Chinese Room Argument, The

Intermediate article

John Searle, University of California, Berkeley, California, USA

CONTENTS

The Chinese room argument is a refutation of 'strong artificial intelligence' (strong AI), the view that an appropriately programmed digital computer capable of passing the Turing test would thereby have mental states and a mind in the same sense in which human beings have mental states and a mind. Strong AI is distinguished from weak AI, which is the view that the computer is a useful tool in studying the mind, just as it is a useful tool in other disciplines ranging from molecular biology to weather prediction.

SUMMARY OF THE ARGUMENT

Strong AI is often expressed in the formula: 'Mind is to brain as program is to hardware.' On this view, the human mind is a program running in the hardware, or 'wetware', of the brain. The Chinese room argument against strong AI proceeds by a thought experiment. If strong AI were true, then one could acquire any cognitive capacity that one does not have by simply implementing the program for that cognitive capacity in a way that would enable one to pass the Turing test.

Imagine that I, who am a native English speaker, unable to speak any Chinese at all, am locked in a room containing several boxes of Chinese symbols (the database). Imagine that I have in the room a set of instructions for manipulating Chinese symbols (the program). I receive, through a window in the room, Chinese symbols which, unknown to me, are in the form of questions. I follow the instructions in the program, and give back through the window Chinese symbols which, unknown to me, are answers to the questions. For the purposes of the thought experiment we may suppose that the programmers get so good at writing the programs, and I get so good at shuffling the symbols, that after a time my answers are indistinguishable from those of the native Chinese speaker. I pass the Turing test for understanding Chinese, and I do so by implementing the program. But I do not understand a word of Chinese. This is the point of the thought experiment: if I do not understand Chinese by virtue of implementing the Chinese-understanding program, then neither does any other digital computer by virtue of doing so.

Why is it that I do not understand Chinese? The answer seems obvious. Though I manipulate the symbols, I have no knowledge of what any of the symbols mean. One can see this by contrasting my manipulation of Chinese symbols with my answering questions in English. Suppose that the people on the outside of the room also submit written questions in English and I submit written answers to the questions. My answers to the questions in Chinese are as good as those of a native Chinese speaker because I have been appropriately programmed. My answers to the questions in English are as good as a native English speaker because I am a native English speaker. From the outside, from the third-person behavioral point of view, my behavior is equally good in Chinese and in English. From the inside it is obviously quite different: in English I understand perfectly both the questions and my answers; while in Chinese I understand nothing – I am just a digital computer.

Construed as a deductive argument, the Chinese room argument has three steps and a conclusion. We may formulate these as follows.

> Computer programs are defined entirely in terms of symbolic or syntactic operations. (1)

The implemented program consists entirely of symbol manipulations. To put this somewhat more precisely: the notion 'same implemented program' defines an equivalence class that is specified entirely in symbolic or syntactic terms, and independently of the physics of the underlying medium. There is nothing more to the implemented program, qua implemented program, than symbol manipulations.

> Minds – actual human minds such as yours and mine – have mental contents or semantics. (2)

For example, when I understand a sentence in English I have more than just symbols going through my head: I know what the symbols mean.

> By themselves, the implemented syntactic steps of the program are neither constitutive of mental content nor sufficient to guarantee the presence of mental content. (3)

This is what was shown by the Chinese room thought experiment. I went through the appropriate syntactic steps, but I had no Chinese thought content, no Chinese semantics, associated with them.

> Conclusion: the implemented computer program is insufficient by itself to constitute or to guarantee the presence of the appropriate mental states. (4)

I went through the right steps of the program, I had the right behavior, but I did not have the appropriate mental states. Therefore, strong AI is false.

The argument rests on two fundamental logical principles: firstly, syntax is not semantics, and secondly, simulation is not duplication. Any problem-solving process that can be described as an effective procedure, that is, a procedure going through a finite number of exactly specifiable discrete steps, can be programmed on a computer. That is why the computer is so powerful: we can represent any domain that we can describe precisely. Thus we can represent the stages of the weather, the flow of money in the economy, or the understanding of Chinese sentences. The syntax of the program states can be used to represent anything. They can be used to represent weather changes, economic developments, and even semantics. But the simulation of the process, whether it be atmospheric, economic, or semantic, is not a duplication of the process. You do not produce a rainstorm by doing a computer simulation of a rainstorm. You do not produce wealth by doing a computer simulation of the production of wealth. And you do not produce understanding and thought processes by doing a computer simulation of understanding and thought processes.

RESPONSES TO THE ARGUMENT

A number of responses have been presented against the Chinese room argument. We will consider four of these.

The Systems Reply

Perhaps the most commonly presented answer is this: the person in the room does not understand, but the person is only an element in a larger system. The system consists of the room, the program, the database, etc. So the understanding should be found in the entire system, not in the person, because the person is only the central processing unit.

Just as we would not say of a single neuron in the brain that it understood English, so we should not say of a single element, the person, in the whole system that that person understands Chinese.

Answer to the systems reply

The answer to this reply is that the reason the person does not understand is that he has no way to attach any meaning to the symbols. But if he has no way to attach meaning to the symbols, neither does the whole room. The whole room has no way to get from the syntax to the semantics any more than the person does. To see this, simply imagine that the person internalizes the entire system. Imagine that the person memorizes the database, memorizes the program, does all of the calculations in his head, and works outdoors in the middle of an open field. In this variation there is nothing in the room that is not in the person, and still there is no understanding in the person.

The Robot Reply

The robot reply is based on a variation of strong AI whereby the unit of understanding is not the computer, but the computer within a motorized system that will be able to process sensory inputs and produce motor outputs computationally. The robot would move about, with video cameras attached to its head; it would take in information from the video cameras, and adjust its movements accordingly.

According to the robot reply, the computer by itself does not have semantic content, but the causal relations between a robot and the external world would be sufficient to give semantic content to the symbols processed by the robot.

Answer to the robot reply

The robot reply tacitly abandons the thesis of strong AI, which is that the implemented computer program by itself is sufficient to guarantee or constitute understanding. The idea behind the robot reply is that the addition of causal relations between the system and the external world would

be sufficient to produce semantic content or understanding.

The answer to the robot reply is that even this amendment to the strong AI thesis will not be sufficient to produce understanding. Imagine that the robot has a very large cranium, and inside the cranium is a room, and I am inside the room. I receive inputs in the form of Chinese symbols. I process them according to the program, and I produce outputs in the form of Chinese symbols. Unknown to me, the input symbols are the product of video cameras and other sensors attached to the outside of the robot. The input stimuli are converted by transduction into Chinese symbols, and the output that I provide is converted into motor output of the entire robot system. But I have no way of understanding what is going on because I have no way of attaching any meaning to any of the symbols, or to anything else that is going on in the robot's brain. I am the robot's homunculus, but unlike the usual homunculi of philosophical literature, I understand nothing, because I have no way of attaching any meaning to any of the symbols that I process.

The robot reply tries to defeat the Chinese room argument by adding causal relations. But the causal relations will produce semantic content only if there is some conscious agent who can become aware of the causal relations. I, as a human being, can become aware of the causal relations between the Chinese symbol for chicken chow mein and the actual food type of chicken chow mein if I can *see* chicken chow mein associated with this symbol. But in the robot, I am just a computer and, like any other computer, I function by processing meaningless symbols. The symbols in the computer brain are meaningless in a way that is quite different from the symbols passing through my mind when I think in English. When I think in English, symbols do indeed go through my mind, but I know what they mean.

The Brain Simulator Reply

Suppose we simulated the actual operations of a Chinese person's brain when that person understands sentences in Chinese. Suppose we produced a perfect computer simulation of all of the synaptic transmissions in the Chinese brain. Then we would have to say that the system understood, otherwise we would have to deny that the Chinese person understood. Since the brain operates, like a computer, with a series of state transitions, there is no reason why we could not produce a perfect replica of these state transitions on a digital computer.

Answer to the brain simulator reply

The computer simulation of the brain is not duplicating the relevant features of the brain. It is merely duplicating the formal pattern. The actual human brain, like any other organ, is a causal mechanism, and it causes consciousness and intentionality by quite specific neurobiological processes. The computer merely produces a model or representation of these processes, but the model or representation lacks the causal features of the original.

To see this, compare the brain to any other organ. We can do a perfect simulation of the digestive processes in the stomach on a digital computer. But even if we have a perfect simulation on the computer, to any degree of accuracy, we do not produce actual digestion. When we run the digestion program, we cannot put a pizza into the computer and expect the computer to digest it. The computational simulation is merely a matter of zeros and ones, not a matter of the enzymes and other chemicals that actually carry out digestion. The situation in the brain is similar. Specific biochemical processes cause consciousness and intentionality. We cannot reproduce those by doing a simulation with zeros and ones, any more than we can reproduce digestion with zeros and ones.

The Parallel Distributed Processing Reply

The Chinese room argument works against the von Neumann symbolic digital computer, but recent developments in computer technology have created new types of computational systems which are immune to the Chinese room argument. These new types of systems are known variously as 'parallel distributed processing' (PDP), 'neural net modeling, 'connectionism' or 'new connectionism'.

PDP systems function in a way that is quite different from the traditional von Neumann system. They have a series of computational processes going on in parallel, distributed over a network. Whereas the traditional von Neumann machine works in a series of discrete steps, PDP systems do massively parallel distributed processing.

Answer to the parallel distributed processing reply

There is an ambiguity in the PDP reply. It is not clear which of the differences between the connectionist machine and the von Neumann machine are being appealed to in order to claim that the connectionist machine is not subject to the Chinese room argument. Either what is claimed is that there is some computational power of the connectionist

machine lacking in the von Neumann machine, or it is claimed that there is some hardware feature of the connectionist architecture which is superior to the von Neumann architecture. But neither of these approaches is successful in answering the Chinese room argument.

According to Church's thesis, there is no computation that can be performed on a connectionist machine that cannot be performed on a von Neumann machine. According to this thesis, any computable function whatever, any problem that can be solved algorithmically, can be computed on a Turing machine. All effective computability is Turing computability. Church's thesis is one of the foundational principles of the modern theory of computation and is universally accepted by the parties to this dispute. It has the consequence that there cannot be any computational power possessed by a PDP system that is not possessed by a von Neumann machine.

The other possibility is that something is being claimed for the connectionist architecture: for the actual structure of the wiring and the hardware. But if this is so, it is no longer strong AI. Strong AI is a thesis about the powers of computation. If it is claimed that the particular hardware of the connectionist machines can duplicate the powers of the brain to cause mental content, then the thesis is no longer strong AI, but is rather a form of speculative neurobiology. The Chinese room argument is not intended to answer any claims in speculative neurobiology, but is intended as a logical thesis about the distinction between the syntax of the implemented program and the semantics of actual human minds.

Either we are to think of the essential feature of the system as being its computational power, or we are to think of it as some causal property of the specific hardware in which the computation is implemented. If it is a matter of computational power, no new power is added by the connectionist architecture. If we are to think of it as an architectural feature, then it is no longer the thesis of strong AI. Actual human brains cause consciousness and other mental phenomena by way of specific neurobiological processes operating in a 'bottom-up' fashion. That is, processes at the level of neurons and synapses cause consciousness and other mental phenomena that are features of much larger elements of the brain system.

The thesis of the PDP reply, if followed to its logical conclusion, would have to be that the connectionist architecture is capable of duplicating and not merely simulating the causal powers of the brain to cause higher-level consciousness, etc., by way of bottom-up causation. Nothing in the neurobiological literature would tend to support this thesis. In any case, it is not strong AI, and in consequence, is irrelevant to answering the Chinese room argument.

COMMON MISUNDERSTANDINGS OF THE CHINESE ROOM ARGUMENT

The Chinese room argument is sometimes misinterpreted, and several of these misinterpretations are common in the literature. Firstly, it is sometimes supposed that the argument is intended to show that 'machines cannot think'. But that is not the point of the argument. The argument assumes that the brain is a machine. The problem with computation is that in the relevant sense it does not name a machine process. It names an abstract mathematical process that we can implement on machines, but computation is not defined in terms of machine processes such as energy transfer. Thus, on the view implicit in the Chinese room argument, the brain is a machine, and brain processes are machine processes. 'Computers' of the ordinary kind are indeed machines, but computation is not essentially a machine process.

Another misinterpretation of the Chinese room argument is that it is attempting to show that only human brains have the power of thinking. But that is not the point of the argument. Whether or not we can build an artifact out of some other type of material capable of producing consciousness is an empirical scientific question. In principle there is no more serious logical obstacle to building an artificial brain than there is to building an artificial heart. The point of the argument is that we do not produce the same causal powers by simply duplicating the formal pattern. The computer gives us a picture, or a model, of thought processes, but it does not actually produce thought processes.

CONCLUSION

In the early days of cognitive science, the computationalist model of cognition was the dominant paradigm. At present there is a gradual paradigm shift away from computational cognitive science towards cognitive neuroscience. As we learn more about the brain we see that cognition is essentially a matter of a certain sort of brain processing. We may be able to simulate this on a digital computer, and we may eventually be able to duplicate it in some other medium. But the Chinese room argument shows that simulation by itself does not guarantee duplication. To guarantee duplication, the artificial

creation of a real mind, we would have to duplicate, and not merely simulate, the powers that actual brains have to cause consciousness and cognition.

Further Reading

Dietrich E (ed.) (1994) *Thinking Computers and Virtual Persons*. San Diego, CA: Academic Press.

Preston J and Bishop M (eds) (2002) *Views Into the Chinese Room: New Essays on John Searle's Arguments Against 'Strong AI'*. Oxford, UK: Oxford University Press.

Searle JR (1980) 'Minds, brains, and programs'. *Behavior and Brain Sciences* **3**: 417–457.

Searle JR (1982) The Chinese room revisited. *Behavior and Brain Sciences* **5**: 345–348.