PH1102E/GEK1067

Week 9

Can computers think?

Computational Theory

A Chinese speaker's brain is a computer which, by following a biologically-encoded program for the manipulation of intrinsically meaningless neurons, manifests an understanding of Chinese.

This is the computational theory of cognition (as it pertains to Chinese language comprehension).

Chinese Room Argument

1. The person in the Chinese Room is computationally equivalent to an ordinary Chinese speaker, in all respects relevant to Chinese language comprehension.
2. According to the computational theory of the mind, anyone who is computationally equivalent to an ordinary Chinese speaker (in all relevant respects) understands Chinese.
3. So if the computational theory is true, it follows that the person in the Chinese Room understands Chinese. (1, 2)
4. But the person in the room does not understand Chinese.

Therefore, the computational theory is false. (3, 4)

I. The computational theory of thought

What distinguishes intelligent beings---beings capable of thought---from other beings? Here is one: intelligent beings are, in effect, computers that run advanced software which allows them to navigate through their environments effectively, identify and solve a whole variety of problems, and, in some cases, communicate by means of language. The difference between "higher" and "lower" animals is that the "higher" ones (like humans) have more powerful hardware (bigger brains) and more sophisticated software (fancier neural networks). The difference between intelligent and unintelligent beings is that the unintelligent beings have no software at all (rocks, tables, pens, etc.), or else only primitive software (oysters, trees, bacteria).

Why the analogy with computers? Well, remember that the cognitive features of the mind---the features that relate to belief, desire, thought, understanding, etc.---are those that mediate environmental input and behavioral output. You receive input through your eyes, ears, nose, tongue, and skin. Events occur in your environment that stimulate these sense organs, causing changes in the rods and cones embedded in your retina, or in the tiny hairs inside your ears, etc. But these changes in your sense organs are not enough to elicit an intelligent response from you. For that, your sense organs must send some kind of signal to your brain. Your brain receives these signals, and does something with them---exactly what, we are still discovering---and eventually sends signals via efferent nerve fibers to your muscles and glands, causing your body to move in certain ways. These motions are the response to the original stimulus.

For example, light reflecting from a tiger enters your eyes, stimulating a certain pattern of rods and cones in your retina. These rods and cones send an electrical impulse to your brain, via the optical nerve. Something (we're not sure exactly what) happens in your brain, and, quickly, your brain sends out electrical impulses to the muscles in your legs, causing them to contract and relax in a coordinated pattern---i.e., causing you to run.

This is a crude story, but it captures the basics of what goes on in you when you react to an environmental stimulus.[1] The point is that we can describe the situation in terms of inputs and outputs. This is what suggests an analogy with computers, which we also describe in terms of inputs and outputs. It's just that a computer typically takes keystrokes (or mouse clicks) as input, and gives electronic text (or something similar) as output. Also, what mediates input and output in an ordinary computer is not a brain, but a network of electrical circuits. But many people think that these differences between a human being and an ordinary desktop computer are mere differences of degree. Your desktop has a mouse, keyboard, webcam, microphone, and internet cable for input. You have eyes, ears, nose, tongue, and skin. Your desktop has an LCD monitor, speakers, and an attached printer for output (as well as the internet cable). You have arms, legs, vocal cords, and other body parts. Your desktop has a motherboard with a CPU; you have a central nervous system with a brain.

The computational theory of thought says that the "something-we're-not-sure-exactly-what" that takes place in your brain between the time you see the tiger and the time you start to run is some kind of computation. It is a computation that we can equate with a certain mental act: a decision on your part to flee. This doesn't mean that we carry out the computation consciously. The fact that we don't know

[1] A less crude story would also take into account inputs from one part of your brain to another part of it, or, from one part of the computer's motherboard to another. Not all inputs come from outside the system, and not all outputs translate into overt motor behavior (or screen output).

how our brains work does not prevent us from thinking, any more than the fact that we don't know how our lungs work prevents us from breathing.

If this picture of intelligence is correct, then there should be nothing in principle to prevent us from building artificial intelligences. It is just a matter of constructing a machine that can receive a sufficiently wide range of input, and process it as quickly and effectively as a human brain processes the input that it receives. The fact that the machine will be made of metal and silicon instead of living cells is irrelevant. All that matters for intelligence is that you have the right input-output architecture.

At least, that's what many people think. But not John Searle. Searle thinks that there is more to having a mind, and, in particular, more to being intelligent (in the broad sense defined earlier) than having an appropriately sophisticated input-output architecture.

II. The Chinese Room argument

Specifically, Searle thinks that there is more to thinking (deliberating, meditating, deciding, understanding, etc.) than running a digital computer program. This is what Searle means when he says that computers can't think. He means that even if we humans are, in some sense, computers, it is not just by virtue of being computers (and running various computer programs) that we are able to think. The fact that we can think proves that we are not just computers---not mere runners-of-computer-programs. And the fact that the machines built by scientists pursuing AI are just computers proves that those machines cannot think. Such, at any rate, is Searle's position.

The "Chinese Room argument" is Searle's attempt to prove that thinking is not purely a matter of computation. To understand the argument, it will help to give explicit definitions of the main concepts the argument involves, especially the concept of a "computer program." Here goes:

> For Searle, a **digital computer** is anything that runs a computer program.
>
> A **program** (as Searle defines it) is any set of **simple rules** for manipulating **symbols**.
>
> A **simple rule** is a rule that requires no intelligence or creativity to apply---a rule that even a mindless mechanism can apply correctly.
>
> A **symbol** is any intrinsically meaningless thing that we can create, erase, copy, store, transmit, or otherwise manipulate.

Note that there is a difference between a "set of simple rules" and a "simple set of rules." A set of simple rules is a collection of rules, where each rule in the collection is, taken on its own, simple, in the above-defined sense. But you could have a set that contained hundreds, or thousands, or millions of simple rules. This would still be a set of simple rules, but it would not necessarily be a simple set of rules. The set itself might be quite complicated, inasmuch as it contains a very large number of rules, many of which might refer to other rules in the set. For example, you might have a rule that says: "Close circuit Number 2, and then proceed to rule #48501," where rule #48501 is some other simple rule in the set. You can also write the rules so as to create operational "loops": one rule referring to another rule, referring to another rule, referring back to the first rule. In this way, a system can quickly become extremely complex, even though the individual rules it follows are all very simple.

Anyway, Searle's main claim is that there cannot be a computer program of which you can say: "Anything that runs this program thinks." In other words, running a computer program, no matter how sophisticated (with loops upon loops upon loops), is never by itself sufficient for thinking.

What Searle actually argues for is a more limited claim. He argues there is no possible computer program of which you can say: "Anything that runs this program speaks and understands Chinese." Of course, speaking and understanding Chinese are just two specific forms of cognitive activity. But it is pretty clear that by adjusting his argument, Searle can put together a perfectly analogous argument for any other form of cognition. Certainly he could recycle the argument in terms of any other language---he just takes Chinese as an example. So we could have the Japanese Room argument, the English Room argument , the Cherokee Room argument, etc.

Given that his objective is to show that running a program is never sufficient for understanding Chinese, all Searle has to do is to show that someone or something that had no understanding of Chinese could nevertheless run whatever programs a Chinese speaker runs when understanding Chinese. If a non-understander of Chinese can run all the programs that a Chinese understander runs, then, obviously, running those programs is not by itself enough for understanding Chinese. After all, if it were enough, then anyone or anything that ran the programs would understand Chinese.

Searle develops his argument against the background of a thought-experiment: a fanciful but perfectly conceivable, and theoretically possible, scenario. This is supposed to be a scenario in which someone (actually, Searle) runs all the programs that ordinary Chinese speakers run when understanding Chinese, without himself understanding Chinese. It's supposed to be an example of someone who fails to understand Chinese, despite running every computer program that could possibly be relevant to understanding Chinese. (It may sound strange to describe an ordinary Chinese speaker as "running a program," but remember that we are using the term "program" in a broad sense, to cover any process or activity that we can construe as arising from the application of simple rules.)

Imagine that Searle is locked in a room. The only other things in the room are a large number of books full of two-column charts, and some large baskets full of slips of paper with markings on them. There is only one door to the room (locked), and it has a letter-slot in it. Through this slot, Searle receives a message written in English (this is the first and last English message that will come throught the slot). The message explains to Searle that he has been recruited to take part in an experiment, and that if he just does what the note instructs him to do, he will be released in due course. The note then tells him that soon, slips of paper with markings on them will start coming through the slot. Each time a slip comes into the room, Searle is to flip through the books until he finds a chart with an identical marking in column A. His job then is to rummage through the baskets to find a slip of paper with a marking that identical to the corresponding mark in column B of the chart. Once he finds the right slip of paper, he feeds it out through the door slot.

Soon slips of paper start coming through the slot, and Searle does what he is told: he looks up the marks on each slip in the books he has been provided, finds the corresponding marks indicated on the chart, looks through the baskets for a slip of paper bearing these marks, and, when he finds it, feeds it out through the slot.

What Searle does not realize is that these "marks" are actually Chinese sentences: the marks on the slips he receives are Chinese questions, and the marks on the slips he gives out are Chinese answers. He also doesn't know that the manuals he consults were written in such a way that the slips of paper he feeds

out through the slot are perfectly natural and sensible answers to the questions written (in Chinese) on the slips of paper that come in through the slot. The manuals are so cleverly composed, in fact, that if the people conducting the Chinese Room experiment invite a Chinese speaker to write down questions (in Chinese) and feed them into the room through the slot, this person will have no idea that the person inside the room who feeds answers back out through the slot is not an ordinary speaker of Chinese. To an outsider, it is just like having a conversation with an ordinary Chinese speaker, just by passing notes instead of face-to-face.

With this thought-experiment as background, Searle advances an argument that is supposed to prove that running a program can never be enough on its own for understanding Chinese. The argument goes like this:

> *The person in the Chinese Room runs whatever programs an ordinary Chinese speaker runs when speaking and understanding Chinese. In other words, to whatever extent we can construe an ordinary Chinese speaker as following simple rules (for taking Chinese symbols as input and giving Chinese symbols as output), we can construe the person in the Chinese Room as following the same rules. And since running a computer program is just a matter of following simple rules for the manipulation of symbols, it follows that in whatever sense an ordinary Chinese speaker runs a computer program when having a conversation in Chinese, the person in the Chinese Room is running a computer program just like that.*
>
> *Now, some people (proponents of "strong AI") say that speaking and understanding Chinese is just a matter of running the right kind of computer program. If this is correct, then surely an ordinary Chinese speaker runs the right kind of computer program; after all, an ordinary Chinese speaker does speak and understand Chinese. So, if it's enough for understanding Chinese that you run the right kind of program, then it's enough for understanding Chinese that you run the kind of program that an ordinary Chinese speaker runs.*
>
> *But remember: the person in the Chinese Room runs the same kind of program that an ordinary Chinese speaker runs. So, if it's enough for understanding Chinese that you run that kind of program, it follows that the person in the Chinese Room understands Chinese.*
> *But the person in the Chinese Room is Searle, and Searle doesn't understand Chinese! So we have to conclude that running the right kind of program (or any kind of program) is not enough for understanding Chinese.*
>
> *To summarize: if running the right kind of computer program were enough for understanding Chinese, then the person in the Chinese Room would understand Chinese, since he is computationally equivalent to an ordinary Chinese speaker. But the person in the Chinese Room does not understand Chinese, since the person in the Chinese Room is Searle. Therefore, running the right kind of computer program is not enough for understanding Chinese. To understand Chinese, you must be more than just a computer.*

III. Objections and replies

There are two main objections people have raised against Searle's argument, the "robot objection" and the "system objection."

A. Robot objection

Proponents of the robot objection deny that the person in the Chinese Room is computationally equivalent to an ordinary Chinese speaker. After all, the manuals only permit Searle (in the Chinese Room) to reproduce one aspect of a true Chinese speaker's behavior: they only permit him to give appropriate answers to questions posed in Chinese. But an ordinary Chinese speaker can do far more than this. He can ask questions of his own; he can respond appropriately to commands, like "Draw a banana" or "Take out the garbage" or "Tell us how you feel." In general, language is not something we use in isolation from the other things we do: it is very much connected with a whole range of behavior. I read a sign that says "MRT STATION CLOSED," and this leads me to make alternate travel arrangements. On Thursday I read an advertisement that says "ALL ITEMS HALF PRICE THIS WEEKEND AT MEGAMART," and, on Saturday, I go shopping at Megamart. We also use language in non-social ways; for example, we talk to ourselves (silently or out loud), we keep diaries, etc.

But as Searle describes it, the person in the Chinese Room does none of these things. All that his manuals allow him to do is answer questions, and a very limited range of questions at that. To reproduce all the computation that takes place in someone who actually understands Chinese, we would have to do two things.

First, we would have to find a way to expand the range of inputs and outputs available to Searle. For example, he would have to receive some analog of sensory input, and be capable of returning some analog of motor output.

Second, we would have to add a lot more manuals, containing instructions telling Searle what to do when he receives one of these new inputs.

How might we arrange this? Well, we could build a robot with a digital camera for eyes, and have it send the images it takes to a monitor in the Chinese Room. Then Searle would be instructed to press certain buttons whenever he sees a sign that reads (in Chinese) "MRT STATION CLOSED." Unbeknownst to Searle, by pressing these buttons, he remotely navigates the robot away from the MRT station, and towards a nearby taxi stand. However, with this arrangement we run the risk that Searle will begin to figure out what's going on. For example, he might notice that whenever he sees a sign with the Chinese word for "CLOSED" on it, his instructions tell him to press certain buttons, and that whenever he presses these buttons, the camera whose images he's seeing moves away from the sign. Based on this, he might conjecture that the marks on the sign are actually words that mean *closed*. (We'd also have to be very careful not to let the camera point at any bilingual signs.) To prevent this sort of thing, we could set the system up so that all Searle receives from the digital camera is raw image data -- strings of 0's and 1's, let's say. Searle's manuals then tell him to press such-and-such buttons based on which strings of numerals come up on his monitor.

The point is that in order to make the person in the Chinese Room computationally equivalent to an ordinary speaker of Chinese, we would have to provide him with many more instruction manuals than we find in Searle's original thought-experiment, and we would have to supply him with far more input---and allow him to give far more output---in accordance with these expanded manuals (or whatever plays the role of the manuals in the enhanced thought-experiment). The person in Searle's Chinese Room does not, it is true, understand Chinese. But that is only because the person in Searle's Chinese Room is not computationally equivalent to an ordinary Chinese speaker: he does not run all the programs that an ordinary speaker of Chinese runs when using the Chinese language, but only a subset of them.

How does Searle reply to this? He replies by saying, in effect: “Fine, add as many manuals as you like. Expand the range of available inputs and outputs as much as you like. Put the Chinese Room in the head of a robot, and integrate my activities inside the Chinese Room with the robot’s sensors and motor control systems. None of this will bring me an inch closer to understanding Chinese. Yet now, even by your own standards, I am computationally equivalent to an ordinary Chinese speaker. So, if understanding Chinese were merely a matter of running the right computer program, I should now be understanding Chinese. But I still don’t.”

B. System objection

Searle’s reply to the robot objection seems sound. It is hard to see how Searle could suddenly come to understand Chinese just be being provided more manuals, and some new---but equally unintelligible to him---forms of input and output.

A different objection can pick up where the robot objection left off. Suppose we have placed the Chinese Room (with Searle and all the manuals, buttons, etc.) into the head of a robot, and connected it with the robot’s sensory apparatus and mechanical body so as to allow Searle---unknowingly---to navigate the robot through its environment. And suppose that the manuals have been augmented so that by following their instructions on how to react to any given input, Searle causes the robot to behave in a way that is entirely indistinguishable from an ordinary Chinese person. It responds to advertisements, writes poetry, keeps a diary, engages in conversation, asks for and follows directions, sends email to its friends, maintains a weblog, etc., etc., etc. We can imagine that the robot’s body is so cleverly constructed that it even looks like an ordinary human being. (We would have to shrink Searle and the Chinese Room down to scale to fit it into a human-scale robot head. As an alternative, we could have Searle controlling the robot by remote control---it doesn’t matter, as long as Searle remains unaware of the consequences of his rule-governed activities inside the room.)

Now, instead of focusing on Searle, stuck inside the room in the robot’s head, let’s focus on the system as a whole. Granted that Searle himself does not understand Chinese, we can still maintain that the whole system of which he is but one part does understand Chinese. At least, we have no better reason to doubt this than we have to doubt that an ordinary Chinese speaker understands Chinese.
Searle’s reply to the system objection is puzzling. He actually seems to give two replies, although neither is very clear.

His first reply is that “there is no way that the system can get from the syntax to the semantics.” But if this is Searle’s reply, he is simply begging the question against his opponent. By this, I mean that he is taking for granted the very thing that he is supposed to be proving.

To see why this is so, think of what Searle means by “syntax” and “semantics.” For a system to “have a syntax” is basically just for it to manipulate symbols on the basis of simple rules (simple, in the sense explained earlier). For it to “have a semantics” is for it to make meaningful use of language---the sort of use that implies actual understanding. So, when Searle says that “there is no way that the system can get from the syntax to the semantics,” what he is saying is that there is no way that the system can make meaningful use of language just by manipulating symbols in accordance with simple rules. But that is just another way of saying that a system cannot understand language just by running a computer program, which is the very thing that Searle is trying to prove! The objection was that the system as a whole understands Chinese; Searle’s reply is that it doesn’t, because it’s just a computer, and computers

can't think. He seems to have temporarily lost sight of the fact that "computers can't think" is something that he is supposed to be proving, not taking for granted.

However, Searle also seems to have a different reply to the system objection. He seems to reason as follows: "You grant that the person in the Chinese Room does not understand Chinese. But neither do the manuals understand Chinese, or the slips of paper, or the buttons, or levers, or anything else that is in the room. The digital camera that serves as the robot's eyes doesn't understand Chinese, and neither does any other part of the robotic mechanism. But if no part of the system understands Chinese, how can it be that the system as a whole understands Chinese?"

I don't know if this is how Searle intends to reply to the system objection; as I said, his reply is quite brief and dismissive. But if this is Searle's intended reply, I think it is a bad one.

It is true that the Chinese Room---and the Chinese Robot---consists of parts no one of which understands Chinese. *But the same is true of an ordinary Chinese speaker*. An ordinary Chinese speaker, according to Searle, is a biological organism, consisting of around 50 trillion cells. None of these cells speaks or understands Chinese. So the fact that none of a system's parts understands Chinese gives us no reason to think that the system as a whole does not understand Chinese.

Against this, one could reply that maybe an ordinary Chinese speaker does not consist only of cells; maybe he or she contains a mind that is not composed of cells or anything physical. That is a possible line to take, but it does not seem to be available to Searle, who emphasizes that (on his view) thought is an essentially biological phenomenon. Anyway, if Searle does want to claim that an ordinary Chinese speaker has some special part that understands Chinese, but that does not consist of non-Chinese-understanding parts, he must give us some reason to believe that this is true.

To be fair, Searle is right to point out that it is amazing and, at least when you first think about it, implausible to suppose that a system like the one he describes (or the more elaborate "robot" system) could understand Chinese, or any other language. What he fails to recognize is that it is equally amazing and *prima facie* implausible to suppose that a human brain could understand Chinese! The Chinese Room is just a bunch of non-Chinese speaking things behaving in accordance with simple rules. But a Chinese brain is also just a bunch of non-Chinese speaking things behaving in accordance with simple rules. So, whatever basis we have for believing that the computation that takes place within a Chinese brain is sufficient for an understanding of Chinese, we have an equal basis for thinking that the computation that takes place in the Chinese Room is sufficient for an understanding of Chinese.

IV. Appendix on "syntax" and "semantics"

Searle spends a lot of time talking about "syntax" and "semantics." By "syntax," he means the non-meaningul aspects of language, such as the shape or sound of the words, the rules for combining the words into allowable (i.e., grammatical) phrases and sentences, etc. The syntactic properties of a language are the properties that it would have even if none of the words of the language had any meaning. Even an invented nonsense language has a syntax: the nonsense words are spelled a certain way, sound a certain way when spoken, and are combined into meaningless phrases and sentences in accordance with specific rules. A language has a "semantics" or semantic properties only insofar as the words, phrases, and sentences of the language have some meaning.

Searle claims that a computer can have syntax but not semantics. By this, he means that a computer can combine and manipulate words in accordance with the rules of grammar etc., but cannot mean anything by the words it combines and manipulates. Unfortunately, Searle sometimes gets sloppy, and says things that suggest that he is arguing that a computer couldn't possibly understand language because it has only syntax, and no semantics. This is not really Searle's argument, but if it were, it would be a lousy one. He would be reasoning in a circle. Again: to say that a computer has no semantics just is to say that it can't understand language. So, to argue that "a computer can't understand language because it has no semantics" is equivalent to arguing that "a computer can't understand language because it can't understand language."

In my lecture, and in these notes, I have given what I take to be the proper interpretation of Searle's argument, which does not commit this error of circular reasoning.