THE CHINESE ROOM ARGUMENT

Minds, Brains, and Programs (1980)

By John Searle

IN:

Heil, PP. 235-52

Introduction

I. Searle's purpose is to refute "Strong" AI

   A. distinguishes Strong vs. Weak AI

      1. Strong AI

         a. a computer programmed in the right way really is a mind

         b. that is, it can understand and have other cognitive states

         c. the programs actually explain human cognition

      2. Weak AI

         a. the computer is a useful tool for the study of the human mind

         b. it helps us formulate and test our hypotheses in a more precise, rigorous way

   B. Searle has no objection to Weak AI - only to Strong AI

II. Schank's scripts

   A. these programs were supposed to allow a computer to be able to "understand" stories (Cf. Clark, p. 30)

   B. Searle chooses this example only because he is most familiar with it; he thinks the same arguments could apply to (235)

      1. Winograd's SHRDLU (manipulates blocks in micro-world)

      2. Weizenbaum's ELIZA, etc. (psychological counseling)

   C. Searle readily admits that a machine running the restaurant script could successfully answer questions about stories about going to the restaurant - see example about the man eating the hamburger, pp. 235-36

D. What he denies is: (236, q.v.)

1. that the computer actually understands what's going on
2. that what the computer and the program do explains how human beings are able to understand the story and answer questions about it

III. The "Chinese room" thought experiment is intended to show this

A. We take it as a given that Searle knows no Chinese

B. Then suppose he is locked in a room and given:

1. a first batch of Chinese symbols
2. a second batch of Chinese symbols, together with a set of rules written in English that he understands for correlating symbols from the first batch with symbols from the second
3. a third batch of Chinese symbols with more rules written in English
   a. for correlating symbols from the third with symbols from the first and second batches
   b. and for giving back responses with Chinese symbols

C. Unbeknownst to Searle (236)

1. the first batch is a script
2. the second batch is a short story
3. the third batch is a set of questions about the story (236-37)
4. the symbols he returns are answers to these questions (237)
5. the instructions in English are his “program”

D. Now let's suppose further that the answers that Searle gives to these questions are as good as those that a native speaker of Chinese would give

E. Nevertheless, Searle would argue

1. that he does not understand the Chinese story
   a. and for the same reason Schank's computer does not understand any stories in any language
   b. in effect, Searle is the computer in the Chinese room
2. that the computer program does not explain human understanding
   a. it is clearly not sufficient

b. Searle doubts that symbol manipulation is even a necessary part of explaining human understanding

1.) some claim that when he actually understands English, he's just doing more of the same as when he's manipulating Chinese symbols

a.) he hasn't actually shown this is false

b.) it gets its plausibility from

(1) the assumption that we can construct a program with the same input/output as a human (237-38)

(2) the assumption that a human speaker at some level can be described this way (238)

2.) given these two assumptions, it's at least possible that Schank's scripts are part of the explanation of how we understand English, but no reason has been given to think so

F. What is it then that allows him to understand English and not Chinese? And could we give this to a machine? He will turn to this after he considers some replies to his Chinese room thought experiment

<u>Replies to the Chinese Room and Searle's Responses</u>

Introduction (238)

A. Searle has received a variety of replies to his thought experiment from AI workers

B. But before he turns to consider the most common ones, he wants to remove a few misunderstandings first

1. People often argue that understanding is a matter of degree, that the law of the excluded middle does not apply to the statement "X understands Y," and so on.

a. Searle agrees (238-39)

b. for instance, he understands English better than French, etc. (239)

c. But finds the point irrelevant to his argument – all he needs for his argument are some clear cases where understanding applies and some clear cases where it doesn't

2. people often use words like "understand," "perceive," and "know" in a metaphorical sense when talking about machines

a. if this were all Schank meant, there would be no argument

b. however, there are some, like Newell and Simon, who say that computers understand in the same way that we do (Cf. Clark, p. 28)

C. Searle wants to defend the claim that in the literal sense of understanding, computers understand absolutely nothing

I. The Systems Reply: although the person in the room may not understand the story, the whole system, including the person, the room, the symbols, the rules, etc. does

Searle's Response

A. imagine that we take the person out of the room and have him memorize all the symbols and the rules: he still doesn't understand Chinese (239-40)

B. An AI enthusiast may reply: (240)

1. the person who memorized the rules and the symbols doesn't understand Chinese in the same way that someone who speaks Chinese does – that is, he doesn't know what the symbols *refer* to (240, q.v.)

2. nevertheless, the man as a system that manipulates Chinese symbols really does understand Chinese

3. and this subsystem is just separate from the subsystem that understands English

C. But Searle thinks:

1. the two subsystems are totally different:

a. the English one knows what's being referred to

b. the Chinese one just links this symbol with that

1.) the Chinese subsystem is no better off than the whole person

2.) in fact, since the rule book is written in English, the Chinese subsystem is really a subsystem within the English subsystem: it's one for talking about Chinese symbols

2. the AI enthusiast has no *independent* reasons for thinking that there is a subsystem that understands Chinese (241)

a. the only motivation for saying that it does is that it passes the Turing test for Chinese

b. but this only calls the Turing test into question

c. that is, although both his English and Chinese subsystems pass the Turing test, that does not mean that they equally understand (241)

d. in sum, the systems reply begs the question by simply assuming without argument that the system understands Chinese (241)

D. the systems reply also leads to some absurd consequences

1. if input, program, output were all that were needed for cognition, we could describe all sorts of things as cognitive systems

2. for example, we could describe the digestive system as a cognitive system

3. it does no good to argue that the digestive system processes food and not information

a. the Chinese symbols carry no information for Searle, either

b. in the Chinese case, the information is only in the eye of the programmers (q.v.)

E. If Strong AI is supposed to be an approach to psychology, it ought to be able to distinguish systems that are genuinely mental from those that are not

1. it must be able to distinguish the way minds work from the ways in which non-minds work (241-42)

2. e.g., McCarthy (1979) says thermostats have beliefs (242)

3. if you end up with a theory that tells you that even things like thermostats have beliefs, that mind is everywhere, there's something wrong with your theory

II. The Robot Reply: suppose we wrote a different sort of program than Schank's scripts and put it in a computer in a robot that had a TV camera and that could walk around and do other cool stuff. That robot would understand. (242-43)

Searle's Response (243)

A. This reply already concedes that cognition is more than just the manipulation of formal symbols and involves causal relations with the real world

B. The perceptual and motor capabilities of the robot don't add understanding to Schank's original program

C. To see this, just replace the computer with Searle in the Chinese room

1. that is, some of the Chinese symbols are coming to Searle from the robot's video camera

2. and the symbols he sends back are going to its motors

3. Searle would be manipulating these symbols and have no idea what's going on with the robot

III. The Brain Simulator Reply

A. instead of a program like Schank's, we write one that simulates the nerve cell firings in somebody's brain when that person understands a story in Chinese

B. we could even throw in parallel processing

C. such a simulation would understand Chinese

Searle's Response

A. this is an odd reply for Strong AI enthusiast or indeed for any sort of functionalist to make

1. the whole idea of Strong AI is supposed to be that we don't need to know how the brain works to know how the mind works (243-44)

2. for Strong AI, the mind is supposed to be like the software (244)

3. if we need to know how the brain works to do AI, why bother with AI?

B. Anyway, this simulation wouldn't necessarily give us understanding

C. to see this, Searle has us change the Chinese room thought experiment in the following way:

1. instead of shuffling symbols around, the man has to turn valves on and off in a complex maze of water pipes
   a. the connections of pipes correspond to synapses
   b. Turning on a valve would then correspond to a synapse firing
2. the man inside still receives Chinese symbols, but when he does he looks them up in rule book that tells him in English which valves to turn
3. as a result of turning the right valves, the system gives an output of Chinese symbols
4. in this new system, neither the man nor the water pipes understands Chinese

D. it does no good to object that the whole system understands Chinese, because we could have the man inside memorize the structure of pipes, etc. -- he'd have the whole system in his head and still not understand Chinese (244)

E. the problem with the brain simulation is that it's simulating the wrong thing about the brain:

1. it's simulating only the formal properties of the neural firings
2. it’s not simulating what really matters about the brain, “namely its causal properties, its ability to produce intentional states” (q.v.)

IV. The Combination Reply: put the first three replies together: a robot with a computer that is running a simulation of the nerve cells firing, who acts just like us, and is a unified system

Searle's Response

A. here he would agree that we would accept that the robot had

intentionality as long as we know nothing more about it (q.v.)

1. we would grant that as long as the robot just acted like us (244-45)
2. if it behaves like us, the rest is irrelevant (245)

B. However, this case doesn't help Strong AI

1. for Strong AI, having the right program, the right inputs, and the right outputs is sufficient for intentionality
2. but in this case, the reason we ascribe intentionality to the robot has nothing to do with its program:
   a. if it looks and acts like us, we assume, until it is shown otherwise, that it has mental states like us
   b. if we knew how to explain its behavior without these assumptions, especially if we knew it were running a program, we would not attribute intentionality to it (245)

C. suppose we replaced the computer in the robot with a man manipulating symbols in accordance with rules -- and we all knew this

1. we would then regard the robot as just a dummy
2. the only intentionality would be in the man and he would not see what comes into the robot's eyes, he would not intend to move its arms, or understand what is said to it or by it
3. neither would the system of the man-inside-the-robot

D. to see his point, contrast this last case to cases where we ascribe intentionality to animals.

1. we do this
   a. because we cannot make sense of the animal's behavior without assuming it has intentionality, and
   b. because it's made out of the same sort of stuff we are, “the same causal stuff” (245, q.v.)
2. if a robot acted like we do (246, q.v.)
   a. we would make similar assumptions about it and ascribe intentionality to it unless we had some reason not to

b. but if we knew its behavior were the result of a formal program, and that the causal properties of its physical substance were not relevant, we'd give up the assumption of intentionality (246, q.v.)

V. The Other-Minds Reply

A. you only know that people understand Chinese by their behavior

B. so if a computer passes the same behavior test, you must attribute understanding to it

Searle's Response

A. the problem is not how Searle knows that other people have mental states but what he is attributing to them when he attributes mental states to them: it’s not just computational processes

B. it's no answer to his argument just to fake anesthesia

VI. The Many Mansions Reply: Searle's whole argument applies to our current technology. But someday we will be able to build devices that have the causal properties that Searle says are needed for intentionality, and then we will have AI.

Searle's Response

A. this argument changes the definition of strong AI to whatever it is that produces and explains intelligence

B. the original claim of strong AI was that the manipulation of formal symbols alone could produce and explain intelligence -- this reply abandons that claim (246, q.v.)

<u>Meaning</u>

I. Here he returns to the questions at the beginning of the paper about (246)

A. what is needed for understanding, say, English

B. whether it could be given to a machine

II. Searle sees no reason we could not give understanding to a machine (247)

A. there's a sense in which we, after all, are machines

1. however, it is not because of some formal program that we understand English or Chinese (q.v.)
2. rather, it's because we are a certain sort of biological organism

B. he imagines that other things made of other stuff could also have intentionality

C. however, he does not think that we could give intentionality to a machine that is understood purely as something that manipulates formal symbols.

1. The only causal powers they have is the power to produce the next step in the formalism
2. other causal properties are irrelevant, since we can always put the same formal system in a different physical realization in which these other causal powers are absent

<u>Conclusion</u>

I. Could a machine think?

Yes -- we are just such machines

II. But could a man-made machine think? (247)

A. again, yes (247-48)

B. if we can duplicate the causes, we can duplicate the effects (248)

III. But could a digital computer think?

If by a digital computer we mean anything that could instantiate a program, the answer is yes, since we could be described that way.

IV. But could a digital computer think solely in virtue of running the right program?

No.

V. Why not?

A. because formal symbol manipulations have no intentionality

1. they are meaningless
2. they aren't even really symbols -- at least, not to the physical

device that manipulates them

3. they have syntax, but no semantics

4. the meaning or intentionality is only in the mind of the programmer or the person who uses the machine

B. it was the point of the thought experiment to show this

1. we put something in the system -- a human being -- that had real intentionality

2. the program added nothing to his intentionality -- it doesn't help him to understand Chinese

C. the very feature of AI that made it attractive, the distinction between the program and its physical realization, is what leads to its downfall. There are three problems with the analogy between mind/brain and program/hardware: (248)

1. all sorts of crazy stuff could instantiate programs that you wouldn't want to ascribe intentionality to

a. Weizenbaum's roll of toilet paper and pile of stones, etc.

b. Again, for Searle, this is “the wrong kind of stuff;” not the stuff that has the "causal powers" of brains (248-49)

2. programs are purely formal but intentional states are not -- they are defined in terms of their contents (249)

3. mental states are produced by the brain but programs are not in the same way produced by the computer

VI. Why then have people believed otherwise? That programs could constitute minds?

A. Searle can't really say (249)

1. after all, nobody ever thought they could get burned by a computer simulation of fire

2. so why would they think a simulation of thinking could actually understand anything?

3. although some may think that the hardest thing for AI is for a computer to feel pain or fall in love, these things are neither

harder nor easier than understanding

4. simulation not the same as duplication

B. but he does offer three reasons for the illusion that AI produces and explains mental phenomena

1. First reason: ambiguity of concept of "information processing" (249)

   a. some people think that both the brain and computers do something called “information processing,” but that other things computers simulate, like storms, don’t do information processing

   b. however, people and computers are processing information in two different senses: the computer is only manipulating formal symbols (250)

   c. the idea of information processing leaves us with a dilemma: (q.v.)

      1.) either we construe information processing such that it implies intentionality or we don't (q.v.)

      2.) if we do, a computer isn't processing information, just manipulating formal symbols

      3.) if we don't, then a computer would process information only in the same sense that a typewriter or thermostat does. It is up to the outside observer to interpret this information

2. Second reason: residual behaviorism

   a. that is, we are tempted to attribute mental states to something that can mimic our input/output patterns

   b. this is what's wrong with the Turing test (q.v.)

3. Third reason: residual dualism

   a. that is, the brain doesn't matter

   b. in AI and in functionalism, only the program matters

   c. we could even run the program on a soul (251)

   d. unless you believe the mind can really be separated from the brain, you couldn't hope to reproduce it just by writing a

program

e. it may not be substance dualism, but it's a form of dualism nevertheless (q.v.)

VII. Could a machine think?

A. Only a machine could think -- only a machine with the same causal powers as a brain

B. This is what's wrong with Strong AI

1. it's just about programs

2. but programs are not machines

C. intentionality is causally dependent on the biology of the brain

1. you couldn't get intelligence out of brain simulation any more than you could get lactation or any other biological phenomenon out of simulation

2. the belief otherwise reflects a kind of dualism

D. it's no help to be told the brain is a digital computer -- it is because everything is

E. the point is that the brain's causal capacity to produce intentionality does not consist in running a program (251-52)