *The Mechanical Mind*. He cites the Churchlands' luminous room analogy, but then goes on to argue that in the course of operating the room, Searle would learn the meaning of the Chinese: "…if Searle had not just memorized the rules and the data, but also started acting in the world of Chinese people, then it is plausible that he would before too long come to realize what these symbols mean."(127). (Rapaport 2006 presses an analogy between Helen Keller and the Chinese Room.) Crane appears to end with a version of the Robot Reply: "Searle's argument itself begs the question by (in effect) just denying the central thesis of AI—that thinking is formal symbol manipulation. But Searle's assumption, none the less, seems to me correct … the proper response to Searle's argument is: sure, Searle-in-the-room, or the room alone, cannot understand Chinese. But if you let the outside world have some impact on the room, meaning or 'semantics' might begin to get a foothold. But of course, this concedes that thinking cannot be simply symbol manipulation." (129)

Margaret Boden 1988 also argues that Searle mistakenly supposes programs are pure syntax. But programs bring about the activity of certain machines: "The inherent procedural consequences of any computer program give it a toehold in semantics, where the semantics in question is not denotational, but causal." (250) Thus a robot might have causal powers that enable it to refer to a hamburger.

Stevan Harnad also finds important our sensory and motor capabilities: "Who is to say that the Turing Test, whether conducted in Chinese or in any other language, could be successfully passed without operations that draw on our sensory, motor, and other higher cognitive capacities as well? Where does the capacity to comprehend Chinese begin and the rest of our mental competence leave off?" Harnad believes that symbolic functions must be grounded in "robotic" functions that connect a system with the world. And he thinks this counts against symbolic accounts of mentality, such as Jerry Fodor's, and, one suspects, the approach of Roger Schank that was Searle's original target. Harnad 2012 (Other Internet Resources) argues that the CRA shows that even with a robot with symbols grounded in the external world, there is still something missing: feeling, such as the feeling of understanding.

4.3 The Brain Simulator Reply

Consider a computer that operates in quite a different manner than the usual AI program with scripts and operations on sentence-like strings of symbols. The Brain Simulator reply asks us to suppose instead the program simulates the actual sequence of nerve firings that occur in the brain of a native Chinese language speaker when that person understands Chinese—every nerve, every firing. Since the computer then works the very same way as the brain of a native Chinese speaker, processing information in just the same way, it will

understand Chinese. Paul and Patricia Churchland have set out a reply along these lines, discussed below.

In response to this, Searle argues that it makes no difference. He suggests a variation on the brain simulator scenario: suppose that in the room the man has a huge set of valves and water pipes, in the same arrangement as the neurons in a native Chinese speaker's brain. The program now tells the man which valves to open in response to input. Searle claims that it is obvious that there would be no understanding of Chinese. (Note however that the basis for this claim is no longer simply that Searle himself wouldn't understand Chinese – it seems clear that now he is just facilitating the causal operation of the system and so we rely on our Leibnizian intuition that water-works don't understand (see also Maudlin 1989).) Searle concludes that a simulation of brain activity is not the real thing.

However, following Pylyshyn 1980, Cole and Foelber 1984, Chalmers 1996, we might wonder about hybrid systems. Pylyshyn writes:

> If more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output function each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that you would eventually stop meaning anything by it. What we outside observers might take to be words would become for you just certain noises that circuits caused you to make.

These cyborgization thought experiments can be linked to the Chinese Room. Suppose Otto has a neural disease that causes one of the neurons in my brain to fail, but surgeons install a tiny remotely controlled artificial neuron, a synron, along side his disabled neuron. The control of Otto's neuron is by John Searle in the Chinese Room, unbeknownst to both of them. Tiny wires connect the artificial neuron to the synapses on the cell-body of his disabled neuron. When his artificial neuron is stimulated by neurons that synapse on his disabled neuron, a light goes on in the Chinese Room. Searle then manipulates some valves and switches in accord with a program. That, via the radio link, causes Otto's artificial neuron to release neuro-transmitters from its tiny artificial vesicles. If Searle's programmed activity causes Otto's artificial neuron to behave just as his disabled natural neuron once did, the behavior of the rest of my nervous system will be unchanged. Alas, Otto's disease progresses; more neurons are replaced by synrons controlled by Searle. Ex hypothesis the rest of the world will not notice the difference; will Otto?

Under the rubric "The Combination Reply", Searle also considers a system with the features of all three of the preceding: a robot with a digital brain simulating computer in its cranium, such that the system as a whole behaves indistinguishably from a human. Since the normal input to the brain is from sense organs, it is natural to suppose that most advocates of the Brain Simulator Reply have in mind such a combination of brain simulation, Robot, and Systems Reply. Some (e.g. Rey 1986) argue it is reasonable to attribute intentionality to such a system as a whole. Searle agrees that it would be reasonable to attribute understanding to such an android system—but only as long as you don't know how it works. As soon as you know the truth—it is a computer, uncomprehendingly manipulating symbols on the basis of syntax, not meaning—you would cease to attribute intentionality to it.

(One assumes this would be true even if it were one's spouse, with whom one had built a life-long relationship, that was revealed to hide a silicon secret. Science fiction stories, including episodes of Rod Serling's television series *The Twilight Zone*, have been based on such possibilities; Steven Pinker (1997) mentions one episode in which the android's secret was known from the start, but the protagonist developed a romantic relationship with the android.)

On its tenth anniversary the Chinese Room argument was featured in the general science periodical *Scientific American*. Leading the opposition to Searle's lead article in that issue were philosophers Paul and Patricia Churchland. The Churchlands agree with Searle that the Chinese Room does not understand Chinese, but hold that the argument itself exploits our ignorance of cognitive and semantic phenomena. They raise a parallel case of "The Luminous Room" where someone waves a magnet and argues that the absence of resulting visible light shows that Maxwell's electromagnetic theory is false. The Churchlands advocate a view of the brain as a connectionist system, a vector transformer, not a system manipulating symbols according to structure-sensitive rules. The system in the Chinese Room uses the wrong computational strategies. Thus they agree with Searle against traditional AI, but they presumably would endorse what Searle calls "the Brain Simulator Reply", arguing that, as with the Luminous Room, our intuitions fail us when considering such a complex system, and it is a fallacy to move from part to whole: "… no neuron in my brain understands English, although my whole brain does."

In his 1991 book, *Microcognition*. Andy Clark holds that Searle is right that a computer running Schank's program does not know anything about restaurants, "at least if by 'know' we mean anything like 'understand'". But Searle thinks that this would apply to any computational model, while Clark, like the Churchlands, holds that Searle is wrong about connectionist models. Clark's interest is thus in the brain-simulator reply. The brain thinks in virtue of its physical properties. What physical properties of the brain are important? Clark answers that what is important about brains are "variable and flexible substructures" which conventional AI systems lack. But that doesn't mean computationalism or functionalism is false. It depends on what level you take the functional units to be. Clark defends "microfunctionalism"—one should look to a fine-grained functional description, e.g. neural net level. Clark cites William Lycan approvingly contra Block's absent qualia objection—yes, there can be absent qualia, if the functional units are made large. But that does not constitute a refutation of functionalism generally. So Clark's views are not unlike the Churchlands', conceding that Searle is right about Schank and symbolic-level processing systems, but holding that he is mistaken about connectionist systems.

Similarly Ray Kurzweil (2002) argues that Searle's argument could be turned around to show that human brains cannot understand—the brain succeeds by manipulating neurotransmitter concentrations and other mechanisms that are in themselves meaningless. In criticism of Searle's response to the Brain Simulator Reply, Kurzweil says: "So if we scale up Searle's Chinese Room to be the rather massive 'room' it needs to be, who's to say that the entire system of a hundred trillion people simulating a Chinese Brain that knows Chinese isn't conscious? Certainly, it would be correct to say that such a system knows Chinese. And we can't say that it is not conscious anymore than we can say that about any other process. We can't know the subjective experience of another entity…."

4.4 The Other Minds Reply

Related to the preceding is The Other Minds Reply: "How do you know that other people understand Chinese or anything else? Only by their behavior. Now the computer can pass the behavioral tests as well as they can (in principle), so if you are going to attribute cognition to other people you must in principle also attribute it to computers. "

Searle's (1980) reply to this is very short:

> The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that I am attributing to them when I attribute cognitive states to them. The thrust of the argument is that it couldn't be just computational processes and their output because the

> computational processes and their output can exist without the cognitive state. It is no answer to this argument to feign anesthesia. In 'cognitive sciences' one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects.

Critics hold that if the evidence we have that humans understand is the same as the evidence we might have that a visiting extra-terrestrial alien understands, which is the same as the evidence that a robot understands, the presuppositions we may make in the case of our own species are not relevant, for presuppositions are sometimes false. For similar reasons, Turing, in proposing the Turing Test, is specifically worried about our presuppositions and chauvinism. If the reasons for the presuppositions regarding humans are pragmatic, in that they enable us to predict the behavior of humans and to interact effectively with them, perhaps the presupposition could apply equally to computers (similar considerations are pressed by Dennett, in his discussions of what he calls the Intentional Stance).

Searle raises the question of just what we are attributing in attributing understanding to other minds, saying that it is more than complex behavioral dispositions. For Searle the additional seems to be certain states of consciousness, as is seen in his 2010 summary of the CRA conclusions. Terry Horgan (2013) endorses this claim: "the real moral of Searle's Chinese room thought experiment is that genuine original intentionality requires the presence of internal states with intrinsic phenomenal character that is inherently intentional…" But this tying of understanding to phenomenal consciousness raises a host of issues.

We attribute limited understanding of language to toddlers, dogs, and other animals, but it is not clear that we are ipso facto attributing unseen states of subjective consciousness – what do we know of the hidden states of exotic creatures? Ludwig Wittgenstein (the Private Language Arugment) and his followers pressed similar points. Altered qualia possibilities, analogous to the inverted spectrum, arise: suppose I ask "what's the sum of 5 and 7" and you respond "the sum of 5 and 7 is 12", but as you heard my question you had the conscious experience of hearing and understanding "what is the sum of 10 and 14", though you were in the computational states appropriate for producing the correct sum and so said "12". Are there certain conscious states that are "correct" for certain functional states? The underlying problem of epiphenomenality is one familiar from inverted spectrum problems – it is difficult to see what subjective consciousness adds if it is not itself functionally important.

In the 30 years since the CRA there has been philosophical interest in zombies – creatures that look like and behave just as normal humans, including linguistic behavior, yet have no subjective consciousness. A difficulty for claiming that subjective states of consciousness are crucial for understanding meaning will arise in these cases of absent qualia: we can't tell the difference between zombies and non-zombies, and so on Searle's account we can't tell the difference between those that really understand English and those that don't. But then there appears to be a distinction without a difference. In any case, Searle's short reply to the Other Minds Reply may be too short.

Descartes argued famously that speech was sufficient for attributing minds and consciousness to others, and argued infamously that it was necessary. Turing was in effect endorsing Descartes' sufficiency condition, at least for intelligence, while substituting written for oral linguistic behavior. Since most of us use dialog as a sufficient condition for attributing understanding, Searle's argument, which holds that speech is a sufficient condition for humans (in all states of sleep-walking, stroke?) but not for anything that doesn't share our biology, an account would appear to be required of what additionally is being attributed, and what can justify the additional attribution. Further, if being con-specific is key on Searle's account, a natural question arises as to what

circumstances would justify us in attributing understanding (or consciousness) to extra-terrestrial aliens who do not share our biology? Offending ET's by withholding attributions of understanding until after a post-mortem may be risky.

Hans Moravec, director of the Robotics laboratory at Carnegie Mellon University, and author of *Robot: Mere Machine to Transcendent Mind*, argues that Searle's position merely reflects intuitions from traditional philosophy of mind that are out of step with the new cognitive science. Moravec endorses a version of the Other Minds reply. It makes sense to attribute intentionality to machines for the same reasons it makes sense to attribute them to humans; his "interpretative position" is similar to the views of Daniel Dennett. Moravec goes on to note that one of the things we attribute to others is the ability to make attributions of intentionality, and then we make such attributions to ourselves. It is such self-representation that is at the heart of consciousness. These capacities appear to be implementation independent, and hence possible for aliens and suitably programmed computers.

Presumably the reason that Searle thinks we can disregard the evidence in the case of robots and computers is that we know that their processing is syntactic, and this fact trumps all other considerations. Indeed, Searle believes this is the larger point that the Chinese Room merely illustrates. This larger point is addressed in the Syntax and Semantics section below.

4.5 The Intuition Reply

Many responses to the Chinese Room argument have noted that, as with Leibniz' Mill, the argument appears to be based on intuition: the intuition that a computer (or the man in the room) cannot think or have understanding. For example, Ned Block (1980) in his original BBS commentary says "Searle's argument depends for its force on intuitions that certain entities do not think." But, Block argues, (1) intuitions sometimes can and should be trumped and (2) perhaps we need to bring our concept of understanding in line with a reality in which certain computer robots belong to the same natural kind as humans. Similarly Margaret Boden (1988) points out that we can't trust our untutored intuitions about how mind depends on matter; developments in science may change our intuitions. Indeed, elimination of bias in our intuitions was what motivated Turing (1950) to propose the Turing Test, a test that was blind to the physical character of the system replying to questions. Some of Searle's critics in effect argue that he has merely pushed the reliance on intuition back, into the room.

Critics argue that our intuitions regarding both intelligence and understanding may be unreliable, and perhaps incompatible even with current science. With regard to understanding, Steven Pinker, in *How the Mind Works* (1997), holds that "… Searle is merely exploring facts about the English word *understand*…. People are reluctant to use the word unless certain stereotypical conditions apply…" But, Pinker claims, nothing scientifically speaking is at stake. Pinker objects to Searle's appeal to the "causal powers of the brain" by noting that the apparent locus of the causal powers is the "patterns of interconnectivity that carry out the right information processing". Pinker ends his discussion by citing a science fiction story in which Aliens, anatomically quite unlike humans, cannot believe that humans think when they discover that our heads are filled with meat. The Aliens' intuitions are unreliable—presumably ours may be so as well.

Other critics are also concerned with how our understanding of understanding bears on the Chinese Room argument. In their paper "A Chinese Room that Understands" AI researchers Simon and Eisenstadt (2002) argue that whereas Searle refutes "logical strong AI", the thesis that a program that passes the Turing Test will *necessarily* understand, Searle's argument does not impugn "Empirical Strong AI"—the thesis that it is possible

to program a computer that convincingly satisfies ordinary criteria of understanding. They hold however that it is impossible to settle these questions "without employing a definition of the term 'understand' that can provide a test for judging whether the hypothesis is true or false". They cite W.V.O. Quine's *Word and Object* as showing that there is always empirical uncertainty in attributing understanding to humans. The Chinese Room is a Clever Hans trick (Clever Hans was a horse who appeared to clomp out the answers to simple arithmetic questions, but it was discovered that Hans could detect unconscious cues from his trainer). Similarly, the man in the room doesn't understand Chinese, and could be exposed by watching him closely. (Simon and Eisenstadt do not explain just how this would be done, or how it would affect the argument.) Citing the work of Rudolf Carnap, Simon and Eisenstadt argue that to understand is not just to exhibit certain behavior, but to use "intensions" that determine extensions, and that one can see in actual programs that they do use appropriate intensions. They discuss three actual AI programs, and defend various attributions of mentality to them, including understanding, and conclude that computers understand; they learn "intensions by associating words and other linguistic structure with their denotations, as detected through sensory stimuli". And since we can see exactly how the machines work, "it is, in fact, easier to establish that a machine exhibits understanding that to establish that a human exhibits understanding…." Thus, they conclude, the evidence for empirical strong AI is overwhelming.

Similarly, Daniel Dennett in his original 1980 response to Searle's argument called it "an intuition pump", a term he came up with in discussing the CRA with Hofstader. Dennett's considered view (2013) is that the CRA is "clearly a fallacious and misleading argument …." (p. 320). Paul Thagard (2013) proposes that for every thought experiment in philosophy there is an equal and opposite thought experiment. Thagard holds that intuitions are unreliable, and the CRA is an example (and that in fact the CRA has now been refuted by the technology of autonomous robotic cars). Dennett has elaborated on concerns about our intuitions regarding intelligence. Dennett 1987 ("Fast Thinking") expressed concerns about the slow speed at which the Chinese Room would operate, and he has been joined by several other commentators, including Tim Maudlin, David Chalmers, and Steven Pinker. The operator of the Chinese Room may eventually produce appropriate answers to Chinese questions. But slow thinkers are stupid, not intelligent—and in the wild, they may well end up dead. Dennett argues that "speed … is 'of the essence' for intelligence. If you can't figure out the relevant portions of the changing environment fast enough to fend for yourself, you are not practically intelligent, however complex you are" (326). Thus Dennett relativizes intelligence to processing speed relative to current environment. Tim Maudlin (1989) disagrees. Maudlin considers the time-scale problem pointed to by other writers, and concludes, contra Dennett, that the extreme slowness of a computational system does not violate any necessary conditions on thinking or consciousness.

Steven Pinker (1997) also holds that Searle relies on untutored intuitions. Pinker endorses the Churchlands' (1990) counterexample of an analogous thought experiment of waving a magnet and not generating light, noting that this outcome would not disprove Maxwell's theory that light consists of electromagnetic waves. Pinker holds that the key issue is speed: "The thought experiment slows down the waves to a range to which we humans no longer see them as light. By trusting our intuitions in the thought experiment, we falsely conclude that rapid waves cannot be light either. Similarly, Searle has slowed down the mental computations to a range in which we humans no longer think of it as understanding (since understanding is ordinarily much faster)" (94–95). Howard Gardiner, a supporter of Searle's conclusions regarding the room, makes a similar point about understanding. Gardiner addresses the Chinese Room argument in his book *The Mind's New Science* (1985, 171–177). Gardiner considers all the standard replies to the Chinese Room argument and concludes that Searle is correct about the room: "…the word understand has been unduly stretched in the case of the Chinese room …." (175).

Thus several in this group of critics argue that speed affects our willingness to attribute intelligence and

understanding to a slow system, such as that in the Chinese Room. The result may simply be that our intuitions regarding the Chinese Room are unreliable, and thus the man in the room, in implementing the program, may understand Chinese despite intuitions to the contrary (Maudlin and Pinker). Or it may be that the slowness marks a crucial difference between the simulation in the room and what a fast computer does, such that the man is not intelligent while the computer system is (Dennett).

5. The Larger Philosophical Issues

5.1 Syntax and Semantics

Searle believes the Chinese Room argument supports a larger point, which explains the failure of the Chinese Room to produce understanding. Searle argued that programs implemented by computers are just syntactical. Computer operations are "formal" in that they respond only to the physical form of the strings of symbols, not to the meaning of the symbols. Minds on the other hand have states with meaning, mental contents. We associate meanings with the words or signs in language. We respond to signs because of their meaning, not just their physical appearance. In short, we understand. But, and according to Searle this is the key point, "Syntax is not by itself sufficient for, nor constitutive of, semantics." So although computers may be able to manipulate syntax to produce appropriate responses to natural language input, they do not understand the sentences they receive or output, for they cannot associate meanings with the words.

Searle (1984) presents a three premise argument that because syntax is not sufficient for semantics, programs cannot produce minds.

1. Programs are purely formal (syntactic).
2. Human minds have mental contents (semantics).
3. Syntax by itself is neither constitutive of, nor sufficient for, semantic content.
4. Therefore, programs by themselves are not constitutive of nor sufficient for minds.

The Chinese Room thought experiment itself is the support for the third premise. The claim that syntactic manipulation is not sufficient for meaning or thought is a significant issue, with wider implications than AI, or attributions of understanding. Prominent theories of mind hold that human cognition generally is computational. In one form, it is held that thought involves operations on symbols in virtue of their physical properties. On an alternative connectionist account, the computations are on "subsymbolic" states. If Searle is right, not only Strong AI but also these main approaches to understanding human cognition are misguided.

As we have seen, Searle holds that the Chinese Room scenario shows that one cannot get semantics from syntax alone. In formal logic systems, a kind of artificial language, rules are given for syntax, and this procedure appears to be quite independent of semantics. The logician specifies the basic symbol set and some rules for manipulating strings to produce new ones. These rules are purely formal or syntactic—they are applied to strings of symbols solely in virtue of their syntax or form. A semantics, if any, for the symbol system must be provided separately. And if one wishes to show that interesting additional relationships hold between the syntactic operations and semantics, such as that the symbol manipulations preserve truth, one must provide sometimes complex meta-proofs to show this. So on the face of it, semantics is quite independent of syntax for artificial languages, and one cannot get semantics from syntax alone. "Formal symbols by themselves can never be enough for mental contents, because the symbols, by definition, have no meaning (or interpretation, or semantics) except insofar as someone outside the system gives it to them" (Searle 1989, 45).

Searle's identification of meaning with interpretation in this passage is important. Searle's point is clearly true of the causally inert formal systems of logicians. When we move from formal systems to computational systems, the situation is more complex. As many of Searle's critics (e.g. Cole 1984, Dennett 1987, Boden 1988, and Chalmers 1996) have noted, a computer running a program is not the same as "syntax alone". A computer is an enormously complex electronic causal system. State changes in the system are physical. One can *interpret* the physical states, e.g. voltages, as syntactic 1's and 0's, but the intrinsic reality is electronic and the syntax is "derived", a product of interpretation. The states are syntactically specified by programmers, but they are fundamentally states of a complex causal system embedded in the real world. This is quite different from the abstract formal systems that logicians study. Dennett notes that no "computer program by itself" (Searle's language)—e.g. a program lying on a shelf—can cause anything, even simple addition, let alone mental states. The program must be running. Chalmers (1996) offers a parody in which it is reasoned that recipes are syntactic, syntax is not sufficient for crumbliness, cakes are crumbly, so implementation of a recipe is not sufficient for making a cake. Dennett (1987) sums up the issue: "Searle's view, then, comes to this: take a material object (any material object) that does not have the power of causing mental phenomena; you cannot turn it in to an object that does have the power of producing mental phenomena simply by programming it—reorganizing the conditional dependencies of transitions between its states." Dennett's view is the opposite: programming "is precisely what could give something a mind". But Dennett claims that in fact it is "empirically unlikely that the right sorts of programs can be run on anything but organic, human brains" (325–6).

A further related complication is that it is not clear that computers perform syntactic operations in quite the same sense that a human does—it is not clear that a computer understands syntax or syntactic operations. A computer does not know that it is manipulating 1's and 0's. A computer does not recognize that its binary data strings have a certain form, and thus that certain syntactic rules may be applied to them, unlike the man inside the Chinese Room. Inside a computer, there is nothing that literally reads input data, or that "knows" what symbols are. Instead, there are millions of transistors that change states. A sequence of voltages causes operations to be performed. We humans may choose to interpret these voltages as binary numerals and the voltage changes as syntactic operations, but a computer does not interpret its operations as syntactic or any other way. So perhaps a computer does not need to make the move from syntax to semantics that Searle objects to; it needs to move from complex causal connections to semantics. Furthermore, perhaps *any* causal system is describable as performing syntactic operations—if we interpret a light square as logical "0" and a dark square as logical "1", then a kitchen toaster *may* be described as a device that rewrites logical "0"s as logical "1"s.

In the 1990s, Searle began to use considerations related to these to argue that computational views are not just false, but lack a clear sense. Computation, or syntax, is "observer-relative", not an intrinsic feature of reality: "… you can assign a computational interpretation to anything" (Searle 2002b, p. 17), even the molecules in the paint on the wall. Since nothing is intrinsically computational, one cannot have a scientific theory that reduces the mental, which is not observer-relative, to computation, which is. "Computation exists only relative to some agent or observer who imposes a computational interpretation on some phenomenon. This is an obvious point. I should have seen it ten years ago, but I did not." (Searle 2002b, p.17, originally published 1993).

Critics note that walls are not computers; unlike a wall, a computer goes through state-transitions that are counterfactually described by a program (Chalmers 1996, Block 2002, Haugeland 2002). In his 2002 paper, Block addresses the question of whether a wall is a computer (in reply to Searle's charge that anything that maps onto a formal system is a formal system, whereas minds are quite different). Block denies that whether or not something is a computer depends entirely on our interpretation. Block notes that Searle ignores the counterfactuals that must be true of an implementing system. Haugeland (2002) makes the similar point that an

implementation will be a causal process that reliably carries out the operations—and they must be the right causal powers. Block concludes that Searle's arguments fail, but he concedes that they "do succeed in sharpening our understanding of the nature of intentionality and its relation to computation and representation" (78).

Rey (2002) also addresses Searle's arguments that syntax and symbols are observer-relative properties, not physical. Searle infers this from the fact that they are not defined in physics; it does not follow that they are observer-relative. Rey argues that Searle also misunderstands what it is to realize a program. Rey endorses Chalmers' reply to Putnam: a realization is not just a structural mapping, but involves causation, supporting counterfactuals. "This point is missed so often, it bears repeating: the syntactically specifiable objects over which computations are defined can and standardly do possess a semantics; it's just that the semantics is not involved in the specification." States of a person have their semantics in virtue of computational organization and their causal relations to the world. Rey concludes: Searle "simply does not consider the substantial resources of functionalism and Strong AI." (222) A plausibly detailed story would defuse negative conclusions drawn from the superficial sketch of the system in the Chinese Room.

John Haugeland (2002) argues that there is a sense in which a processor must intrinsically understand the commands in the programs it runs: it executes them in accord with the specifications. "The only way that we can make sense of a computer as executing a program is by understanding its processor as responding to the program prescriptions as meaningful" (385). Thus operation symbols have meaning to a system. Haugeland goes on to draw a distinction between narrow and wide system. He argues that data can have semantics in the wide system that includes representations of external objects produced by transducers. In passing, Haugeland makes the unusual claim, argued for elsewhere, that genuine intelligence and semantics presuppose "the capacity for a kind of commitment in how one lives" which is non-propositional—that is, love (cp. Steven Spielberg's 2001 film *Artificial Intelligence: AI*).

To Searle's claim that syntax is observer-relative, that the molecules in a wall might be interpreted as implementing the Wordstar program (an early word processing program) because "there is some pattern in the molecule movements which is isomorphic with the formal structure of Wordstar" (Searle 1990b, p. 27), Haugeland counters that "the very idea of a complex syntactical token … presupposes specified processes of writing and reading.…" The tokens must be systematically producible and retrievable. So no random isomorphism or pattern somewhere (e.g. on some wall) is going to count, and hence syntax is not observer-relative.

With regard to the question of whether one can get semantics from syntax, William Rapaport has for many years argued for "syntactic semantics", a view in which understanding is a special form of syntactic structure in which symbols (such as Chinese words) are linked to concepts, themselves represented syntactically. Others believe we are not there yet. AI futurist (*The Age of Spiritual Machines*) Ray Kurzweil holds in a 2002 follow-up book that it is red herring to focus on traditional symbol-manipulating computers. Kurzweil agrees with Searle that existent computers do not understand language—as evidenced by the fact that they can't engage in convincing dialog. But that failure does not bear on the capacity of future computers based on different technology. Kurzweil claims that Searle fails to understand that future machines will use "chaotic emergent methods that are massively parallel". This claim appears to be similar to that of connectionists, such as Andy Clark, and the position taken by the Churchlands in their 1990 *Scientific American* article.

Apart from Haugeland's claim that processors understand program instructions, Searle's critics can agree that computers no more understand syntax than they understand semantics, although, like all causal engines, a

computer has syntactic descriptions. And while it is often useful to programmers to treat the machine as if it performed syntactic operations, it is not always so: sometimes the characters programmers use are just switches that make the machine do something, for example, make a given pixel on the computer display turn red, or make a car transmission shift gears. Thus it is not clear that Searle is correct when he says a digital computer is just "a device which manipulates symbols". Computers are complex causal engines, and syntactic descriptions are useful in order to structure the causal interconnections in the machine. AI programmers face many tough problems, but one can hold that they do not have to get semantics from syntax. If they are to get semantics, they must get it from causality.

Two main approaches have developed that explain meaning in terms of causal connections. The internalist approaches, such as Schank's and Rapaport's conceptual representation approaches, and also Conceptual Role Semantics, hold that a state of a physical system gets its semantics from causal connections to other states of the same system. Thus a state of a computer might represent "kiwi" because it is connected to "bird" and "flightless" nodes, and perhaps also to images of prototypical kiwis. The state that represents the property of being "flightless" might get its content from a Negation-operator modifying a representation of "capable of airborne self-propulsion", and so forth, to form a vast connected conceptual network, a kind of mental dictionary.

Externalist approaches developed by Dennis Stampe, Fred Dretske, Hilary Putnam, Jerry Fodor, Ruth Millikan, and others, hold that states of a physical system get their content through causal connections to the external reality they represent. Thus, roughly, a system with a KIWI concept is one that has a state it uses to represent the presence of kiwis in the external environment. This kiwi-representing state can be any state that is appropriately causally connected to the presence of kiwis. Depending on the system, the kiwi representing state could be a state of a brain, or of an electrical device such as a computer, or even of a hydraulic system. The internal representing state can then in turn play a causal role in the determining the behavior of the system. For example, Rey (1986) endorses an indicator semantics along the lines of the work of Dennis Stampe (1977) and Fodor's *Psychosemantics*. Semantics that emphasize causal connection with the world fit well with the Robot Reply. A computer in a robot body might have just the causal connections that could allow its inner syntactic states to have the semantic property of representing states of things in its environment.

Thus there are at least two families of theories (and marriages of the two, as in Block 1986) about how semantics might depend upon causal connections. Both of these attempt to provide accounts that are substance neutral: states of suitably organized causal systems can have content, no matter what the systems are made of. On these theories a computer could have states that have meaning. It is not necessary that the computer be aware of its own states and know that they have meaning, nor that any outsider appreciate the meaning of the states. On either of these accounts meaning depends upon the (possibly complex) causal connections, and digital computers are systems designed to have states that have just such complex causal dependencies. It should be noted that Searle does not subscribe to these theories of semantics. Instead, Searle's discussions of linguistic meaning have often centered on the notion of *intentionality*.

5.2 Intentionality

Intentionality is the property of being about something, having content. In the 19th Century, psychologist Franz Brentano re-introduced this term from Medieval philosophy and held that intentionality was the "mark of the mental". Beliefs and desires are intentional states: they have propositional content (one believes that p, one desires that p, where sentences substitute for "p"). Searle's views regarding intentionality are complex; of relevance here is that he makes a distinction between the original or intrinsic intentionality of genuine mental

states, and the derived intentionality of language. A written or spoken sentence only has derivative intentionality insofar as it is interpreted by someone. It appears that on Searle's view, original intentionality can at least potentially be conscious. Searle then argues that the distinction between original and derived intentionality applies to computers. We can interpret the states of a computer as having content, but the states themselves do not have original intentionality. Many philosophers endorse this intentionality dualism, including Fodor (2009), despite his many differences with Searle.

In a section of her 1988 book, *Computer Models of the Mind*, Margaret Boden notes that intentionality is not well-understood—reason to not put too much weight on arguments that turn on intentionality. Furthermore, insofar as we understand the brain, we focus on informational functions, not unspecified causal powers of the brain: "…from the psychological point of view, it is not the biochemistry as such which matters but the information-bearing functions grounded in it." (241) Responders to Searle have argued that he displays substance chauvinism, in holding that brains understand but systems made of silicon with comparable information processing capabilities cannot, even in principle. Papers on both sides of the issue appeared, such as J. Maloney's 1987 paper "The Right Stuff", defending Searle, and R. Sharvy's 1985 critique, "It Ain't the Meat, it's the Motion". AI proponents such as Kurzweil (1999, see also Richards 2002) have continued to hold that AI systems can potentially have such mental properties as understanding, intelligence, consciousness and intentionality, and will exceed human abilities in these areas.

Other critics of Searle's position take intentionality more seriously than Boden does, but deny his dualistic distinction between original and derived intentionality. Dennett (1987, e.g.) argues that all intentionality is derived. Attributions of intentionality—to animals, other people, even ourselves—are instrumental and allow us to predict behavior, but they are not descriptions of intrinsic properties. As we have seen, Dennett is concerned about the slow speed of things in the Chinese Room, but he argues that once a system is working up to speed, it has all that is needed for intelligence and derived intentionality—and derived intentionality is the only kind that there is, according to Dennett. A machine can be an intentional system because intentional explanations work in predicting the machine's behavior. Dennett also suggests that Searle conflates intentionality with awareness of intentionality. In his syntax-semantic arguments, "Searle has apparently confused a claim about the underivability of semantics from syntax with a claim about the underivability of the consciousness of semantics from syntax" (336). We might also worry that Searle conflates meaning and interpretation, and that Searle's original or underived intentionality is just second-order intentionality, a representation of what an intentional object means. Dretske and others have seen intentionality as information-based. One state of the world, including a state in a computer, may carry information about other states in the world, and this informational aboutness is a mind-independent feature of states. Hence it is a mistake to hold that conscious attributions of meaning are the source of intentionality.

Others have noted that Searle's discussion has shown a shift from issues of intentionality and understanding to issues of consciousness. Searle links intentionality to awareness of intentionality, in that intentional states are at least potentially conscious. In his 1996 book, *The Conscious Mind*, David Chalmers notes that although Searle originally directs his argument against machine intentionality, it is clear from later writings that the real issue is consciousness, which Searle holds is a necessary condition of intentionality. It is consciousness that is lacking in digital computers. Chalmers uses thought experiments to argue that it is implausible that one system has some basic mental property (such as having qualia) that another system lacks, if it is possible to imagine transforming one system into the other, either gradually (as replacing neurons one at a time by digital circuits), or all at once, switching back and forth between flesh and silicon.

A second strategy regarding the attribution of intentionality is taken by externalist critics who in effect argue that

intentionality is an intrinsic feature of states of physical systems that are causally connected with the world in the right way, independently of interpretation (see the preceding Syntax and Semantics section). Fodor's semantic externalism is influenced by Fred Dretske, but they come to different conclusions with regard to the semantics of states of computers. Over a period of years, Dretske developed an historical account of meaning or mental content that would preclude attributing beliefs and understanding to most machines. But Dretske (1985) agrees with Searle that adding machines don't literally add; *we* do the adding, using the machines. Dretske emphasizes the crucial role of natural selection and learning in producing states that have genuine content. Human built systems will be, at best, like Swampmen (beings that result from a lightning strike in a swamp and by chance happen to be a molecule by molecule copy of some human being, say, you)—they appear to have intentionality or mental states, but do not, because such states require the right history. AI states will generally be counterfeits of real mental states; like counterfeit money, they may appear perfectly identical but lack the right pedigree. But Dretske's account of belief appears to make it distinct from conscious awareness of the belief or intentional state (if that is taken to require a higher order thought), and so would allow attribution of intentionality to systems that can learn.

Howard Gardiner endorses Zenon Pylyshyn's criticisms of Searle's view of the relation of brain and intentionality, as supposing that intentionality is somehow a stuff "secreted by the brain", and Pylyshyn's own counter-thought experiment in which one's neurons are replaced one by one with integrated circuit workalikes (see also Chalmers (1996), for exploration of neuron replacement scenarios). Gardiner holds that Searle owes us a more precise account of intentionality than Searle has given so far, and until then it is an open question whether AI can produce it, or whether it is beyond its scope. Gardiner concludes with the possibility that the dispute between Searle and his critics is not scientific, but (quasi?) religious.

5.3 Mind and Body

Several critics have noted that there are metaphysical issues at stake in the original argument. The Systems Reply draws attention to the metaphysical problem of the relation of mind to body. It does this in holding that understanding is a property of the system as a whole, not the physical implementer. The Virtual Mind Reply holds that minds or persons—the entities that understand and are conscious—are more abstract than any physical system, and that there could be a many-to-one relation between minds and physical systems. Thus larger issues about personal identity and the relation of mind and body are in play in the debate between Searle and some of his critics.

Searle's view is that the problem the relation of mind and body "has a rather simple solution. Here it is: Conscious states are caused by lower level neurobiological processes in the brain and are themselves higher level features of the brain." (Searle 2002b, p.9) In his early discussion of the CR, Searle spoke of the causal powers of the brain. Thus his view appears to be that brain states cause consciousness and understanding, and "consciousness is just a feature of the brain" (ibid).

Consciousness and understanding are features of persons, so it appears that Searle accepts a metaphysics in which I, my conscious self, am identical with my brain—a form of mind-brain identity theory. This very concrete metaphysics is reflected in Searle's original presentation of the CR argument, in which Strong AI was described by him as the claim that "the appropriately programmed computer really is a mind" (Searle 1980). This is an identity claim, and has odd consequences. If A and B are identical, any property of A is a property of B. Computers are physical objects. Some computers weigh 6 lbs and have stereo speakers. So the claim that Searle called Strong AI would entail that some minds weigh 6 lbs and have stereo speakers. However it seems

to be clear that while humans may weigh 150 pounds; human minds do not weigh 150 pounds. This suggests that neither bodies nor machines can literally be minds. It appears that minds are more abstract than that, and that at least one version of the claim that Searle calls Strong AI, the version that says that computers are minds, is metaphysically untenable on the face of it, apart from any thought-experiments.

Searle's CR argument was thus directed against the claim that a computer is a mind, that a suitably programmed digital computer understands language, or that its program does. Searle's thought experiment appeals to our strong intuition that someone who did exactly what the computer does would not thereby come to understand Chinese. As noted above, many critics have held that Searle is quite right on this point—no matter how you program a computer, the computer will not literally be a mind and the computer will not understand natural language. This however cannot show that something else understands—it cannot show that AI cannot *produce* understanding of natural language, for this is a different claim. It is not the claim that the computer understands language, or that the program or even the system does. It is the claim that AI creates understanding, with the thing doing the understanding unspecified. This understanding mind might not be identical with the computer, the program, nor the system consisting of computer and program. Hauser (2002) accuses Searle of Cartesian bias in his inference from "it seems to me quite obvious that I understand nothing" to the conclusion that I really understand nothing. Normally, if one understands English or Chinese, one knows that one does—but not necessarily. Searle lacks the normal introspective awareness of understanding—but this, while abnormal, is not conclusive.

Functionalism is a theory of the relation of minds to bodies that was developed in the two decades prior to Searle's CRA. Functionalism is an alternative to the identity theory that is implicit in much of Searle's discussion, as well as to the dominant behaviorism of the mid-Twentieth Century. If functionalism is correct, there appears to be no intrinsic reason why a computer couldn't have mental states. Hence the CRA's conclusion that a computer is intrinsically incapable of mental states is an important consideration against functionalism. Julian Baggini (2009, 37) writes that Searle "came up with perhaps the most famous counter-example in history – the Chinese room argument – and in one intellectual punch inflicted so much damage on the then dominant theory of functionalism that many would argue it has never recovered."

Functionalists hold that a mental state *is* what a mental state *does*—the causal (or "functional") role that the state plays determines what state it is. A functionalist might hold that pain, for example, is a state that is typically caused by damage to the body, is located in a body-image, and is aversive. Functionalists distance themselves both from behaviorists and identity theorists. In contrast with the former, functionalists hold that the *internal* causal processes are important for the possession of mental states. Thus functionalists may reject the Turing Test as too behavioristic. In contrast with identity theorists, functionalists hold that mental states might be had by a variety of physical systems (or non-physical, as in Cole and Foelber 1984, in which a mind changes from a material to an immaterial implementation, neuron by neuron). Thus while an identity theorist will identify pain with certain neuron firings, a functionalist will identify pain with something more abstract and higher level, a functional role that might be had by many different types of underlying system. Functionalists accuse identity theorists of substance chauvinism. However, functionalism remains controversial: functionalism is vulnerable to the Chinese Nation type objections discussed above, and functionalists notoriously have trouble explaining qualia, a problem highlighted by the apparent possibility of an inverted spectrum, where qualitatively different states might have the same functional role (e.g. Block 1978, Maudlin 1989, Cole 1990).

Searle's 2010 statement of the conclusion of the CRA has it showing that computational accounts cannot explain consciousness. There has been considerable interest in the decades since 1980 in determining what does explain

consciousness, and this has been an extremely active research area across disciplines. One interest has been in the neural correlates of consciousness. This bears directly on Searle's claim that consciousness is intrinsically biological and not computational or information processing. There is no definitive answer yet, though some recent work on anesthesia suggests that consciousness is lost when cortical (and cortico-thalamic) connections and information flow are disrupted (e.g.Hudetz 2012, a review article).

In general, if the basis of consciousness is confirmed to be at the relatively abstract level of information flow through neural networks, it will be friendly to functionalism, and if it is turns out to be lower and more biological (or sub-neuronal), it will be friendly to Searle's account.

These controversial biological and metaphysical issues bear on the central inference in the Chinese Room argument. From the intuition that in the CR thought experiment he would not understand Chinese by running a program, Searle infers that there is no understanding created by running a program. Clearly, whether that inference is valid or not turns on a metaphysical question about the identity of persons and minds. If the person understanding is not identical with the room operator, then the inference is unsound.

5.4 Simulation, duplication and evolution

In discussing the CRA, Searle argues that there is an important distinction between simulation and duplication. No one would mistake a computer simulation of the weather for weather, or a computer simulation of digestion for real digestion. It is just as serious a mistake to confuse a computer simulation of understanding with understanding.

On the face of it, this seems true. But two problems emerge. It is not clear that we can always make the distinction between simulations and the real thing. Hearts are biological, if anything is. Are artificial hearts simulations of hearts? Or are they functional duplicates of hearts, hearts made from different materials? Walking is normally a biological phenomenon performed using limbs. Do those with artificial limbs walk? Or do they simulate walking? Do robots walk? If the properties that are needed to be certain kind of thing are high-level properties, anything sharing those properties will be a thing of that kind, even if it differs in its lower level properties. Chalmers (1996) offers a principle governing when simulation is replication. Chalmers suggests that, contra Searle and Harnad (1989), a simulation of *X* can be an *X*, namely when the property of being an *X* is an organizational invariant, a property that depends only on the functional organization of the underlying system, and not on any other details.

Copeland (2002) argues that the Church-Turing thesis does not entail that the brain (or every machine) can be simulated by a universal Turing machine, for the brain (or other machine) might have primitive operations that are not simple clerical routines that can be carried out by hand. Sprevak 2007 raises a related point. Turing's 1938 Princeton thesis described such machines ("O-machines"). If the brain is such a machine, then: "There is no possibility of Searle's Chinese Room Argument being successfully deployed against the functionalist hypothesis that the brain instantiates an O-machine…." (120). Copeland then turns to the Brain Simulator Reply. He argues that Searle correctly notes that one cannot infer from *X* simulates *Y*, and *Y* has property *P*, to the conclusion that therefore *X* has *Y*'s property *P* for arbitrary *P*. But Copeland claims that Searle himself commits the simulation fallacy in extending the CR argument from traditional AI to apply against computationalism. The contrapositive of the inference is logically equivalent—*X* simulates *Y*, *X* does not have *P* therefore *Y* does not—where *P* is understands Chinese. The faulty step is: the CR operator *S* simulates a neural net *N*, it is not the case that *S* understands Chinese, therefore it is not the case that *N* understands Chinese. Copeland also notes results by

Siegelmann and Sontag (1994) showing that some connectionist networks cannot be simulated by a universal Turing Machine (in particular, where connection weights are real numbers).

There is another problem with the simulation-duplication distinction, arising from the process of evolution. Searle wishes to see original intentionality and genuine understanding as properties only of certain biological systems, presumably the product of evolution. Computers merely simulate these properties. At the same time, in the Chinese Room scenario, Searle maintains that a system can exhibit behavior just as complex as human behavior, simulating any degree of intelligence and language comprehension that one can imagine, and simulating any ability to deal with the world, yet not understand a thing. He also says that such behaviorally complex systems might be implemented with very ordinary materials, for example with tubes of water and valves.

This creates a biological problem, beyond the Other Minds problem noted by early critics of the CR argument. While *we* may presuppose that others have minds, evolution makes no such presuppositions. The selection forces that drive biological evolution select on the basis of behavior. Evolution can select for the ability to use information about the environment creatively and intelligently, as long as this is manifest in the behavior of the organism. If there is no overt difference in behavior in any set of circumstances between a system that understands and one that does not, evolution cannot select for genuine understanding. And so it seems that on Searle's account, minds that genuinely understand meaning have no advantage over creatures that merely process information, using purely computational processes. Thus a position that implies that simulations of understanding can be just as biologically adaptive as the real thing, leaves us with a puzzle about how and why systems with "genuine" understanding could evolve. Original intentionality and genuine understanding become epiphenomenal.

Conclusion

As we have seen, since its appearance in 1980 the Chinese Room argument has sparked discussion across disciplines. Despite the extensive discussion there is still no consensus as to whether the argument is sound. At one end we have Julian Baggini's (2009) assessment that Searle "came up with perhaps the most famous counter-example in history – the Chinese room argument – and in one intellectual punch inflicted so much damage on the then dominant theory of functionalism that many would argue it has never recovered." Whereas philosopher Daniel Dennett (2013, p. 320) concludes that the Chinese Room argument is "clearly a fallacious and misleading argument". Hence there is no consensus as to whether the argument is a proof that limits the aspirations of Artificial Intelligence or computational accounts of mind.

Meanwhile work in artificial intelligence and natural language processing has continued. The CRA led Stevan Harnad and others on a quest for "symbol grounding" in AI. Many in philosophy (Dretske, Fodor, Millikan) worked on naturalistic theories of mental content. Speculation about the nature of consciousness continues in many disciplines. And computers have moved from the lab to the pocket.

At the time of Searle's construction of the argument, personal computers were very limited hobbyist devices. Weizenbaum's 'Eliza' and a few text 'adventure' games were played on DEC computers; these included limited parsers. More advanced parsing of language was limited to computer researchers such as Schank. Much changed in the next quarter century; billions now carry computers in their pockets. Has the Chinese Room argument moderated claims by those who produce AI and natural language systems? Some manufacturers linking devices to the "internet of things" make modest claims: appliance manufacturer LG says the second decade of the 21st century brings the "experience of conversing" with major appliances. That may or may not be the same as conversing. Apple is less cautious than LG in describing the capabilities of its "virtual personal assistant"

application called 'Siri': Apple says of Siri that "It understands what you say. It knows what you mean." IBM is quick to claim its much larger 'Watson' system is superior in language abilities to Siri. In 2011 Watson beat human champions on the television game show 'Jeopardy', a feat that relies heavily on language abilities and inference. IBM goes on to claim that what distinguishes Watson is that it "knows what it knows, and knows what it does not know." This appears to be claiming a form of reflexive self-awareness or consciousness for the Watson computer system. Thus the claims of strong AI now are hardly chastened, and if anything some are stronger and more exuberant. At the same time, as we have seen, many others believe that the Chinese Room Argument showed once and for all that at best computers can simulate human cognition.

Though separated by three centuries, Leibniz and Searle had similar intuitions about the systems they consider in their respective thought experiments, Leibniz' Mill and the Chinese Room. In both cases they consider a complex system composed of relatively simple operations, and note that it is impossible to see how understanding or consciousness could result. These simple arguments do us the service of highlighting the serious problems we face in understanding meaning and minds. The many issues raised by the Chinese Room argument may not be settled until there is a consensus about the nature of meaning, its relation to syntax, and about the biological basis of consciousness. There continues to be significant disagreement about what processes create meaning, understanding, and consciousness, as well as what can be proven a priori by thought experiments.

Bibliography

Apple Inc., 2014, 'IOS 7 Siri', accessed 1/10/2014.

Baggini, J., 2009, 'Painting the bigger picture', *The Philosopher's Magazine*, 8: 37–39.

Block, N., 1978, 'Troubles with Functionalism', in C. W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology*, Minneapolis: University of Minnesota Press. (Reprinted in many anthologies on philosophy of mind and psychology.)

——, 1986, 'Advertisement for a Semantics for Psychology', *Midwest Studies in Philosophy* (Volume X), P.A. French, *et al*. (eds.), Minneapolis: University of Minnesota Press, 615–678.

——, 2002, 'Searle's Arguments Against Cognitive Science', in Preston and Bishop (eds.) 2002.

Boden, M., 1988, *Computer Models of the Mind*, Cambridge: Cambridge University Press; pp. 238–251 were excerpted and published as 'Escaping from the Chinese Room', in *The Philosophy of Artificial Intelligence*, ed M. A. Boden, New York: Oxford University Press, 1990.

Cam, P., 1990, 'Searle on Strong AI', *Australasian Journal of Philosophy*, 68: 103–8.

Chalmers, D., 1992, 'Subsymbolic Computation and the Chinese Room', in J. Dinsmore (ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap*, Hillsdale, NJ: Lawrence Erlbaum.

——, 1996, *The Conscious Mind*, Oxford: Oxford University Press.

——, 1996a, "Does a Rock Implement Every Finite-State Automaton', *Synthese* 108: 309–33.

——, 1996b, 'Minds, machines, and mathematics', *Psyche*, 2: 11–20.

Churchland, P., 1985, 'Reductionism, Qualia, and the Direct Introspection of Brain States', *The Journal of Philosophy*, LXXXII: 8–28.

Churchland, P. and Churchland, P., 1990, 'Could a machine think?', *Scientific American*, 262(1): 32–37.

Clark, A., 1991, *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, MA: MIT Press.

Cole, D., 1984, 'Thought and Thought Experiments', *Philosophical Studies*, 45: 431–44.

——, 1990, ' Functionalism and Inverted Spectra ', *Synthese*, 82: 202–222.

——, 1991a, ' Artificial Intelligence and Personal Identity', *Synthese*, 88: 399–417.

——, 1991b, ‘Artificial Minds: Cam on Searle’, *Australasian Journal of Philosophy*, 69: 329–33.

——, 1994, ‘The Causal Powers of CPUs’, in E. Dietrich (ed.), *Thinking Computers and Virtual Persons*, New York: Academic Press

Cole, D. and Foelber, R., 1984, Contingent Materialism’, *Pacific Philosophical Quarterly*, 65(1): 74–85.

Copeland, J., 2002, ‘The Chinese Room from a Logical Point of View’, in Preston and Bishop (eds.) 2002, 104–122.

Crane, Tim., 1996, *The Mechanical Mind*: *A Philosophical Introduction to Minds, Machines and Mental Representation*, London: Penguin.

Davis, Lawrence, 2001, ‘Functionalism, the Brain, and Personal Identity’, *Philosophical Studies*, 102(3): 259–279.

Dennett, D., 1978, ‘Toward a Cognitive Theory of Consciousness’, in *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, MA: MIT Press.

——, 1981, ‘Where am I?’ in *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, MA: MIT Press, pp. 310–323.

——, 1987, ‘Fast Thinking’, in *The Intentional Stance*, Cambridge, MA: MIT Press, 324–337.

——, 1997, ‘Consciousness in Humans and Robot Minds,’ in M. Ito, Y. Miyashita and E.T. Rolls (eds.), *Cognition, computation, and consciousness*, New York: Oxford University Press, pp. 17–29.

——, 2013, *Intuition Pumps and Other Tools for Thought*, New York: W.W.Norton and Co.

Double, R., 1983, ‘Searle, Programs and Functionalism’, *Nature and System*, 5: 107–14.

Dretske, F. 1985, ‘Presidential Address’ (Central Division Meetings of the American Philosophical Association), *Proceedings and Addresses of the American Philosophical Association*, 59(1): 23–33.

Fodor, J., 1987, *Psychosemantics*, Cambridge, MA: MIT Press.

——, 1991, ‘Yin and Yang in the Chinese Room’, in D. Rosenthal (ed.), *The Nature of Mind*, New York: Oxford University Press.

——, 1992, *A Theory of Content and other essays*, Cambridge, MA: MIT Press.

——, 2009, ‘Where is my Mind?’, *London Review of Books*, (31)3: 13–15.

Ford, J., 2010, ‘Helen Keller was never in a Chinese Room’, *Minds and Machines*, VOLUME: PAGES.

Gardiner, H., 1987, *The Mind's New Science: A History of the Cognitive Revolution*, New York: Basic Books.

Hanley, R., 1997, *The Metaphysics of Star Trek*, New York: Basic Books.

Harnad, S., 1989, ‘Minds, Machines and Searle’, *Journal of Experimental and Theoretical Artificial Intelligence*, 1: 5–25.

——, 2002, ‘Minds, Machines, and Searle2: What's Right and Wrong about the Chinese Room Argument’, in Preston and Bishop (eds.) 2002, 294–307.

Haugeland, J., 2002, ‘Syntax, Semantics, Physics’, in Preston and Bishop (eds.) 2002, 379–392.

Hauser, L., 1997, ‘Searle's Chinese Box: Debunking the Chinese Room Argument’, *Minds and Machines*, 7: 199–226.

——, 2002, ‘Nixin’ Goes to China’, in Preston and Bishop (eds.) 2002, 123–143.

Hayes, P., Harnad, S., Perlis, D. & Block, N., 1992, ‘Virtual Symposium on Virtual Mind’, *Minds and Machines*, 2(3): 217–238.

Hofstadter, D., 1981, ‘Reflections on Searle’, in Hofstadter and Dennett (eds.), *The Mind's I*, New York: Basic Books, pp. 373–382.

Horgan, T., 2013, ‘Original Intentionality is Phenomenal Intentionality’, *The Monist* 96: 232–251.

Hudetz, A., 2012, ‘General Anesthesia and Human Brain Connectivity’, *Brain Connect*, 2(6): 291–302.

Jackson, F., 1986, ‘What Mary Didn't Know’, *Journal of Philosophy*, LXXXIII: 291–5.

Kaernbach, C., 2005, 'No Virtual Mind in the Chinese Room', *Journal of Consciousness Studies*, 12(11): 31–42.

Kurzweil, R., 2000, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, New York: Penguin.

——, 2002, 'Locked in his Chinese Room', in Richards 2002, 128–171.

Maloney, J., 1987, 'The Right Stuff', *Synthese*, 70: 349–72.

Maudlin, T., 1989, 'Computation and Consciousness', *Journal of Philosophy*, LXXXVI: 407–432.

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

Moravec, H., 1999, *Robot: Mere Machine to Transcendent Mind*, New York: Oxford University Press.

Nute, D., 2011, 'A Logical Hole the Chinese Room Avoids', *Minds and Machines*, 21: 431–3; this is a reply to Shaffer 2009.

Penrose, R., 2002, 'Consciousness, Computation, and the Chinese Room' in Preston and Bishop (eds.) 2002, 226–249.

Pinker, S., 1997, *How the Mind Works*, New York: Norton.

Preston, J. and M. Bishop (eds.), 2002, *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, New York: Oxford University Press.

Pylyshyn, Z., 1980, Reply to Searle,*Behavioral and Brain Sciences*, 3.

Rapaport, W., 1984, 'Searle's Experiments with Thought', *Philosophy of Science*, 53: 271–9.

—— 2006, "How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room", *Minds and Machines*, 16(4): 381–436.

Rey, G., 1986, 'What's Really Going on in Searle's "Chinese Room"', *Philosophical Studies*, 50: 169–85.

——, 2002, 'Searle's Misunderstandings of Functionalism and Strong AI', in Preston and Bishop (eds.) 2002, 201–225.

Richards, J. W. (ed.), 2002, *Are We Spiritual Machines: Ray Kurzweil vs. the Critics of Strong AI*, Seattle: Discovery Institute.

Rosenthal, D. (ed), 1991, *The Nature of Mind*, Oxford and NY: Oxford University Press.

Schank, R. and Abelson, R., 1977, *Scripts, Plans, Goals, and Understanding*, Hillsdale, NJ: Lawrence Erlbaum.

Schank, R. and P. Childers, 1985, *The Cognitive Computer: On Language, Learning, and Artificial Intelligence*, New York: Addison-Wesley.

Schweizer, P., 2012, 'The Externalist Foundations of a Truly Total Turing Test', *Minds and Machines*, 22: 191–212.

Searle, J., 1980, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3: 417–57 [Preprint available online]

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

——, 1989, 'Artificial Intelligence and the Chinese Room: An Exchange', *New York Review of Books*, 36: 2 (February 16, 1989).

——, 1990a, 'Is the Brain's Mind a Computer Program?', *Scientific American*, 262(1): 26–31.

——, 1990b, 'Presidential Address', *Proceedings and Addresses of the American Philosophical Association*, 64: 21–37.

——, 1998, 'Do We Understand Consciousness?' (Interview with Walter Freeman), *Journal of Consciousness Studies*, 6: 5–6.

——, 1999, 'The Chinese Room', in R.A. Wilson and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.

——, 2002a, 'Twenty-one Years in the Chinese Room', in Preston and Bishop (eds.) 2002, 51–69.

——, 2002b, 'The Problem of Consciousness', in *Consciousness and Language*, Cambridge: Cambridge University Press, 7–17.
——, 2010, ' Why Dualism (and Materialism) Fail to Account for Consciousness ' in Lee, Richard E. (ed) *Questioning Nineteenth Century Assumptions about Knowledge, III: Dualism*. NY: SUNY Press.
Shaffer, M., 2009, 'A Logical Hole in the Chinese Room', *Minds and Machines*, 19(2): 229–235.
Sharvy, R., 1985, 'It Ain't the Meat It's the Motion', *Inquiry*, 26: 125–134.
Simon, H. and Eisenstadt, S., 2002, 'A Chinese Room that Understands', in Preston and Bishop (eds.) 2002, 95–108.
Sloman, A. and Croucher, M., 1980, 'How to turn an information processor into an understanding', *Brain and Behavioral Sciences*, 3: 447–8.
Sprevak, M., 2007, 'Chinese Rooms and Program Portability', *British Journal for the Philosophy of Science*, 58(4): 755–776.
Stampe, Dennis, 1977, 'Towards a Causal Theory of Linguistic Representation', in P. French, T. Uehling, H. Wettstein, (eds.) *Contemporary Perspectives in the Philosophy of Language*, (Midwest Studies in Philosophy, Volume 2), Minneapolis: University of Minnesota Press, pp. 42–63.
Thagard, P., 1986, 'The Emergence of Meaning: An Escape from Searle's Chinese Room', *Behaviorism*, 14: 139–46.
——, 2013, 'Thought Experiments Considered Harmful', *Perspectives on Science*, 21: 122–139.
Turing, A., 1948, 'Intelligent Machinery: A Report', London: National Physical Laboratory.
——, 1950, 'Computing Machinery and Intelligence', *Mind*, 59: 433–460.
Weiss, T., 1990, 'Closing the Chinese Room', *Ratio*, 3: 165–81.

Academic Tools

How to cite this entry.
Preview the PDF version of this entry at the Friends of the SEP Society.
Look up this entry topic at the Indiana Philosophy Ontology Project (InPhO).
Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

- Papers on the Chinese Room Argument, at PhilPapers.org.
- Annotated Chinese Room Bibliography, by L. Hauser.
- Harnad, S., 2012, 'Alan Turing and the 'Hard' and 'Easy' Problem of Cognition: Doing and Feeling," *Turing100: Essays in Honour of Centenary Turing Year 2012*, available online.
- Searle, J., Failures of Computationalism (Searle's reply to Harnad, and Harnad's response)

Related Entries

computation: in physical systems | consciousness: and intentionality | consciousness: representational theories of | emergent properties | epiphenomenalism | functionalism | information: biological | information: semantic conceptions of | intentionality | mental content: causal theories of | mental content: externalism about | mental content: teleological theories of | mental representation | mind: computational theory of | multiple realizability |

neuroscience, philosophy of | other minds | thought experiments | Turing, Alan | Turing test | zombies

Library of Congress Catalog Data: ISSN 1095-5054