Chinese room

From Wikipedia, the free encyclopedia

The **Chinese room** argument comprises a thought experiment and associated arguments by John Searle (1980), which attempts to show that a symbol-processing machine like a computer can never be properly described as having a "mind" or "understanding", regardless of how intelligently it may behave.

Contents

Chinese room thought experiment

Searle's thought experiment begins with this hypothetical premise: suppose that artificial intelligence research has succeeded in constructing a computer that behaves as if it understands Chinese. It takes Chinese characters as input and, by following the instructions of a computer program, produces other Chinese characters, which it presents as output. Suppose, says Searle, that this computer performs its task so convincingly that it comfortably passes the Turing test: it convinces a human Chinese speaker that the program is itself a human Chinese speaker. To all of the questions that the human asks, it makes appropriate responses, such that any Chinese speaker would be convinced that he or she is talking to another Chinese-speaking human being.

Some proponents of artificial intelligence would conclude that the computer "understands" Chinese. [1] This conclusion, a position he refers to as strong AI, is the target of Searle's argument.

Searle then asks the reader to suppose that he is in a closed room and that he has a book with an English version of the aforementioned computer program, along with sufficient paper, pencils,

erasers and filing cabinets. He can receive Chinese characters (perhaps through a slot in the door), process them according to the program's instructions, and produce Chinese characters as output. As the computer had passed the Turing test this way, it is fair, says Searle, to deduce that the human operator will be able to do so as well, simply by running the program manually.

Searle asserts that there is no essential difference between the role the computer plays in the first case and the role the human operator plays in the latter. Each is simply following a program, step-by-step, which simulates intelligent behavior. And yet, Searle points out, the human operator does not understand a word of Chinese. Since it is obvious that he does not understand Chinese, Searle argues, we must infer that the computer does not understand Chinese either.

Searle argues that without "understanding" (what philosophers call "intentionality"), we cannot describe what the machine is doing as "thinking". Because it does not think, it does not have a "mind" in anything like the normal sense of the word, according to Searle. Therefore, he concludes, "strong AI" is mistaken.

History

Searle's argument first appeared in his paper "Minds, Brains, and Programs", published in *Behavioral and Brain Sciences* in 1980.[1] It eventually became the journal's "most influential target article",[2] generating an enormous number of commentaries and responses in the ensuing decades.

Most of the discussion consists of attempts to refute it. "The overwhelming majority," notes *BBS* editor Stevan Harnad, "still think that the Chinese Room Argument is dead wrong."[3] The sheer volume of the literature that has grown up around it inspired Pat Hayes to quip that the field of cognitive science ought to be redefined as "the ongoing research program of showing Searle's Chinese Room Argument to be false."[2]

Despite the controversy (or perhaps because of it), the paper has become "something of a classic in cognitive science," according to Harnad.[3] Varol Akman agrees, and has described Searle's paper as "an exemplar of philosophical clarity and purity".[4]

Searle's targets: "strong AI" and computationalism

Although the Chinese Room argument was originally presented in reaction to the statements of AI researchers, philosophers have come to view it as an important part of the philosophy of mind. It is a challenge to functionalism and the computational theory of mind,[5] and is related to such questions as the mind-body problem,[6] the problem of other minds,[7] the symbol-grounding problem and the hard problem of consciousness.[8]

Strong AI

Searle identified a philosophical position he calls "strong AI":

> The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds.[9]

The definition hinges on the distinction between *simulating* a mind and *actually having* a mind. Searle writes that "according to Strong AI, the correct simulation really is a mind. According to Weak AI, the correct simulation is a model of the mind."[10]

The position is implicit in some of the statements of early AI researchers and analysts. For example,

in 1955, AI founder Herbert Simon declared that "there are now in the world machines that think, that learn and create"[11] and claimed that they had "solved the venerable mind-body problem, explaining how a system composed of matter can have the properties of mind."[12] John Haugeland wrote that "AI wants only the genuine article: *machines with minds*, in the full and literal sense. This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely, we are, at root, *computers ourselves*."[13]

Searle also ascribes the following positions to advocates of strong AI:

- AI systems can be used to explain the mind;[14]
- The study of the brain is irrelevant to the study of the mind;[15] and
- The Turing test is adequate for establishing the existence of mental states.[16]

Strong AI as philosophy

Stevan Harnad argues that Searle's depictions of strong AI can be reformulated as "recognizable tenets of *computationalism*, a position (unlike 'strong AI') that is actually held by many thinkers, and hence one worth refuting."[17] Computationalism[18] is the position in the philosophy of mind which argues that the mind can be accurately described as an information-processing system.

Each of the following, according to Harnad, is a "tenet" of computationalism:[19]

- Mental states are computational states (which is why computers can have mental states and help to explain the mind);
- Computational states are implementation-independent — in other words, it is the software that determines the computational state, not the hardware (which is why the brain, being hardware, is irrelevant); and that
- Since implementation is unimportant, the only empirical data that matters is how the system functions; hence the Turing test is definitive. This last point is a version of functionalism.

Searle accuses strong AI of dualism, the idea that the mind and the body are made up of different "substances". He writes that "strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter."[20] He rejects any form of dualism, writing that "brains cause minds"[21] and that "actual human mental phenomena [are] dependent on actual physical-chemical properties of actual human brains",[20] a position called "biological naturalism" (as opposed to alternatives like behaviourism, functionalism, identity theory and dualism).[22]

Searle's argument centers on "understanding" — that is, mental states with what philosophers call "intentionality" — and does not directly address other closely related ideas, such as "intelligence" or "consciousness". David Chalmers has argued that, to the contrary, "it is fairly clear that consciousness is at the root of the matter".[23]

Strong AI v. AI research

Searle's argument does not limit the intelligence with which machines can behave or act; indeed, it fails to address this issue directly, leaving open the possibility that a machine could be built that *acts* intelligently but does not have a mind or intentionality in the same way that brains do.

Since the primary mission of artificial intelligence research is only to create useful systems that act intelligently, Searle's arguments are not usually considered an issue for AI research. Stuart Russell and Peter Norvig observe that most AI researchers "don't care about the strong AI hypothesis—as

long as the program works, they don't care whether you call it a simulation of intelligence or real intelligence."[24]

Searle's "strong AI" should not be confused with "strong AI" as defined by Ray Kurzweil and other futurists,[25] who use the term to describe machine intelligence that rivals human intelligence. Kurzweil is concerned primarily with the *amount* of intelligence displayed by the machine, whereas Searle's argument sets no limit on this, as long as it is understood that it is merely a simulation and not the real thing.

Replies

Replies to Searle's argument may be classified according to what they claim to show:[26]

- Those which identify *who* speaks Chinese;
- Those which demonstrate how meaningless symbols can become meaningful;
- Those which suggest that the Chinese room should be redesigned in some way; and
- Those which demonstrate the ways in which Searle's argument is misleading.

Some of the arguments (robot and brain simulation, for example) fall into multiple categories.

System and virtual mind replies: finding the mind

These two replies attempt to answer the question: since the man in the room doesn't speak Chinese, *where* is the "mind" that does? These replies address the key ontological issues of mind vs. body and simulation vs. reality.

Systems reply.[27] The "systems reply" argues that it is the *whole system* that understands Chinese, consisting of the room, the book, the man, the paper, the pencil and the filing cabinets. While the man by himself can only understand English, the complete system can understand Chinese. The man is part of the system, just as the hippocampus is a part of the brain. The fact that the man doesn't understand Chinese is irrelevant and is no more surprising than the fact that the hippocampus understands nothing by itself.

Searle responds to this position by asking what happens if the man memorizes the rules and keeps track of everything in his head. Then, Searle argues, the only component of the system is the man himself. Searle argues that if the man doesn't understand Chinese then the system (which Searle says consists only of the man) doesn't understand Chinese either and the fact that the man *appears* to understand Chinese proves nothing.[28]

Searle suggests with his response that by memorizing the program, the program has become part of the man—but for the program, which understands Chinese, the man is still simply providing the hardware on which it runs. This type of elaboration of the system reply is called the *virtual mind reply*.

Virtual mind reply.[29] A more subtle version of the systems reply is that the Chinese-speaking mind in Searle's room is a "virtual mind", similar to the virtual machines used in computer science. A fundamental property of computing machinery is that one machine can "implement" another: any (Turing complete) computer can do a step-by-step simulation of any other machine.[30] In this way, a machine can simultaneously be two machines at once: for example, it can be a Macintosh and a word processor at the same time. A virtual machine depends on the hardware (in that if you turn off the Macintosh, you turn off the word processor as well), yet is different from the hardware. (This is how the position resists dualism: there can be two machines in the same place, both made of the same substance, if one of them is virtual.) A virtual machine is also "implementation independent" in that

it doesn't matter what sort of hardware it runs on: a PC, a Macintosh, a supercomputer, a brain or Searle in his Chinese room.[31]

To clarify the distinction between the systems reply and virtual mind reply, David Cole notes that a program could be written that implements two minds at once—for example, one speaking Chinese and the other Korean. While there is only one system and only one man in the room, there may be an unlimited number of "virtual minds."[32]

Searle would respond that such a mind is only a simulation. He writes: "No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched."[33] Nicholas Fearn responds that, for some things, simulation is as good as the real thing. "When we call up the pocket calculator function on a desktop computer, the image of a pocket calculator appears on the screen. We don't complain that 'it isn't *really* a calculator', because the physical attributes of the device do not matter."[34] The question is, is the human mind like the pocket calculator, essentially composed of information? Or is it like the rainstorm, which can't be duplicated using digital information alone? (The issue of simulation is also discussed in the article synthetic intelligence.)

What they do and don't prove. These replies provide an explanation of exactly who it is that understands Chinese. If there is something *besides* the man in the room that can understand Chinese, Searle can't argue that (1) the man doesn't understand Chinese, therefore (2) nothing in the room understands Chinese. This, according to those who make this reply, shows that Searle's argument fails to prove that "strong AI" is false.[35]

However, the replies, by themselves, do not prove that strong AI is *true*, either: they provide no evidence that the system (or the virtual mind) understands Chinese, other than the hypothetical premise that it passes the Turing Test. As Searle writes "the systems reply simply begs the question by insisting that system must understand Chinese."[28]

Robot and semantics replies: finding the meaning

As far as the man in the room is concerned, the symbols he writes are just meaningless "squiggles." But if the Chinese room really "understands" what it's saying, then the symbols must get their meaning from somewhere. These arguments attempt to connect the symbols to the things they symbolize. These replies address Searle's concerns about intentionality, symbol grounding and syntax vs. semantics.

Robot reply.[36] Suppose that instead of a room, the program was placed into a robot that could wander around and interact with its environment. This would allow a "causal connection" between the symbols and things they represent. Hans Moravec comments: 'If we could graft a robot to a reasoning program, we wouldn't need a person to provide the meaning anymore: it would come from the physical world."[37]

Searle's reply is to suppose that, unbeknownst to the individual in the Chinese room, some of the inputs he was receiving came directly from a camera mounted on a robot, and some of the outputs were used to manipulate the arms and legs of the robot. Nevertheless, the person in the room is still just following the rules, and *does not know what the symbols mean.* Searle writes "he doesn't *see* what comes into the robot's eyes."[38] (See Mary's room for a similar thought experiment.)

Derived meaning.[39] Some respond that the room, as Searle describes it, *is* connected to the world: through the Chinese speakers that it is "talking" to and through the programmers who designed the knowledge base in his file cabinet. The symbols he manipulates *are already meaningful*, they're just not meaningful to *him*.

Searle complains that the symbols only have a "derived" meaning, like the meaning of words in books. The meaning of the symbols depends on the conscious understanding of the Chinese speakers and the programmers outside the room. The room, according to Searle, has no understanding of its own.[40]

Commonsense knowledge / contextualist reply.[41] Some have argued that the meanings of the symbols would come from a vast "background" of commonsense knowledge encoded in the program and the filing cabinets. This would provide a "context" that would give the symbols their meaning.

Searle agrees that this background exists, but he does not agree that it can be built into programs. Hubert Dreyfus has also criticized the idea that the "background" can be represented symbolically. [42]

What they do and don't prove. To each of these suggestions, Searle's response is the same: no matter how much knowledge is written into the program and no matter how the program is connected to the world, he is still in the room manipulating symbols according to rules. His actions are syntactic and this can never explain to him what the symbols stand for. Searle writes "syntax is insufficient for semantics."[43]

However, for those who accept that Searle's actions simulate a mind, separate from his own, the important question is not what the symbols mean *to Searle*, what is important is what they mean *to the virtual mind.* While Searle is trapped in the room, the virtual mind is not: it is connected to the outside world through the Chinese speakers it speaks to, through the programmers who gave it world knowledge, and through the cameras and other sensors that roboticists can supply.

Brain simulation and connectionist replies: redesigning the room

These arguments are all versions of the systems reply that identify a particular *kind* of system as being important. They try to outline what kind of a system would be able to pass the Turing test and give rise to conscious awareness in a machine. (Note that the "robot" and "commonsense knowledge" replies above also specify a certain kind of system as being important.)

Brain simulator reply.[44] Suppose that the program simulated in fine detail the action of every neuron in the brain of a Chinese speaker. This strengthens the intuition that there would be no significant difference between the operation of the program and the operation of a live human brain.

Searle replies that such a simulation will not have reproduced the important features of the brain — its causal and intentional states. Searle is adamant that "human mental phenomena [are] dependent on actual physical-chemical properties of actual human brains."[20]

Two variations on the brain simulator reply are:

> **China brain**.[45] What if we ask each citizen of China to simulate one neuron, using the telephone system to simulate the connections between axons and dendrites? In this version, it seems obvious that no individual would have any understanding of what the brain might be saying.
>
> **Brain replacement scenario**.[46] In this, we are asked to imagine that engineers have invented a tiny computer that simulates the action of an individual neuron. What would happen if we replaced one neuron at a time? Replacing one would clearly do nothing to change conscious awareness. Replacing all of them would create a digital computer that simulates a brain. If Searle is right, then conscious awareness must disappear during the procedure (either

gradually or all at once). Searle's critics argue that there would be no point during the procedure when he can claim that conscious awareness ends and mindless simulation begins. [47]

Connectionist replies.[48] Closely related to the brain simulator reply, this claims that a massively parallel connectionist architecture would be capable of understanding.

Combination reply.[49] This response combines the robot reply with the brain simulation reply, arguing that a brain simulation connected to the world through a robot body could have a mind.

What they do and don't prove. Arguments such as these (and the robot and commonsense knowledge replies above) recommend that Searle's room be redesigned. Searle's replies all point out that, however the program is written or however it is connected to the world, it is still being *simulated* by a simple step by step Turing complete machine (or machines). These machines are still just like the man in the room: they understand nothing and don't speak Chinese. They are merely manipulating symbols without knowing what they mean.

Searle also argues that, if features like a robot body or a connectionist architecture are *required*, then strong AI (as he understands it) has been abandoned.[50] Either (1) Searle's room can't pass the Turing test, because formal symbol manipulation (by a Turing complete machine) is not enough, or (2) Searle's room *could* pass the Turing test, but the Turing test is not sufficient to determine if the room has a "mind." Either way, it denies one or the other of the positions Searle thinks of "strong AI", proving his argument.

The brain arguments also suggests that computation can't provide an *explanation* of the human mind (another aspect of what Searle thinks of as "strong AI"). They assume that there is no simpler way to describe the mind than to create a program that is just as mysterious as the brain was. He writes "I thought the whole idea of strong AI was that we don't need to know how the brain works to know how the mind works."[51]

Other critics don't argue that these improvements are *necessary* for the Chinese room to pass the Turing test or to have a mind. They accept the premise that the room as Searle describes it does, in fact, have a mind, but they argue that it is difficult to see—Searle's description is correct, but *misleading*. By redesigning the room more realistically they hope to make this more obvious. In this case, these arguments are being used as appeals to intuition (see next section). Searle's intuition, however, is never shaken. He writes: "I can have any formal program you like, but I still understand nothing."[52]

In fact, the room can just as easily be redesigned to *weaken* our intuitions. Ned Block's "blockhead" argument (Block 1981) suggests that the program could, in theory, be rewritten into a simple lookup table of rules of the form "if the user writes *S*, reply with *P* and goto X". Any program can be rewritten (or "refactored") into this form, even a brain simulation.[53] In the blockhead scenario, the entire mental state is hidden in the letter X, which represents a memory address—a number associated with the next rule. It is hard to visualize that an instant of our conscious experience can be captured in a single large number, yet this is exactly what "strong AI" claims.

Speed, complexity and other minds: appeals to intuition

The following arguments (and the intuitive interpretations of the arguments above) do not directly explain how a Chinese speaking mind could exist in Searle's room, or how the symbols he manipulates could become meaningful. However, by raising doubts about Searle's intuitions they support other positions, such as the system and robot replies.

Speed and complexity replies.[54] The speed at which our brains process information is (by some estimates) 100,000,000,000 operations per second.[55] Several critics point out that the man in the room would probably take millions of years to respond to a simple question, and would require "filing cabinets" of astronomical proportions. This brings the clarity of Searle's intuition into doubt.

An especially vivid version of the speed and complexity reply is from Paul and Patricia Churchland. They propose this analogous thought experiment:

> **Churchland's luminous room.**[56] Suppose a philosopher finds it inconceivable that light is caused by waves of electromagnetism. He could go into a dark room and wave a magnet up and down. He would see no light, of course, and he could claim that he had proved light is not a magnetic wave and that he has refuted Maxwell's equations. The problem is that he would have to wave the magnet up and down something like 450,000,000,000,000 times a second in order to see anything.

Several of the replies above address the issue of complexity. The connectionist reply emphasizes that a working artificial system would have to be as complex and as interconnected as the human brain. The commonsense knowledge reply emphasizes that any program that passed a Turing test would have to be "an extraordinarily supple, sophisticated, and multilayered system, brimming with 'world knowledge' and meta-knowledge and meta-meta-knowledge," as Daniel Dennett explains.[57]

Stevan Harnad is critical of speed and complexity replies when they stray beyond addressing our intuitions. He writes "Some have made a cult of speed and timing, holding that, when accelerated to the right speed, the computational may make a phase transition into the mental. It should be clear that is not a counterargument but merely an *ad hoc* speculation (as is the view that it is all just a matter of ratcheting up to the right degree of 'complexity.')"[58]

Other minds reply.[59] This reply points out that Searle's argument is a version of the problem of other minds, applied to machines. There is no way we can determine if other people's subjective experience is the same as our own. We can only study their behavior (i.e., by giving them our own Turing test). Critics of Searle argue that he is holding the Chinese room to a higher standard than we would hold an ordinary person.

Nils Nilsson writes "If a program behaves *as if* it were multiplying, most of us would say that it is, in fact, multiplying. For all I know, Searle may only be behaving *as if* he were thinking deeply about these matters. But, even though I disagree with him, his simulation is pretty good, so I'm willing to credit him with real thought."[60]

Alan Turing (writing 30 years before Searle presented his argument) noted that people never consider the problem of other minds when dealing with each other. He writes that "instead of arguing continually over this point it is usual to have the polite convention that everyone thinks."[61] The Turing test simply extends this "polite convention" to machines. He doesn't intend to solve the problem of other minds (for machines or people) and he doesn't think we need to.[62]

Searle believes that there are "causal properties" in our neurons that give rise to the mind. However, these causal properties can't be detected by anyone outside the mind, otherwise the Chinese Room couldn't pass the Turing test—the people outside would be able to tell there wasn't a Chinese speaker in the room by detecting their causal properties. Since they can't detect causal properties, they can't detect the existence of the mental. Russell & Norvig (2003) argue that this implies the human mind, as Searle describes it, is epiphenomenal: that it "casts no shadow." To make this point clear, Daniel Dennett suggests this version of the "other minds" reply:

Dennett's reply from natural selection.[63] Suppose that, by some mutation, a human being is born that does not have Searle's "causal properties" but nevertheless acts exactly like a human being. (This sort of animal is called a "zombie" in thought experiments in the philosophy of mind). This new animal would reproduce just as any other human and eventually there would be more of these zombies. Natural selection would favor the zombies, since their design is (we could suppose) a bit simpler. Eventually the humans would die out. So therefore, if Searle is right, it's most likely that human beings (as we see them today) are actually "zombies," who nevertheless insist they are conscious. This suggests it's unlikely that Searle's "causal properties" would have ever evolved in the first place. Nature has no incentive to create them.

Searle disagrees with this analysis and insists that we must "presuppose the reality and knowability of the mental."[64] and that "The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't ... what we wanted to know is what distinguishes the mind from thermostats and livers."[38] He takes it as obvious that we can detect the presence of other minds and dismisses this reply as being off the point.

What they do and don't prove. These arguments apply only to our intuitions. (As do the arguments above which are intended to make it seem more plausible that the Chinese room contains a mind, which can include the robot, commonsense knowledge, brain simulation and connectionist replies.) They do not directly prove that a machine can or can't have a mind.

However, some critics believe that Searle's argument relies entirely on intuitions. Ned Block writes "Searle's argument depends for its force on intuitions that certain entities do not think."[65] Daniel Dennett describes the Chinese room argument as an "intuition pump"[66] and writes "Searle's thought experiment depends, illicitly, on your imagining too simple a case, an irrelevant case, and drawing the 'obvious' conclusion from it."[67]

These arguments, if accepted, prevent Searle from claiming that his conclusion is obvious by undermining the intuitions that his certainty requires.

Formal arguments

Searle has produced a more formal version of the argument of which the Chinese Room forms a part. He presented the first "excessively crude"[68] version in 1984. The version given below is from 1990.[69]

The part of the argument which should be controversial is A3 and it is this point which the Chinese room thought experiment is intended to prove.[70]

He begins with three axioms:

(A1) "Programs are formal (syntactic)."
A program uses syntax to manipulate symbols and pays no attention to the semantics of the symbols. It knows where to put the symbols and how to move them around, but it doesn't know what they stand for or what they mean. For the program, the symbols are just physical objects like any others.

(A2) "Minds have mental contents (semantics)."
Unlike the symbols used by a program, our thoughts have meaning: they represent things and we know what it is they represent.

(A3) "Syntax by itself is neither constitutive of nor sufficient for semantics."
This is what the Chinese room argument is intended to prove: the Chinese room has syntax (because there is a man in there moving symbols around). The Chinese room has no semantics (because, according to Searle, there is no one or nothing in the room that understands what the symbols mean). Therefore, having syntax is not enough to generate semantics.

Searle posits that these lead directly to this conclusion:

(C1) Programs are neither constitutive of nor sufficient for minds.
This should follow without controversy from the first three: Programs don't have semantics. Programs have only syntax, and syntax is insufficient for semantics. Every mind has semantics. Therefore programs are not minds.

This much of the argument is intended to show that artificial intelligence will never produce a machine with a mind by writing programs that manipulate symbols. The remainder of the argument addresses a different issue. Is the human brain running a program? In other words, is the computational theory of mind correct?[71] He begins with an axiom that is intended to express the basic modern scientific consensus about brains and minds:

(A4) Brains cause minds.

Searle claims that we can derive "immediately" and "trivially"[72] that:

(C2) Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains.
Brains must have something that causes a mind to exist. Science has yet to determine exactly what it is, but it must exist, because minds exist. Searle calls it "causal powers". "Causal powers" is whatever the brain uses to create a mind. If anything else can cause a mind to exist, it must have "equivalent causal powers". "Equivalent causal powers" is whatever *else* that could be used to make a mind.

And from this he derives the further conclusions:

(C3) Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program.
This follows from C1 and C2: Since no program can produce a mind, and "equivalent causal powers" produce minds, it follows that programs do not have "equivalent causal powers."

(C4) The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program.
Since programs do not have "equivalent causal powers", "equivalent causal powers" produce minds, and brains produce minds, it follows that brains do not use programs to produce minds.

Notes

1. ^ **a b** Searle 1980
2. ^ **a b** (Harnad 2001, p. 1) Harnad edited *BBS* during the years which saw the introduction and popularisation of the Chinese Room argument.
3. ^ **a b** Harnad 2001, p. 2

4. ^ In Akman's review of *Mind Design II* (http://www.google.com/search?client=safari&rls=en&q=cogprints.org/539/0/md2.ps&ie=UTF-8&oe=UTF-8)
5. ^ Harnad (2005) holds that the Searle's argument is against the thesis that "has since come to be called 'computationalism,' according to which cognition is just computation, hence mental states are just computational states". Cole (2004) agrees that "the argument also has broad implications for functionalist and computational theories of meaning and of mind".
6. ^ See the "Systems reply" below.
7. ^ See the "Other minds reply" below.
8. ^ The relationship between Searle's argument and consciousness is detailed in Chalmers 1996
9. ^ This version is from Searle (1999), and is also quoted in Dennett 1991, p. 435. Searle's original formulation was "The appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states." (Searle 1980, p. 1). Strong AI is defined similarly by Russell & Norvig (2003, p. 947): "The assertion that machines could possibly act intelligently (or, perhaps better, act as if they were intelligent) is called the 'weak AI' hypothesis by philosophers, and the assertion that machines that do so are actually thinking (as opposed to simulating thinking) is called the 'strong AI' hypothesis."
10. ^ Searle 2008
11. ^ Quoted in Russell & Norvig 2003, p. 21. Simon, together with Allen Newell and Cliff Shaw, had just completed the first "AI" program, the Logic Theorist.
12. ^ Quoted in Crevier 1993, p. 46 and Russell & Norvig 2003, p. 17.
13. ^ Haugeland 1986, p. 2. (Italics his)
14. ^ "Partisans of strong AI," Searle writes, "claim that in this question and answer sequence the machine is not only simulating a human ability but also (1) that the machine can literally be said to *understand* the story and provide the answers to questions, and (2) that what the machine and its program do *explains* the human ability to understand the story and answer questions about it." (Searle 1980, p. 2)
15. ^ Searle believes that "strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter." (Searle 1980, p. 13) He writes elsewhere, "I thought the whole idea of strong AI was that we don't need to know how the brain works to know how the mind works." (Searle 1980, p. 8) This position owes its phrasing to Harnad (2001).
16. ^ "One of the points at issue," writes Searle, "is the adequacy of the Turing test." (Searle 1980, p. 6)
17. ^ Harnad 2001, p. 3 (Italics his)
18. ^ Computationalism is associated with Jerry Fodor and Hilary Putnam. (Horst 2005, p. 1) Harnad (2001) also cites Allen Newell and Zenon Pylyshyn. Pinker (1997) also advocates a version of computationalism.
19. ^ Harnad 2001, pp. 3-5
20. ^ ***a b c*** Searle 1980, p. 13
21. ^ Searle 1990, p. 29
22. ^ Hauser 2006, p. 8
23. ^ Chalmers 1996, p. 322, quoted in Larry Hauser's annotated bibliography (http://host.uniroma3.it/progetti/kant/field/chinesebiblio.html).
24. ^ Russell & Norvig 2003, p. 947
25. ^ (Kurzweil 2005, p. 260) or see Advanced Human Intelligence (http://crnano.typepad.com/crnblog/2005/08/advanced_human_.html)
26. ^ Cole (2004, pp. 5-6) combines the middle two categories.
27. ^ Searle 1980, pp. 5-6, Cole 2004, pp. 6-7, Hauser 2006, pp. 2-3, Russell & Norvig 2003, p. 959, Dennett 1991, p. 439, Hearn 2007, p. 44, Crevier 1993, p. 269. This position is held by (according to Cole (2004, p. 6)) Ned Block, Jack Copeland, Daniel Dennett, Jerry Fodor, John Haugeland, Ray Kurzweil, and Georges Rey, among others.
28. ^ ***a b*** Searle 1980, p. 6
29. ^ Cole (2004, pp. 7-9) ascribes this position to Marvin Minsky, Tim Maudlin, David Chalmers and David Cole.
30. ^ This is the point of the universal Turing machine and the Church-Turing thesis: what makes a system Turing complete is its ability to do a step-by-step simulation of any other machine.
31. ^ The terminology "implementation independent" is due to Harnad (2001, p. 4).
32. ^ Cole 2004, p. 8
33. ^ Searle 1980, p. 12
34. ^ Hearn 2007, p. 47
35. ^ Cole (2004, p. 21) writes "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."

36. ^ Searle 1980, p. 7, Cole 2004, pp. 9-11, Hauser 2006, p. 3, Hearn 2007, p. 44. Cole (2004, p. 9) ascribes this position to Margaret Boden, Tim Crane, Daniel Dennett, Jerry Fodor, Stevan Harnad, Hans Moravec and Georges Rey
37. ^ Quoted in Crevier 1993, p. 272. Cole (2004, p. 18) calls this the "externalist" account of meaning.
38. ^ ***a b*** Searle 1980, p. 7
39. ^ Hauser 2006, p. 11, Cole 2004, p. 19. This argument is supported by Daniel Dennett and others.
40. ^ Searle distinguishes between "intrinsic" intentionality and "derived" intentionality. "Intrinsic" intentionality is the kind that involves "conscious understanding" like you would have in a human mind. Daniel Dennett doesn't agree that there is a distinction. Cole (2004, p. 19) writes "derived intentionality is all there is, according to Dennett."
41. ^ Cole 2004, p. 18 (where he calls this the "internalist" approach to meaning.) Proponents of this position include Roger Schank, Doug Lenat, Marvin Minsky and (with reservations) Daniel Dennett, who writes "The fact is that any program [that passed a Turing test] would have to be an extraordinarily supple, sophisticated, and multilayered system, brimming with 'world knowledge' and meta-knowledge and meta-meta-knowledge." (Dennett 1997, p. 438)
42. ^ Dreyfus 1979. See "the epistemological assumption".
43. ^ Searle 1984. He also writes "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, p. 45 quoted in Cole 2004, p. 16.
44. ^ Searle 1980, pp. 7-8, Cole 2004, pp. 12-13, Hauser 2006, pp. 3-4, Churchland & Churchland 1990. Cole (2004, p. 12) ascribes this position to Paul Churchland, Patricia Churchland and Ray Kurzweil.
45. ^ Cole 2004, p. 4, Hauser 2006, p. 11. Early versions of this argument were put forward in 1974 by Lawrence Davis and in 1978 by Ned Block. Block's version used walky talkies and was called the "Chinese Gym". Churchland & Churchland (1990) described this scenario as well.
46. ^ Russell Norvig, pp. 956-8, Cole 2004, p. 20, Moravec 1988, p. ? CHECK, Kurzweil 2005, p. 262 CHECK, Crevier 1993, pp. 271 and 279 CHECK. An early version of this argument was put forward by Clark Glymour in the mid-70s and was touched on by Zenon Pylyshyn in 1980. Moravec (1988) presented a vivid version of it, and it is now associated with Ray Kurzweil's version of transhumanism.
47. ^ Searle predicts that, while going through the brain prosthesis, "you find, to your total amazement, that you are indeed losing control of you external behavior. You find, for example, that when doctors test your vision, you hear them say 'We are holding up a red object in front of you; pleas tell us what you see.' You want to cry out 'I can't see anything. I'm going totally blind.' But you hear your voice saying in a way that is completely out your control, 'I see a read object in front of me.' ... [Y]our conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same." Searle 1992 quoted in Russell & Norvig 2003, p. 957.
48. ^ Cole (2004, pp. 12 & 17) ascribes this position to Andy Clark and Ray Kurzweil. Hauser (2006, p. 7) associates this position with Paul and Patricia Churchland.
49. ^ Searle 1980, pp. 8-9, Hauser 2006, p. 11,
50. ^ Searle (1980, p. 7) writes that the robot reply "tacitly concedes that cognition is not solely a matter of formal symbol manipulation." Harnad (2001, p. 14) makes the same point, writing: "Now just as it is no refutation (but rather an affirmation) of the CRA to deny that [the Turing test] is a strong enough test, or to deny that a computer could ever pass it, it is merely special pleading to try to save computationalism by stipulating ad hoc (in the face of the CRA) that implementational details do matter after all, and that the computer's is the 'right' kind of implementation, whereas Searle's is the 'wrong' kind."
51. ^ Searle 1980, p. 8
52. ^ Searle 1980, p. 3
53. ^ That is, any program running on a machine with a finite amount memory.
54. ^ Cole 2004, pp. 14-15, Crevier 1993, pp. 269-270, Pinker, p. 95. Cole (2004, p. 14) ascribes this "speed" position to Daniel Dennett, Tim Maudlin, David Chalmers, Steven Pinker, Paul Churchland, Patricia Churchland and others. Dennett (1991, p. 438) points out the complexity of world knowledge.
55. ^ Crevier 1993, p. 269
56. ^ Churchland & Churchland 1990, Cole 2004, p. 12, Crevier 1993, p. 270, Hearn 2007, pp. 45-46, Pinker 1997, p. 94
57. ^ (Dennett 1991, p. 438)
58. ^ Harnad 2001, p. 7. Critics of the "phase transition" form of this argument include Harnad, Tim Maudlin, Daniel Dennett and Cole (2004, p. 14). This "phase transition" idea is a version of strong emergentism (what Daniel Dennett derides as "Woo woo West Coast emergence" (Crevier 1993, p. 275)). Harnad accuses Churchland and Patricia Churchland of espousing strong emergentism and Kurzweil (2005) seems to also agree with strong emergentism.

59. ^ Searle 1980, Cole 2004, p. 13, Hauser 2006, pp. 4-5, Nilsson 1984. Turing (1950, pp. 11-12) makes this reply to what he calls "The Argument from Consciousness." Cole (2004, pp. 12-13) ascribes this position to Daniel Dennett, Ray Kurzweil and Hans Moravec.
60. ^ Nilsson 1984
61. ^ Turing 1950, p. 11
62. ^ One of Turing's motivations for devising the Turing test is to avoid precisely the kind of philosophical problems that Searle is interested in. He writes "I do not wish to give the impression that I think there is no mystery ... [but] I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper." (Turing 1950, p. 12) Although Turing is discussing consciousness (not the mind or understanding or intentionality), Norvig & Russell (2003, p. 952-953) argue that Turing's comments apply the Chinese room.
63. ^ Cole 2004, p. 22, Crevier 1993, p. 271, Harnad 2004, p. 4
64. ^ Searle 1980, p. 10
65. ^ Quoted in Cole 2004, p. 13.
66. ^ Dennett 1991, pp. 437 & 440
67. ^ Dennett 1991, p. 438
68. ^ Searle 1984
69. ^ Searle 1984, Searle 1990. The wording of each axiom and conclusion if from Searle (1990). This version is based on Hauser 2006, p. 5. (A1-3) and (C1) are described as 1,2,3 and 4 in Cole 2004, p. 5.
70. ^ Churchland & Churchland (1990, p. 34) explain that the Chinese Room argument is intended to "shore up axiom 3".
71. ^ Harnad (2001) argues that Searle's primary target is computationalism.
72. ^ Searle 1990

References

- Block, Ned (1981), "Psychologism and (http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Psychologism.htm), *The Philosophical Review* **90**: 5–43, doi:10.2307/2184371 (http://dx.doi.org/10.2307/2184371), http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Psychologism.htm.
- Chalmers, David (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press.
- Churchland, Paul; Churchland, Patricia (January 1990), "Could a machine think?", *Scientific American* **262**: 32–39
- Cole, David (Fall 2004), "The Chinese Room (http://plato.stanford.edu/archives/fall2004/entries/chinese-room/), in Zalta, Edward N., *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/fall2004/entries/chinese-room/. *Page numbers above refer to a standard pdf print of the article.*
- Crevier, Daniel (1993), *AI: The Tumultuous Search for Artificial Intelligence*, New York, NY: BasicBooks, ISBN 0-465-02997-3.
- Dennett, Daniel (1991). *Consciousness Explained*. The Penguin Press. ISBN 0-7139-9037-6..
- Fearn, Nicholas (2007), *The Latest Answers to the Oldest Questions: A Philosophical Adventure with the World's Greatest Thinkers*, New York: Grove Press
- Harnad, Stevan (2001), "What's Wrong and Right About Searle's Chinese Room (http://cogprints.org/4023/), in M.; Preston, J., *Essays on Searle's Chinese Room Argument*, Oxford University Press, http://cogprints.org/4023/. *Page numbers above refer to a standard pdf print of the article.*
- Harnad, Stevan (2005), "Searle's Chinese Room (http://eprints.ecs.soton.ac.uk/10424/01/chineseroom.html), *Encyclopedia of Philosophy*, Macmillan, http://eprints.ecs.soton.ac.uk/10424/01/chineseroom.html. *Page numbers above refer to a standard pdf print of the article.*
- Hauser, Larry (1997), "Searle's Chinese Box: Debunking the Chinese Room Argument", *Minds and Machines* **7**: 199–226, doi:10.1023/A:1008255830248 (http://dx.doi.org/10.1023/A:1008255830248). *Page numbers above refer to a standard pdf print of the article.*

- Hauser, Larry (2006), "Searle's Chinese Room" (http://www.iep.utm.edu/c/chineser.htm), *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/c/chineser.htm. *Page numbers above refer to a standard pdf print of the article.*
- Kurzweil, Ray (2005), *The Singularity is Near*, Viking Press
- Moravec, Hans (1988), *Mind Children*, Harvard University Press
- Nilsson, Nils (1984), *A Short Rebuttal to Searle* (http://ai.stanford.edu/~nilsson/OnlinePubs-Nils/General%20Essays/OtherEssays-Nils/searle.pdf), http://ai.stanford.edu/%7Enilsson/OnlinePubs-Nils/General%20Essays/OtherEssays-Nils/searle.pdf
- Russell, Stuart J.; Norvig, Peter (2003), *Artificial Intelligence: A Modern Approach* (http://aima.cs.berkeley.edu/) (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2, http://aima.cs.berkeley.edu/.
- Pinker, Steven (1997), *How the Mind Works*, New York, NY: W. W. Norton & Company, Inc., ISBN 0-393-31848-6
- Searle, John (1980), "Minds, Brains and (http://web.archive.org/web/20071210043312/http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html), *Behavioral and Brain Sciences* **3** (3): 417–457, http://web.archive.org/web/20071210043312/http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html, retrieved May 13, 2009. *Page numbers above refer to a standard pdf print of the article. See also Searle's original draft (http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html).*
- Searle, John (1983), "Can Computers Think?", in Chalmers, David, *Philosophy of Mind: Classical and Contemporary Readings*, Oxford: Oxford University Press, pp. 669–675, ISBN 0-19-514581-X.
- Searle, John (1984), *Minds, Brains and Science: The 1984 Reith Lectures*, Harvard University Press, ISBN 0-67457631-4 paperback: ISBN 0-67457633-0.
- Searle, John (January 1990), "Is the Brain's Mind a Computer Program?", *Scientific American* **262**: 26–31.
- Searle, John (1992), *The Rediscovery of the Mind*, Cambridge, Massachusetts: M.I.T. Press.
- Searle, John (1999), *Mind, language and society*, New York, NY: Basic Books, ISBN 0465045219, OCLC 43689264 231867665 43689264 (http://www.worldcat.org/oclc/231867665)
- Turing, Alan (October 1950), "Computing Machinery and (http://loebner.net/Prizef/TuringArticle.html), *Mind* **LIX** (236): 433–460, doi:10.1093/mind/LIX.236.433 (http://dx.doi.org/10.1093/mind/LIX.236.433), ISSN 0026-4423 (http://www.worldcat.org/issn/0026-4423), http://loebner.net/Prizef/TuringArticle.html, retrieved 2008-08-18. *Page numbers above refer to a standard pdf print of the article.*

Further reading

- Wikibooks: Consciousness Studies
- The Chinese Room Argument (http://globetrotter.berkeley.edu/people/Searle/searle-con4.html), part 4 of the September 2, 1999 interview with Searle Philosophy and the Habits of Critical Thinking (http://globetrotter.berkeley.edu/people/Searle/searle-con0.html) in the Conversations With History series
- Understanding the Chinese Room (http://www.zompist.com/searle.html), Mark Rosenfelder
- A Refutation of John Searle's "Chinese Room Argument" (http://www.anti-state.com/article.php?article_id=247), by Bob Murphy
- Kugel, P. (2004). "The Chinese room is a trick". *Behavioral and Brain Sciences* **27**. doi:10.1017/S0140525X04210044 (http://dx.doi.org/10.1017/S0140525X04210044). , PDF at author's homepage (http://www.cs.bc.edu/~kugel/Publications/Searle%206.pdf), critical paper based on the assumption that the CR cannot use its inputs (which are in Chinese) to change its program (which is in English).
- Wolfram Schmied (2004). "Demolishing Searle's Chinese Room". *arXiv:cs.AI/0403009 (http://www.arxiv.org/abs/cs.AI/0403009)* [cs.AI].

- John Preston and Mark Bishop, "Views into the Chinese Room", Oxford University Press, 2002. Includes chapters by John Searle, Roger Penrose, Stevan Harnad and Kevin Warwick.
- Margaret Boden, "Escaping from the Chinese room", Cognitive Science Research Papers No. CSRP 092, University of Sussex, School of Cognitive Sciences, 1987, OCLC 19297071, online PDF (http://doi.library.cmu.edu/10.1184/OCLC/19297071), "an excerpt from a chapter" in the then unpublished "Computer Models of Mind: : Computational Approaches in Theoretical Psychology", ISBN 052124868X (1988); reprinted in Boden (ed.) "The Philosophy of Artificial Intelligence" ISBN 0198248547 (1989) and ISBN 0198248555 (1990); Boden "Artificial Intelligence in Psychology: Interdisciplinary Essays" ISBN 0262022850, MIT Press, 1989, chapter 6; reprinted in Heil, pp. 253–266 (1988) (possibly abridged); J. Heil (ed.) "Philosophy of Mind: A Guide and Anthology", Oxford University Press, 2004, pages 253-266 (same version as in "Artificial Intelligence in Psychology")