Chinese room

The **Chinese room argument** holds that a digital computer executing a program cannot be shown to have a "mind", "understanding" or "consciousness",[a] regardless of how intelligently or human-like the program may make the computer behave. The argument was first presented by philosopher John Searle in his paper, "Minds, Brains, and Programs", published in *Behavioral and Brain Sciences* in 1980. It has been widely discussed in the years since.[1] The centerpiece of the argument is a thought experiment known as the *Chinese room*.[2]

The argument is directed against the philosophical positions of functionalism and computationalism,[3] which hold that the mind may be viewed as an information-processing system operating on formal symbols, and that simulation of a given mental state is sufficient for its presence. Specifically, the argument is intended to refute a position Searle 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."[b]

Although it was originally presented in reaction to the statements of artificial intelligence (AI) researchers, it is not an argument against the goals of mainstream AI research, because it does not limit the amount of intelligence a machine can display.[4] The argument applies only to digital computers running programs and does not apply to machines in general.[5]

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 live Chinese speaker. To all of the questions that the person asks, it makes appropriate responses, such that any Chinese speaker would be convinced that they are talking to another Chinese-speaking human being.

The question Searle wants to answer is this: does the machine *literally* "understand" Chinese? Or is it merely *simulating* the ability to understand Chinese?[6][c] Searle calls the first position "strong AI" and the latter "weak AI".[d]

John Searle in December 2005

Searle then supposes that he is in a closed room and has a book with
an English version of the computer program, along with sufficient papers, pencils, erasers, and filing cabinets. Searle could receive Chinese characters through a slot in the door, process them according to the program's instructions, and produce Chinese characters as output. If the computer had passed the Turing test this way, it follows, says Searle, that he would do so as well, simply by running the program manually.

Searle asserts that there is no essential difference between the roles of the computer and himself in the experiment. Each simply follows a program, step-by-step, producing a behavior which is then interpreted by the user as demonstrating intelligent conversation. However, Searle himself would not be able to understand

the conversation. ("I don't speak a word of Chinese,"[9] he points out.) Therefore, he argues, it follows that the computer would not be able to understand the conversation either.

Searle argues that, without "understanding" (or "intentionality"), we cannot describe what the machine is doing as "thinking" and, since it does not think, it does not have a "mind" in anything like the normal sense of the word. Therefore, he concludes that the "strong AI" hypothesis is false.

History

Gottfried Leibniz made a similar argument in 1714 against mechanism (the position that the mind is a machine and nothing more). Leibniz used the thought experiment of expanding the brain until it was the size of a mill.[10] Leibniz found it difficult to imagine that a "mind" capable of "perception" could be constructed using only mechanical processes.[e] In the 1961 short story "The Game" by Anatoly Dneprov, a stadium of people act as switches and memory cells implementing a program to translate a sentence of Portuguese, a language that none of them knows.[11] In 1974, Lawrence Davis imagined duplicating the brain using telephone lines and offices staffed by people, and in 1978 Ned Block envisioned the entire population of China involved in such a brain simulation. This thought experiment is called the China brain, also the "Chinese Nation" or the "Chinese Gym".[12]

The Chinese Room Argument was introduced in Searle's 1980 paper "Minds, Brains, and Programs", published in *Behavioral and Brain Sciences*.[13] It eventually became the journal's "most influential target article",[1] generating an enormous number of commentaries and responses in the ensuing decades, and Searle has continued to defend and refine the argument in many papers, popular articles and books. David Cole writes that "the Chinese Room argument has probably been the most widely discussed philosophical argument in cognitive science to appear in the past 25 years".[14]

Most of the discussion consists of attempts to refute it. "The overwhelming majority", notes *BBS* editor Stevan Harnad,[f] "still think that the Chinese Room Argument is dead wrong".[15] The sheer volume of the literature that has grown up around it inspired Pat Hayes to comment 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".[16]

Searle's argument has become "something of a classic in cognitive science", according to Harnad.[15] Varol Akman agrees, and has described the original paper as "an exemplar of philosophical clarity and purity".[17]

Philosophy

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

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.[b]

The definition depends 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."[7]

The claim is implicit in some of the statements of early AI researchers and analysts. For example, in 1955, AI founder Herbert A. Simon declared that "there are now in the world machines that think, that learn and create"[23](Simon, together with Allen Newell and Cliff Shaw, had just completed the first "AI" program, the Logic Theorist), and claimed that they had "solved the venerable mind–body problem, explaining how a system composed of matter can have the properties of mind."[24] 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*."[25]

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

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

Strong AI as computationalism or functionalism

In more recent presentations of the Chinese room argument, Searle has identified "strong AI" as "computer functionalism" (a term he attributes to Daniel Dennett).[3][30] Functionalism is a position in modern philosophy of mind that holds that we can define mental phenomena (such as beliefs, desires, and perceptions) by describing their functions in relation to each other and to the outside world. Because a computer program can accurately represent functional relationships as relationships between symbols, a computer can have mental phenomena if it runs the right program, according to functionalism.

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."[31] Computationalism[j] 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:[34]

- 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.

Strong AI vs. biological naturalism

Searle holds a philosophical position he calls "biological naturalism": that consciousness[a] and understanding require specific biological machinery that are found in brains. He writes "brains cause minds"[5] and that "actual human mental phenomena [are] dependent on actual physical–chemical properties of actual human

brains".[35] Searle argues that this machinery (known to neuroscience as the "neural correlates of consciousness") must have some causal powers that permit the human experience of consciousness.[36] Searle's belief in the existence of these powers has been criticized.[k]

Searle does not disagree with the notion that machines can have consciousness and understanding, because, as he writes, "we are precisely such machines".[5] Searle holds that the brain is, in fact, a machine, but that the brain gives rise to consciousness and understanding using machinery that is non-computational. If neuroscience is able to isolate the mechanical process that gives rise to consciousness, then Searle grants that it may be possible to create machines that have consciousness and understanding. However, without the specific machinery required, Searle does not believe that consciousness can occur.

Biological naturalism implies that one cannot determine if the experience of consciousness is occurring merely by examining how a system functions, because the specific machinery of the brain is essential. Thus, biological naturalism is directly opposed to both behaviorism and functionalism (including "computer functionalism" or "strong AI").[37] Biological naturalism is similar to identity theory (the position that mental states are "identical to" or "composed of" neurological events); however, Searle has specific technical objections to identity theory.[38][l] Searle's biological naturalism and strong AI are both opposed to Cartesian dualism,[37] the classical idea that the brain and mind are made of different "substances". Indeed, Searle accuses strong AI of dualism, writing that "strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter."[26]

Consciousness

Searle's original presentation emphasized "understanding"—that is, mental states with what philosophers call "intentionality"—and did not directly address other closely related ideas such as "consciousness". However, in more recent presentations Searle has included consciousness as the real target of the argument.[3]

> Computational models of consciousness are not sufficient by themselves for consciousness. The computational model for consciousness stands to consciousness in the same way the computational model of anything stands to the domain being modelled. Nobody supposes that the computational model of rainstorms in London will leave us all wet. But they make the mistake of supposing that the computational model of consciousness is somehow conscious. It is the same mistake in both cases.[39]
>
> — John R. Searle, Consciousness and Language, p. 16

David Chalmers writes "it is fairly clear that consciousness is at the root of the matter" of the Chinese room.[40]

Colin McGinn argues that the Chinese room provides strong evidence that the hard problem of consciousness is fundamentally insoluble. The argument, to be clear, is not about whether a machine can be conscious, but about whether it (or anything else for that matter) can be shown to be conscious. It is plain that any other method of probing the occupant of a Chinese room has the same difficulties in principle as exchanging questions and answers in Chinese. It is simply not possible to divine whether a conscious agency or some clever simulation inhabits the room.[41]

Searle argues that this is only true for an observer *outside* of the room. The whole point of the thought experiment is to put someone *inside* the room, where they can directly observe the operations of consciousness. Searle claims that from his vantage point within the room there is nothing he can see that could imaginably give rise to consciousness, other than himself, and clearly he does not have a mind that can speak Chinese.

Applied ethics

Patrick Hew used the Chinese Room argument to deduce requirements from military command and control systems if they are to preserve a commander's moral agency. He drew an analogy between a commander in their command center and the person in the Chinese Room, and analyzed it under a reading of Aristotle's notions of "compulsory" and "ignorance". Information could be "down converted" from meaning to symbols, and manipulated symbolically, but moral agency could be undermined if there was inadequate 'up conversion' into meaning. Hew cited examples from the USS *Vincennes* incident.[42]

Sitting in the combat information center aboard a warship – proposed as a real-life analog to the Chinese Room

Computer science

The Chinese room argument is primarily an argument in the philosophy of mind, and both major computer scientists and artificial intelligence researchers consider it irrelevant to their fields.[4] However, several concepts developed by computer scientists are essential to understanding the argument, including symbol processing, Turing machines, Turing completeness, and the Turing test.

Strong AI vs. AI research

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."[4] The primary mission of artificial intelligence research is only to create useful systems that *act* intelligently, and it does not matter if the intelligence is "merely" a simulation.

Searle does not disagree that AI research can create machines that are capable of highly intelligent behavior. The Chinese room argument leaves open the possibility that a digital machine could be built that *acts* more intelligently than a person, but does not have a mind or intentionality in the same way that brains do.

Searle's "strong AI" should not be confused with "strong AI" as defined by Ray Kurzweil and other futurists,[43] who use the term to describe machine intelligence that rivals or exceeds human intelligence. Kurzweil is concerned primarily with the *amount* of intelligence displayed by the machine, whereas Searle's argument sets no limit on this. Searle argues that even a super-intelligent machine would not necessarily have a mind and consciousness.

Turing test

The Chinese room implements a version of the Turing test.[45] Alan Turing introduced the test in 1950 to help answer the question "can machines think?" In the standard version, a human judge engages in a natural language conversation with a human and a machine designed to generate performance indistinguishable from that of a human being. All participants are separated from one another. If the judge cannot reliably tell the machine from the human, the machine is said to have passed the test.

Turing then considered each possible objection to the proposal "machines can think", and found that there are simple, obvious answers if the question is de-mystified in this way. He did not, however, intend for the test to measure for the presence of "consciousness" or "understanding". He did not believe this was relevant to the

issues that he was addressing. He wrote:

> I do not wish to give the impression that I think there is
> no mystery about consciousness. There is, for instance,
> something of a paradox connected with any attempt to
> localise it. 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.[45]

To Searle, as a philosopher investigating in the nature of mind and
consciousness, these are the relevant mysteries. The Chinese room is
designed to show that the Turing test is insufficient to detect the
presence of consciousness, even if the room can behave or function as
a conscious mind would.

Symbol processing

The Chinese room (and all modern computers) manipulate physical
objects in order to carry out calculations and do simulations. AI
researchers Allen Newell and Herbert A. Simon called this kind of
machine a physical symbol system. It is also equivalent to the formal
systems used in the field of mathematical logic.

The "standard interpretation" of the Turing Test, in which player C, the interrogator, is given the task of trying to determine which player – A or B – is a computer and which is a human. The interrogator is limited to using the responses to written questions to make the determination. Image adapted from Saygin, et al. 2000.[44]
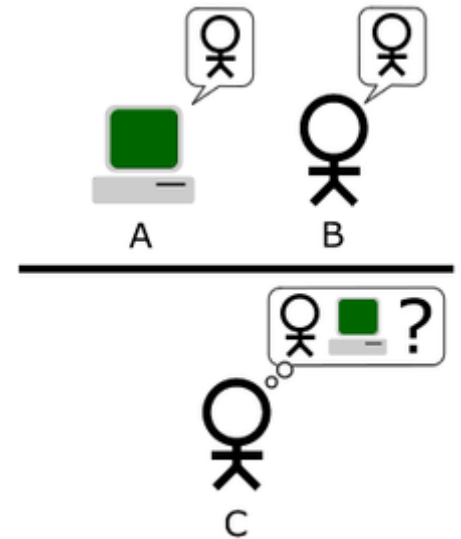
Searle emphasizes the fact that this kind of symbol manipulation is
syntactic (borrowing a term from the study of grammar). The
computer manipulates the symbols using a form of syntax rules, without any knowledge of the symbol's
semantics (that is, their meaning).

Newell and Simon had conjectured that a physical symbol system (such as a digital computer) had all the
necessary machinery for "general intelligent action", or, as it is known today, artificial general intelligence.
They framed this as a philosophical position, the physical symbol system hypothesis: "A physical symbol
system has the necessary and sufficient means for general intelligent action."[46][47] The Chinese room
argument does not refute this, because it is framed in terms of "intelligent action", i.e. the external behavior of
the machine, rather than the presence or absence of understanding, consciousness and mind.

Chinese room and Turing completeness

The Chinese room has a design analogous to that of a modern computer. It has a Von Neumann architecture,
which consists of a program (the book of instructions), some memory (the papers and file cabinets), a CPU
which follows the instructions (the man), and a means to write symbols in memory (the pencil and eraser). A
machine with this design is known in theoretical computer science as "Turing complete", because it has the
necessary machinery to carry out any computation that a Turing machine can do, and therefore it is capable of
doing a step-by-step simulation of any other digital machine, given enough memory and time. Alan Turing
writes, "all digital computers are in a sense equivalent."[48] The widely accepted Church–Turing thesis holds
that any function computable by an effective procedure is computable by a Turing machine.

The Turing completeness of the Chinese room implies that it can do whatever any other digital computer can
do (albeit much, much more slowly). Thus, if the Chinese room does not or can not contain a Chinese-
speaking mind, then no other digital computer can contain a mind. Some replies to Searle begin by arguing

that the room, as described, cannot have a Chinese-speaking mind. Arguments of this form, according to Stevan Harnad, are "no refutation (but rather an affirmation)"[49] of the Chinese room argument, because these arguments actually imply that *no* digital computers can have a mind.[28]

There are some critics, such as Hanoch Ben-Yami, who argue that the Chinese room cannot simulate all the abilities of a digital computer, such as being able to determine the current time.[50]

Complete argument

Searle has produced a more formal version of the argument of which the Chinese Room forms a part. He presented the first version in 1984. The version given below is from 1990.[51][m] The only 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.[n]

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 thought experiment 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 no programs are minds.

This much of the argument is intended to show that artificial intelligence can 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?[g] 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"[52] 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.

Replies

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

- 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
- Those which contend that Searle's argument is misleading
- Those which argue that the argument makes false assumptions about subjective conscious experience and therefore proves nothing

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

Systems and virtual mind replies: finding the mind

These 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. All of the replies that identify the mind in the room are versions of "the system reply".

System reply

The basic version argues that it is the "whole system" that understands Chinese.[57][p] While the man understands only English, when he is combined with the program, scratch paper, pencils and file cabinets, they form a system that can understand Chinese. "Here, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part" Searle explains.[29] The fact that man does not understand Chinese is irrelevant, because it is only the system as a whole that matters.

Searle notes that (in this simple version of the reply) the "system" is nothing more than a collection of ordinary physical objects; it grants the power of understanding and consciousness to "the conjunction of that person and bits of paper"[29] without making any effort to explain how this pile of objects has become a conscious, thinking being. Searle argues that no reasonable person should be satisfied with the reply, unless they are "under the grip of an ideology;"[29] In order for this reply to be remotely plausible, one must take it for granted that consciousness can be the product of an information processing "system", and does not require anything resembling the actual biology of the brain.

Searle then responds by simplifying this list of physical objects: he asks what happens if the man memorizes the rules and keeps track of everything in his head? Then the whole system consists of just one object: the man himself. Searle argues that if the man doesn't understand Chinese then the system doesn't understand Chinese either because now "the system" and "the man" both describe exactly the same object.[29]

Critics of Searle's response argue that the program has allowed the man to have two minds in one head. If we assume a "mind" is a form of information processing, then the theory of computation can account for two computations occurring at once, namely (1) the computation for universal programmability (which is the function instantiated by the person and note-taking materials *independently* from any particular program contents) and (2) the computation of the Turing machine that is described by the program (which is instantiated by everything *including* the specific program).[59] The theory of computation thus formally explains the open possibility that the second computation in the Chinese Room could entail a human-equivalent semantic understanding of the Chinese inputs. The focus belongs on the program's Turing machine rather than on the person's.[60] However, from Searle's perspective, this argument is circular. The question at issue is whether consciousness is a form of information processing, and this reply requires that we make that assumption.

More sophisticated versions of the systems reply try to identify more precisely what "the system" is and they differ in exactly how they describe it. According to these replies, the "mind that speaks Chinese" could be such things as: the "software", a "program", a "running program", a simulation of the "neural correlates of consciousness", the "functional system", a "simulated mind", an "emergent property", or "a virtual mind" (Marvin Minsky's version of the systems reply, described below).

Virtual mind reply

> The term "virtual" is used in computer science to describe an object that appears to exist "in" a computer (or computer network) only because software makes it appear to exist. The objects "inside" computers (including files, folders, and so on) are all "virtual", except for the computer's electronic components. Similarly, Minsky argues, a computer may contain a "mind" that is virtual in the same sense as virtual machines, virtual communities and virtual reality.[q]

> To clarify the distinction between the simple systems reply given above and virtual mind reply, David Cole notes that two simulations could be running on one system at the same time: one speaking Chinese and one speaking Korean. While there is only one system, there can be multiple "virtual minds," thus the "system" cannot be the "mind".[64]

Searle responds that such a mind is, at best, a simulation, and 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."[65] 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."[66] The question is, is the human mind like the pocket calculator,

essentially composed of information? Or is the mind like the rainstorm, something other than a computer, and not realizable in full by a computer simulation? (The issue of simulation is also discussed in the article synthetic intelligence.)

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.[r]

However, the thought experiment is not intended to be a *reductio ad absurdum*, but rather an example that requires explanation. Searle is not asserting that the situation is impossible, but rather that it is difficult or impossible to explain how this system can have subjective conscious experience.[68] The system reply succeeds in showing that it is *not impossible* but fails to show how the system would have consciousness; the replies, by themselves, 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 the system must understand Chinese."[29]

Robot and semantics replies: finding the meaning

As far as the person in the room is concerned, the symbols are just meaningless "squiggles." But if the Chinese room really "understands" what it is 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

> 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.[69][s] 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."[71][t]

> Searle's reply is to suppose that, unbeknownst to the individual in the Chinese room, some of the inputs 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."[73] (See Mary's room for a similar thought experiment.)

Derived meaning

> 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 Searle manipulates *are already meaningful*, they're just not meaningful to *him*.[74][u]

> Searle says 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, like a book, has no understanding of its own.[v]

Commonsense knowledge / contextualist reply

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.[72][w]

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.[77]

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."[78][x]

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 identify some special technology that would create conscious understanding 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

Suppose that the program simulated in fine detail the action of every neuron in the brain of a Chinese speaker.[80][y] 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 does not reproduce 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."[26] Moreover, he argues:

> [I]magine that instead of a monolingual man in a room shuffling symbols we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols, he looks up in the program, written in English, which valves he has to turn on and off. Each water connection corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings, that is after turning on all the right faucets, the Chinese answers pop out at the output end of the series of pipes. Now where is the understanding in this system? It takes Chinese as input, it simulates the formal structure of the synapses of the Chinese brain, and it gives Chinese as output. But the man certainly doesn't understand Chinese, and neither do the water pipes, and if we are tempted to adopt what I think is the absurd view that somehow the conjunction of man and water pipes understands, remember that in principle the man can internalize the formal structure of the water pipes and do all the "neuron firings" in his imagination.[82]

Two variations on the brain simulator reply are the China brain and the brain-replacement scenario.

China brain

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.[83][z] It is also obvious that this system would be functionally equivalent to a brain, so if consciousness is a function, this system would be conscious.

Brain replacement scenario

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.[85][aa] Searle predicts that, while going through the brain prosthesis, "you find, to your total amazement, that you are indeed losing control of your 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; please 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 outside of your control, 'I see a red object in front of me.' [...] [Y]our conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same."[87] (See Ship of Theseus for a similar thought experiment.)

Connectionist replies

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

Combination reply

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.[90]

Many mansions / wait till next year reply

Better technology in the future will allow computers to understand.[27][ac] Searle agrees that this is possible, but considers this point irrelevant. His argument is that a machine using a program to manipulate formally defined elements can not produce understanding. Searle's argument, if correct, rules out only this particular design. Searle agrees that there may be other designs that would cause a machine to have conscious understanding.

These arguments (and the robot or commonsense knowledge replies) identify some special technology that would help create conscious understanding in a machine. They may be interpreted in two ways: either they claim (1) this technology is required for consciousness, the Chinese room does not or cannot implement this

technology, and therefore the Chinese room cannot pass the Turing test or (even if it did) it would not have conscious understanding. Or they may be claiming that (2) it is easier to see that the Chinese room has a mind if we visualize this technology as being used to create it.

In the first case, where features like a robot body or a connectionist architecture are required, Searle claims that strong AI (as he understands it) has been abandoned.[ad] The Chinese room has all the elements of a Turing complete machine, and thus is capable of simulating any digital computation whatsoever. If Searle's room can't pass the Turing test then there is no other digital technology that could pass the Turing test. If Searle's room *could* pass the Turing test, but still does not have a mind, then 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 as "strong AI", proving his argument.

The brain arguments in particular deny strong AI if 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."[27] If computation does not provide an *explanation* of the human mind, then strong AI has failed, according to Searle.

Other critics hold that the room as Searle described it does, in fact, have a mind, however 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).

In fact, the room can just as easily be redesigned to *weaken* our intuitions. Ned Block's Blockhead argument[91] 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". At least in principle, any program can be rewritten (or "refactored") into this form, even a brain simulation.[ae] 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 one's conscious experience can be captured in a single large number, yet this is exactly what "strong AI" claims. On the other hand, such a lookup table would be ridiculously large (to the point of being physically impossible), and the states could therefore be *extremely* specific.

Searle argues that however the program is written or however the machine is connected to the world, the mind is being *simulated* by a simple step-by-step digital machine (or machines). These machines are always 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 writes: "I can have any formal program you like, but I still understand nothing."[9]

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

Several 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."[92] Daniel Dennett describes the Chinese room argument as a misleading "intuition pump"[93] 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."[93]

Some of the arguments above also function as appeals to intuition, especially those that 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. Several of the replies above also address the specific issue of complexity. The connectionist reply emphasizes that a working artificial intelligence 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.[76]

Speed and complexity replies

The speed at which human brains process information is (by some estimates) 100 billion operations per second.[94] 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.[95][af]

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

"Consider a dark room containing a man holding a bar magnet or charged object. If the man pumps the magnet up and down, then, according to Maxwell's theory of artificial luminance (AL), it will initiate a spreading circle of electromagnetic waves and will thus be luminous. But as all of us who have toyed with magnets or charged balls well know, their forces (or any other forces for that matter), even when set in motion produce no luminance at all. It is inconceivable that you might constitute real luminance just by moving forces around!"[84] The problem is that he would have to wave the magnet up and down something like 450 trillion times per second in order to see anything.[97]

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.')"[98][ag]

Searle argues that his critics are also relying on intuitions, however his opponents' intuitions have no empirical basis. He writes that, in order to consider the "system reply" as remotely plausible, a person must be "under the grip of an ideology".[29] The system reply only makes sense (to Searle) if one assumes that any "system" can have consciousness, just by virtue of being a system with the right behavior and functional parts. This assumption, he argues, is not tenable given our experience of consciousness.

Other minds and zombies: meaninglessness

Several replies argue that Searle's argument is irrelevant because his assumptions about the mind and consciousness are faulty. Searle believes that human beings directly experience their consciousness, intentionality and the nature of the mind every day, and that this experience of consciousness is not open to question. He writes that we must "presuppose the reality and knowability of the mental."[101] These replies question whether Searle is justified in using his own experience of consciousness to determine that it is more

than mechanical symbol processing. In particular, the other minds reply argues that we cannot use our experience of consciousness to answer questions about other minds (even the mind of a computer), and the epiphenomena reply argues that Searle's consciousness does not "exist" in the sense that Searle thinks it does.

Other minds reply

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.[102][ah]

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."[104]

Alan Turing anticipated Searle's line of argument (which he called "The Argument from Consciousness") in 1950 and makes the other minds reply.[105] He 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."[106] 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.[ai]

Eliminative Materialism reply

Several philosophers argue that consciousness, as Searle describes it, does not exist. This position is sometimes referred to as eliminative materialism: the view that consciousness is a property that can be reduced to a strictly mechanical description, and that our experience of consciousness is, as Daniel Dennett describes it, a "user illusion".[109] Other mental properties, such as original intentionality (also called "meaning", "content", and "semantic character"), is also commonly regarded as something special about beliefs and other propositional attitudes. Eliminative materialism maintains that propositional attitudes such as beliefs and desires, among other intentional mental states that have content, do not exist. If eliminative materialism is the correct scientific account of human cognition then the assumption of the Chinese room argument that "minds have mental contents (semantics)" must be rejected.[110]

Stuart Russell and Peter Norvig argue that, if we accept Searle's description of intentionality, consciousness and the mind, we are forced to accept that consciousness is epiphenomenal: that it "casts no shadow", that it is undetectable in the outside world. They argue that Searle must be mistaken about the "knowability of the mental", and in his belief that there are "causal properties" in our neurons that give rise to the mind. They point out that, by Searle's own description, 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. In short, Searle's "causal properties" and consciousness itself is undetectable, and anything that cannot be detected either does not exist or does not matter.[111]

Daniel Dennett provides this extension to the "epiphenomena" argument.

Dennett's reply from natural selection

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 is most likely that human beings (as we see them today) are actually "zombies", who nevertheless insist they are conscious. It is impossible to know whether we are all zombies or not. Even if we are all zombies, we would still believe that we are not.[112]

Searle disagrees with this analysis and argues 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."[73] He takes it as obvious that we can detect the presence of consciousness and dismisses these replies as being off the point.

Newton's flaming laser sword reply

Mike Alder argues that the entire argument is frivolous, because it is non-verificationist: not only is the distinction between *simulating* a mind and *having* a mind ill-defined, but it is also irrelevant because no experiments were, or even can be, proposed to distinguish between the two.[113]

In popular culture

The Chinese room argument is a central concept in Peter Watts's novels *Blindsight* and (to a lesser extent) *Echopraxia*.[114] It is also a central theme in the video game *Virtue's Last Reward*, and ties into the game's narrative. In Season 4 of the American crime drama *Numb3rs* there is a brief reference to the Chinese room.

The Chinese Room is also the name of a British independent video game development studio best known for working on experimental first-person games, such as *Everybody's Gone to the Rapture*, or *Dear Esther*.[115]

In the 2016 video game The Turing Test, the Chinese Room thought experiment is explained to the player by an AI.

See also

- Computational models of language acquisition
- Emergent behavior
- No true Scotsman
- Philosophical zombie
- Sorites paradox
- *I Am a Strange Loop*

Notes

a. The section consciousness of this article discusses the relationship between the Chinese room argument and consciousness.

b. This version is from Searle's *Mind, Language and Society*[20] and is also quoted in Daniel Dennett's *Consciousness Explained*.[21] 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."[22] Strong AI is defined similarly by Stuart Russell and Peter Norvig: "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."[4]

c. 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."[7] He also writes: "On the Strong AI view, the appropriately programmed computer does not just simulate having a mind; it literally has a mind."[8]

d. Searle writes: "Partisans of strong AI 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 *explains* the human ability to understand the story and answer questions about it."[6]

e. Note that Leibniz' was objecting to a "mechanical" theory of the mind (the philosophical position known as mechanism.) Searle is objecting to an "information processing" view of the mind (the philosophical position known as "computationalism"). Searle accepts mechanism and rejects computationalism.

f. Harnad edited *BBS* during the years which saw the introduction and popularisation of the Chinese Room argument.

g. Stevan Harnad 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".[18] David Cole agrees that "the argument also has broad implications for functionalist and computational theories of meaning and of mind".[19]

h. Searle believes that "strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter." [26] 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." [27] This position owes its phrasing to Stevan Harnad.[28]

i. "One of the points at issue," writes Searle, "is the adequacy of the Turing test."[29]

j. Computationalism is associated with Jerry Fodor and Hilary Putnam,[32] and is held by Allen Newell,[28] Zenon Pylyshyn[28] and Steven Pinker,[33] among others.

k. See the replies to Searle under Meaninglessness, below

l. Larry Hauser writes that "biological naturalism is either confused (waffling between identity theory and dualism) or else it *just is* identity theory or dualism."[37]

m. The wording of each axiom and conclusion are from Searle's presentation in *Scientific American*.[52][53] (A1-3) and (C1) are described as 1,2,3 and 4 in David Cole.[54]

n. Paul and Patricia Churchland write that the Chinese room thought experiment is intended to "shore up axiom 3".[55]

o. David Cole combines the second and third categories, as well as the fourth and fifth.[56]

p. This position is held by Ned Block, Jack Copeland, Daniel Dennett, Jerry Fodor, John Haugeland, Ray Kurzweil, and Georges Rey, among others.[58]

q. The virtual mind reply is held by Marvin Minsky, Tim Maudlin, David Chalmers and David Cole.[61] The reply was introduced by Marvin Minsky.[62][63]

r. David Cole 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."[67]

s. This position is held by Margaret Boden, Tim Crane, Daniel Dennett, Jerry Fodor, Stevan Harnad, Hans Moravec, and Georges Rey, among others.[70]

t. David Cole calls this the "externalist" account of meaning.[72]

u. The derived meaning reply is associated with Daniel Dennett and others.

v. 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. David Cole writes "derived intentionality is all there is, according to Dennett."[75]

w. David Cole describes this as the "internalist" approach to meaning.[72] 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." [76]

x. Searle 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."[79]

y. The brain simulation reply has been made by Paul Churchland, Patricia Churchland and Ray Kurzweil.[81]

z. Early versions of this argument were put forward in 1974 by Lawrence Davis and in 1978 by Ned Block. Block's version used walkie talkies and was called the "Chinese Gym". Paul and Patricia Churchland described this scenario as well.[84]

aa. An early version of the brain replacement scenario was put forward by Clark Glymour in the mid-70s and was touched on by Zenon Pylyshyn in 1980. Hans Moravec presented a vivid version of it,[86] and it is now associated with Ray Kurzweil's version of transhumanism.

ab. The connectionist reply is made by Andy Clark and Ray Kurzweil,[88] as well as Paul and Patricia Churchland.[89]

ac. Searle (2009) uses the name "Wait 'Til Next Year Reply".

ad. Searle writes that the robot reply "tacitly concedes that cognition is not solely a matter of formal symbol manipulation." [73] Stevan Harnad 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."[49]

ae. That is, any program running on a machine with a finite amount memory.

af. Speed and complexity replies are made by Daniel Dennett, Tim Maudlin, David Chalmers, Steven Pinker, Paul Churchland, Patricia Churchland and others.[96] Daniel Dennett points out the complexity of world knowledge.[76]

ag. Critics of the "phase transition" form of this argument include Stevan Harnad, Tim Maudlin, Daniel Dennett and David Cole.[96] This "phase transition" idea is a version of strong emergentism (what Daniel Dennett derides as "Woo woo West Coast emergence"[99]). Harnad accuses Churchland and Patricia Churchland of espousing strong emergentism. Ray Kurzweil also holds a form of strong emergentism.[100]

ah. The "other minds" reply has been offered by Daniel Dennett, Ray Kurzweil and Hans Moravec, among others.[103]

ai. 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." [107] Although Turing is discussing consciousness (not the mind or understanding or intentionality), Stuart Russell and Peter Norvig argue that Turing's comments apply the Chinese room.[108]

Citations

1. Harnad 2001, p. 1.
2. Roberts 2016.
3. Searle 1992, p. 44.
4. Russell & Norvig 2003, p. 947.
5. Searle 1980, p. 11.
6. Searle 1980, p. 2.
7. Searle 2009, p. 1.
8. Searle 2004, p. 66.
9. Searle 1980, p. 3.
10. Cole 2004, 2.1; Leibniz 1714, section 17.
11. "A Russian Chinese Room story antedating Searle's 1980 discussion" (http://www.hardproblem.ru/en/posts/Events/a-russian-chinese-room-story-antedating-searle-s-1980-discussion/). *Center for Consciousness Studies*. June 15, 2018.
12. Cole 2004, 2.3.
13. Searle 1980.
14. Cole 2004, p. 2; Preston & Bishop 2002
15. Harnad 2001, p. 2.
16. Harnad 2001, p. 1; Cole 2004, p. 2
17. Akman 1998.
18. Harnad 2005, p. 1.
19. Cole 2004, p. 1.
20. Searle 1999, p. .
21. Dennett 1991, p. 435.
22. Searle 1980, p. 1.
23. Quoted in Russell & Norvig 2003, p. 21.
24. Quoted in Crevier 1993, p. 46 and Russell & Norvig 2003, p. 17.
25. Haugeland 1985, p. 2 (Italics his)
26. Searle 1980, p. 13.
27. Searle 1980, p. 8.
28. Harnad 2001.
29. Searle 1980, p. 6.
30. Searle 2004, p. 45.
31. Harnad 2001, p. 3 (Italics his)
32. Horst 2005, p. 1.
33. Pinker 1997.
34. Harnad 2001, pp. 3–5.
35. Searle 1990a, p. 29.
36. Searle 1990b.
37. Hauser 2006, p. 8.
38. Searle 1992, chpt. 5.
39. Searle 2002.
40. Chalmers 1996, p. 322.
41. McGinn 2000.
42. Hew 2016.
43. Kurzweil 2005, p. 260.
44. Saygin, Cicekli & Akman 2000.
45. Turing 1950.
46. Newell & Simon 1976, p. 116.
47. Russell & Norvig 2003, p. 18.
48. Turing 1950, p. 442.
49. Harnad 2001, p. 14.
50. Ben-Yami 1993.
51. Searle 1984; Searle 1990a.
52. Searle 1990a.
53. Hauser 2006, p. 5.
54. Cole 2004, p. 5.
55. Churchland & Churchland 1990, p. 34.
56. Cole 2004, pp. 5–6.
57. Searle 1980, pp. 5–6; Cole 2004, pp. 6–7; Hauser 2006, pp. 2–3, Russell & Norvig 2003, p. 959, Dennett 1991, p. 439; Fearn 2007, p. 44; Crevier 1993, p. 269.
58. Cole 2004, p. 6.
59. Yee 1993, p. 44.
60. Yee 1993, pp. 42–47.
61. Cole 2004, pp. 7–9.
62. Minsky 1980, p. 440.
63. Cole 2004, p. 7.
64. Cole 2004, p. 8.
65. Searle 1980, p. 12.
66. Fearn 2007, p. 47.
67. Cole 2004, p. 21.
68. Searle 2004, p. 63.
69. Searle 1980, p. 7; Cole 2004, pp. 9–11; Hauser 2006, p. 3; Fearn 2007, p. 44.
70. Cole 2004, p. 9.
71. Quoted in Crevier 1993, p. 272
72. Cole 2004, p. 18.
73. Searle 1980, p. 7.
74. Hauser 2006, p. 11; Cole 2004, p. 19.
75. Cole 2004, p. 19.
76. Dennett 1991, p. 438.
77. Dreyfus 1979, "The epistemological assumption".

78. Searle 1984.
79. Motzkin & Searle 1989, p. 45.
80. Searle 1980, pp. 7–8; Cole 2004, pp. 12–13; Hauser 2006, pp. 3–4; Churchland & Churchland 1990.
81. Cole 2004, p. 12.
82. Searle 1980, p. .
83. Cole 2004, p. 4; Hauser 2006, p. 11.
84. Churchland & Churchland 1990.
85. Russell & Norvig 2003, pp. 956–958; Cole 2004, p. 20; Moravec 1988; Kurzweil 2005, p. 262; Crevier 1993, pp. 271 and 279.
86. Moravec 1988.
87. Searle 1992 quoted in Russell & Norvig 2003, p. 957.
88. Cole 2004, pp. 12 & 17.
89. Hauser 2006, p. 7.
90. Searle 1980, pp. 8–9; Hauser 2006, p. 11.
91. Block 1981.
92. Quoted in Cole 2004, p. 13.
93. Dennett 1991, pp. 437–440.
94. Crevier 1993, p. 269.
95. Cole 2004, pp. 14–15; Crevier 1993, pp. 269–270; Pinker 1997, p. 95.
96. Cole 2004, p. 14.
97. Churchland & Churchland 1990; Cole 2004, p. 12; Crevier 1993, p. 270; Fearn 2007, pp. 45–46; Pinker 1997, p. 94.
98. Harnad 2001, p. 7.
99. Crevier 1993, p. 275.
100. Kurzweil 2005.
101. Searle 1980, p. 10.
102. Searle 1980, p. 9; Cole 2004, p. 13; Hauser 2006, pp. 4–5; Nilsson 1984.
103. Cole 2004, pp. 12–13.
104. Nilsson 1984.
105. Turing 1950, pp. 11–12.
106. Turing 1950, p. 11.
107. Turing 1950, p. 12.
108. Russell & Norvig 2003, pp. 952–953.
109. Dennett 1991,.
110. "Eliminative Materialism" (https://plato.stanford.edu/entries/materialism-eliminative/). *Stanford Encyclopedia of Philosophy*. March 11, 2019.
111. Russell & Norvig 2003.
112. Cole 2004, p. 22; Crevier 1993, p. 271; Harnad 2005, p. 4.
113. Alder 2004.
114. Whitmarsh 2016.
115. "Home" (http://www.thechineseroom.co.uk/). *The Chinese Room*. Retrieved 2018-04-27.

References

- Akman, Varol (1998). "Book Review — *John Haugeland (editor), Mind Design II: Philosophy, Psychology, and Artificial Intelligence*" (http://cogprints.org/539/index.html). Retrieved 2018-10-02 – via Cogprints.

- Alder, Mike (2004). "Newton's Flaming Laser Sword" (http://www.philosophynow.org/issues/46/Newtons_Flaming_Laser_Sword). *Philosophy Now*. **46**: 29–33. Archived (https://web.archive.org/web/20180326114344/http://www.philosophynow.org/issues/46/Newtons_Flaming_Laser_Sword) from the original on 2018-03-26. Also available at "Newton's Flaming Laser Sword" (https://web.archive.org/web/20111114041242/http://school.maths.uwa.edu.au/~mike/Newtons%20Flaming%20Laser%20Sword.pdf) (PDF). Archived from the original (http://school.maths.uwa.edu.au/~mike/Newtons%20Flaming%20Laser%20Sword.pdf) (PDF) on 2011-11-14.

- Ben-Yami, Hanoch (1993), "A Note on the Chinese Room", *Synthese*, **95** (2): 169–72, doi:10.1007/bf01064586 (https://doi.org/10.1007%2Fbf01064586)

- Block, Ned (1981), "Psychologism and Behaviourism" (http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Psychologism.htm), *The Philosophical Review*, **90** (1): 5–43, CiteSeerX 10.1.1.4.5828 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.5828), doi:10.2307/2184371 (https://doi.org/10.2307%2F2184371), JSTOR 2184371 (https://www.jstor.org/stable/2184371)

- Chalmers, David (March 30, 1996), *The Conscious Mind: In Search of a Fundamental Theory* (https://books.google.com/books?id=XtgiH-feUylC), Oxford University Press, ISBN 978-0-19-

983935-3

- Churchland, Paul; Churchland, Patricia (January 1990), "Could a machine think?", *Scientific American*, **262** (1): 32–39, Bibcode:1990SciAm.262a..32C (https://ui.adsabs.harvard.edu/abs/1990SciAm.262a..32C), doi:10.1038/scientificamerican0190-32 (https://doi.org/10.1038%2Fscientificamerican0190-32), PMID 2294584 (https://pubmed.ncbi.nlm.nih.gov/2294584)
- Cole, David (Fall 2004), "The Chinese Room Argument", in Zalta, Edward N. (ed.), *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 978-0-7139-9037-9
- Dreyfus, Hubert (1979), *What Computers* Still *Can't Do*, New York: MIT Press, ISBN 978-0-262-04134-8
- 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 Argument", in M.; Preston, J. (eds.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (http://cogprints.org/4023/), Oxford University Press

  Page numbers above refer to a standard PDF print of the article.

- Harnad, Stevan (2005), "Searle's Chinese Room Argument", *Encyclopedia of Philosophy* (http://eprints.ecs.soton.ac.uk/10424/01/chineseroom.html), Macmillan

  Page numbers above refer to a standard PDF print of the article.

- Haugeland, John (1985), *Artificial Intelligence: The Very Idea*, Cambridge, Mass.: MIT Press, ISBN 978-0-262-08153-5
- Haugeland, John (1981), *Mind Design* (https://archive.org/details/minddesignphilos00haug), Cambridge, Mass.: MIT Press, ISBN 978-0-262-08110-8
- Hauser, Larry (2006), "Searle's Chinese Room", *Internet Encyclopedia of Philosophy* (http://www.iep.utm.edu/c/chineser.htm)

  Page numbers above refer to a standard PDF print of the article.

- Hew, Patrick Chisan (September 2016). "Preserving a combat commander's moral agency: The Vincennes Incident as a Chinese Room" (https://philpapers.org/rec/HEWPAC). *Ethics and Information Technology*. **18** (3): 227–235. doi:10.1007/s10676-016-9408-y (https://doi.org/10.1007%2Fs10676-016-9408-y).
- Kurzweil, Ray (2005), *The Singularity is Near*, Viking Press
- Horst, Steven (Fall 2005), "The Computational Theory of Mind", in Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu/archives/fall2005/entries/computational-mind/)
- Leibniz, Gottfried (1714), *Monadology* (https://web.archive.org/web/20110703072430/http://www.philosophy.leeds.ac.uk/GMR/moneth/monadology.html), George MacDonald Ross (trans.), archived from the original (http://www.philosophy.leeds.ac.uk/GMR/moneth/monadology.html) on 2011-07-03
- Moravec, Hans (1988), *Mind Children: The Future of Robot and Human Intelligence* (https://books.google.com/books?id=56mb7XuSx3QC), Harvard University Press

- Minsky, Marvin (1980), "Decentralized Minds", *Behavioral and Brain Sciences*, **3** (3): 439–40, doi:10.1017/S0140525X00005914 (https://doi.org/10.1017%2FS0140525X00005914)
- McGinn, Collin (2000), *The Mysterious Flame: Conscious Minds In A Material World* (https://books.google.com/books?id=qB0lg0u3BEkC&pg=PA194), Basic Books, p. 194, ISBN 978-0786725168
- Moor, James, ed. (2003), *The Turing Test: The Elusive Standard of Artificial Intelligence*, Dordrecht: Kluwer Academic Publishers, ISBN 978-1-4020-1205-1
- Motzkin, Elhanan; Searle, John (February 16, 1989), *Artificial Intelligence and the Chinese Room: An Exchange* (http://www.nybooks.com/articles/archives/1989/feb/16/artificial-intelligence-and-the-chinese-room-an-ex/?pagination=false), New York Review of Books
- Newell, Allen; Simon, H. A. (1976), "Computer Science as Empirical Inquiry: Symbols and Search", *Communications of the ACM*, **19** (3): 113–126, doi:10.1145/360018.360022 (https://doi.org/10.1145%2F360018.360022)
- Nikolić, Danko (2015), "Practopoiesis: Or how life fosters a mind", *Journal of Theoretical Biology*, **373**: 40–61, arXiv:1402.5332 (https://arxiv.org/abs/1402.5332), doi:10.1016/j.jtbi.2015.03.003 (https://doi.org/10.1016%2Fj.jtbi.2015.03.003), PMID 25791287 (https://pubmed.ncbi.nlm.nih.gov/25791287)
- Nilsson, Nils (1984), *A Short Rebuttal to Searle* (https://ai.stanford.edu/%7Enilsson/OnlinePubs-Nils/General%20Essays/OtherEssays-Nils/searle.pdf) (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
- Pinker, Steven (1997), *How the Mind Works*, New York, NY: W. W. Norton & Company, Inc., ISBN 978-0-393-31848-7
- Preston, John; Bishop, Mark, eds. (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (https://books.google.com/books?id=N7msQgAACAAJ), Oxford University Press, ISBN 978-0-19-825057-9
- Roberts, Jacob (2016). "Thinking Machines: The Search for Artificial Intelligence" (https://web.archive.org/web/20180819152455/https://www.sciencehistory.org/distillations/magazine/thinking-machines-the-search-for-artificial-intelligence). *Distillations*. **2** (2): 14–23. Archived from the original (https://www.sciencehistory.org/distillations/magazine/thinking-machines-the-search-for-artificial-intelligence) on 2018-08-19. Retrieved 2018-03-22.
- Saygin, A. P.; Cicekli, I.; Akman, V. (2000), "Turing Test: 50 Years Later" (https://web.archive.org/web/20110409073501/http://crl.ucsd.edu/~saygin/papers/MMTT.pdf) (PDF), *Minds and Machines*, **10** (4): 463–518, doi:10.1023/A:1011288000451 (https://doi.org/10.1023%2FA%3A1011288000451), hdl:11693/24987 (https://hdl.handle.net/11693%2F24987), archived from the original (http://crl.ucsd.edu/~saygin/papers/MMTT.pdf) (PDF) on 2011-04-09, retrieved 2015-06-05. Reprinted in Moor (2003, pp. 23–78).
- Searle, John (1980), "Minds, Brains and Programs" (https://web.archive.org/web/20071210043312/http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html), *Behavioral and Brain Sciences*, **3** (3): 417–457, doi:10.1017/S0140525X00005756 (https://doi.org/10.1017%2FS0140525X00005756), retrieved 2009-05-13

  Page numbers above refer to a standard PDF print of the article. See also Searle's original draft (https://web.archive.org/web/20010221025515/http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html).

- Searle, John (1983), "Can Computers Think?", in Chalmers, David (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, Oxford: Oxford University Press, pp. 669–675, ISBN 978-0-19-514581-6

- Searle, John (1984), *Minds, Brains and Science: The 1984 Reith Lectures* (https://archive.org/details/mindsbrainsscien0000sear), Harvard University Press, ISBN 978-0-674-57631-5 paperback: ISBN 0-674-57633-0.
- Searle, John (January 1990a), "Is the Brain's Mind a Computer Program?", *Scientific American*, vol. 262 no. 1, pp. 26–31, Bibcode:1990SciAm.262a..26S (https://ui.adsabs.harvard.edu/abs/1990SciAm.262a..26S), doi:10.1038/scientificamerican0190-26 (https://doi.org/10.1038%2Fscientificamerican0190-26), PMID 2294583 (https://pubmed.ncbi.nlm.nih.gov/2294583)
- Searle, John (November 1990b), "Is the Brain a Digital Computer?" (https://web.archive.org/web/20121114093932/https://mywebspace.wisc.edu/lshapiro/web/Phil554_files/SEARLE-BDC.HTM), *Proceedings and Addresses of the American Philosophical Association*, **64** (3): 21–37, doi:10.2307/3130074 (https://doi.org/10.2307%2F3130074), JSTOR 3130074 (https://www.jstor.org/stable/3130074), archived from the original (https://mywebspace.wisc.edu/lshapiro/web/Phil554_files/SEARLE-BDC.HTM) on 2012-11-14
- Searle, John (1992), *The Rediscovery of the Mind* (https://books.google.com/books?id=eoh8e52wo_oC), Cambridge, Massachusetts: M.I.T. Press, ISBN 978-0-262-26113-5
- Searle, John (1999), *Mind, language and society* (https://archive.org/details/mindlanguagesoci00sear), New York, NY: Basic Books, ISBN 978-0-465-04521-1, OCLC 231867665 (https://www.worldcat.org/oclc/231867665)
- Searle, John (November 1, 2004), *Mind: a brief introduction* (https://books.google.com/books?id=oSm8JUHJXqcC), Oxford University Press, Inc., ISBN 978-0-19-515733-8
- Searle, John (2002), *Consciousness and Language* (https://books.google.com/books?id=bvxhV-1Duz8C&pg=PA16), Cambridge University Press, p. 16, ISBN 978-0521597449
- Searle, John (2009), "Chinese room argument", *Scholarpedia*, **4** (8): 3100, Bibcode:2009SchpJ...4.3100S (https://ui.adsabs.harvard.edu/abs/2009SchpJ...4.3100S), doi:10.4249/scholarpedia.3100 (https://doi.org/10.4249%2Fscholarpedia.3100)
- Turing, Alan (October 1950), "Computing Machinery and Intelligence", *Mind*, **LIX** (236): 433–460, doi:10.1093/mind/LIX.236.433 (https://doi.org/10.1093%2Fmind%2FLIX.236.433), ISSN 0026-4423 (https://www.worldcat.org/issn/0026-4423)

  Page numbers above refer to a standard PDF print of the article.

- Whitmarsh, Patrick (2016). " "Imagine You're a Machine": Narrative Systems in Peter Watts's *Blindsight* and *Echopraxia*". *Science Fiction Studies*. **43** (2): 237–259. doi:10.5621/sciefictstud.43.2.0237 (https://doi.org/10.5621%2Fsciefictstud.43.2.0237).
- Yee, Richard (1993), "Turing Machines And Semantic Symbol Processing: Why Real Computers Don't Mind Chinese Emperors" (http://lyceumphilosophy.com/Lyceum-5-1.pdf) (PDF), *Lyceum*, **5** (1): 37–59

  Page numbers above and diagram contents refer to the Lyceum PDF print of the article.

Further reading

- General presentations of the argument :

  - "Chinese Room Argument" (http://www.iep.utm.edu/chineser). *Internet Encyclopedia of Philosophy*.
  - The Chinese Room Argument (https://plato.stanford.edu/entries/chinese-room/). *Stanford Encyclopedia of Philosophy*
  - Understanding the Chinese Room (http://www.zompist.com/searle.html), Mark Rosenfelder
- Sources involving John Searle :

- Chinese room argument (http://www.scholarpedia.org/article/Chinese_room_argument) by John Searle on Scholarpedia
- 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
- John R. Searle, "What Your Computer Can't Know" (review of Luciano Floridi, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*, Oxford University Press, 2014; and Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014), *The New York Review of Books*, vol. LXI, no. 15 (October 9, 2014), pp. 52–55.

- Criticism of the argument :

  - 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 (https://doi.org/10.1017%2FS0140525X04210044). S2CID 56032408 (https://api.semanticscholar.org/CorpusID:56032408)., 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 (https://arxiv.org/abs/cs.AI/0403009).
  - 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 (https://www.worldcat.org/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 0-521-24868-X (1988); reprinted in Boden (ed.) "The Philosophy of Artificial Intelligence" ISBN 0-19-824854-7 (1989) and ISBN 0-19-824855-5 (1990); Boden "Artificial Intelligence in Psychology: Interdisciplinary Essays" ISBN 0-262-02285-0, 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")