The Chinese Room Is A Trick

Peter Kugel,Computer Science Department
Boston College,Chestnut Hill, MA 02467-3808,USA

Abstract: In an attempt to convince us that computers can't have mental states, Searle (1980) asks us to imagine a "Chinese room" that imitates a computer that seems to understand Chinese. Then he asks "Where is the understanding in the room?" It's a trick. Like the magician who makes coins disappear, Searle is asking us to look in the wrong place. There is no understanding in the room because its computer imitation is too weak to let it in. Improve the accuracy of its computer imitation and it can handle understanding. And its Chinese can improve too. Abracadabra!

In his article "Minds, Brains, and Programs", Searle (1980) argues that, although computers can *seem* to have mental states, they can't *really* have them. To support his claim, he asks us to imagine a "Chinese room" that (1) simulates what computers can do (2) to produce the appearance of understanding Chinese (3) without having anything that corresponds to "understanding" inside.

Most of those who have argued against Searle – and there have been many – have accepted (1) and (2) and looked for "understanding" somewhere in the room. That's a mistake because Searle is right. It's not there.

It's not missing because computers can't have it. It's missing because claim (1) – that the room can do everything computers can – is false. The room's computer-imitation is so poor that claim (2) – that the room can do a good job of faking fluent Chinese – is also false. To see how limited its (apparent) understanding of Chinese is, consider the following dialog, translated into English for my (and, presumably, most readers') convenience:

Me: "From here on in I'm going to use the word 'bad' to mean 'good' as it does in some contemporary American slang. Got it?"
Room: "Yes."
Me: "Would you say that an A was a bad grade?"
Room : "No." (Gotcha!)

The reason the room can't handle this sort of "conversation" is that it has no way to remember what happens. Its script cannot write on itself. (It can only produce Chinese squiggles, but the script is written in English squiggles so Searle can read it.)

If we allowed the script to write on itself, it could "remember" and it could change what it does as a result of what it "experiences". That would make it a lot more complicated but it is, I claim, precisely what it needs to achieve intentionality. And intentionality, according to Searle (1980) and Brentano (1924), is what distinguishes mental states from physical ones. And, if it had the machinery to produce intentionality, it could be made to understand.According to Searle (1980), internal states have intentionality if they are "directed at or about objects and states of affairs in the world." What this means, it seems to me, is

that they can change in certain ways when what they are "directed at" changes. For example, my thoughts about the Chinese room have the intentionality that the Chinese room's "thoughts" about me lack because my thoughts about the room can change appropriately when I learn something new about it. The room's thoughts about me lack intentionality because they cannot change appropriately or I would not have "caught it" in my "bad" conversation.

That's also true of other mental states. What gives my belief that "All swans are white" intentionality, for example, is that it can change appropriately after I see a few black swans, perhaps to "All swans are black or white."

Not all changes produced by experience are sufficiently rich to count toward intentionality. A supermarket scanner that changes its internal state in response to the UPC code on a bag of cookies lacks intentionality, whereas a Chinese child that changes its internal state in response to the Chinese translation of "I brought home a bag of cookies" has it, as any parent knows.

The Chinese room and the scanner lack intentionality because they only have what I have called "fake intelligence" (Kugel 2002) – the ability to apply the rules (programs, scripts) they have been given. In contrast, a child has intentionality because it can adjust, or even originate, its own rules on the basis of its experiences.

It is not easy to make the distinction between changing your program and changing your data precise. But, if philosophers could clarify it (and I believe they can) and if computer scientists could implement programs based on their accounts, I would be willing to call computer states in those programs "mental" – if they could adjust their computer's programs flexibly enough.

Searle might not. He might still object that the Chinese room, changing its programs in response to its experiences, lacked intentionality because Searle, inside the room, lacked it. Well, he might, but the intentionality achieved by a human mind does not percolate down to the individual neurons and the intentionality of the computer need not work its way down to the machine's central processing unit, which is the role that Searle plays in the Chinese room.

Searle might also complain that the resulting "understanding" would not feel, to the computer, the way understanding does to a person. How important such "feel" is is an interesting question. It is bears some relation to the question "Can people who were born blind, understand the word 'yellow'?" I believe that, to some degree, they can. Searle might disagree.

Bless him. If using the same term for both kinds of states bothers Searle, I would be willing to limit my use of the term "mental states" to refer to what human beings have (or at least to what I have because I can't be sure that yours "feel" the way mine "feel") and call what computers have "intentional states".

But I believe that Searle is right when he argues that machines will have to have intentional states before they can be really intelligent. The ability to remember what happened, and to change the way you think in response, is crucial to both intelligence and understanding. You understand this commentary

to the degree that it changes what you can do – paraphrase it, use its ideas at a cocktail party, argue against them in a discussion, use them to produce programs that understand, or what have you. The programs that Searle (rightly) criticizes lack this ability.

I want to thank Searle for reminding us that this ability is important. But you have to wonder how we could have forgotten it in the first place.

References

Brentano, F. (1924) *Psychologie vom empirischen Standpunkt*, translated as *Psychology from an Empirical Standpoint*. Routledge and Kegan Paul.

Kugel, P. (2002) Computing machines can't be intelligent (…and Turing said so), *Minds and Machines* 12(4): (to appear in November, 2002).

Searle, J.R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417-57.