A Modal Defence of Strong AI

Steffen Borge

John Searle has argued that the aim of strong AI to create a thinking computer is misguided. Searle's "Chinese Room Argument" purports to show that syntax does not suffice for semantics and that computer programs as such must fail to have intrinsic intentionality But we are not mainly interested in the program itself, but rather the implementation of the program in some material. It does not follow by necessity from the fact that computer programs are defined syntactically that the implementation of them cannot suffice for semantics. Perhaps our world is a world in which any implementation of the right computer program will create a system with intrinsic intentionality, in which case Searle's "Chinese Room Scenario" is empirically (nomically) impossible. But perhaps our world is a world in which Searle's "Chinese Room Scenario" is empirically (nomically) possible, and the silicon basis of modern-day computers is one kind of material unsuited to give you intrinsic intentionality. The metaphysical question turns out to be a question of what kind of world we are in, and I argue that in this respect we do not know our model address. The "Model Address Argument" does not ensure that strong AI will succeed, but it shows that Searle's challenge to the research program of strong AI fails in its objectives.

Alan Turing suggested that if a system has decent conversational skills, it has a mind (Turing, 1950). Turing envisaged an imitation game played by three systems, a human, a machine and an interrogator communicating with each other via a teleprinter where both subjects present themselves as the human. If a machine passes the Turing test (the interrogator cannot tell them apart), it should be credited with having a mind. The main thesis of strong Artificial Intelligence (AI) is that behavioural (and functional) features of a system suffice for having mental states. John Searle's Chinese Room Argument aims to show that this thesis is false and that strong AI as a research program is moribund.

Searle invites us to consider the following thought-experiment (Searle, 1980, pp. 417-418). Searle, a non-Chinese speaker, is locked in a room together with an instruction manual in English that pairs Chinese symbols with other Chinese symbols. The book enables Searle to respond to any incoming string of Chinese characters by sending out another string of Chinese characters. Unbeknownst to Searle the incoming strings of symbols represent questions, while the outgoing strings of symbols represent answers to these questions. Searle's performance is, for the people outside the room, indistinguishable from that of a native Chinese speaker. Searle passes the Turing test for speaking Chinese. But Searle cannot be said to speak Chinese since he has no understanding of the activity he is engaged in.

By parity of reasoning the same holds for the machines in Turing's imitation game. The proponents of strong AI that claim that having a mind is nothing but the implementation of a certain computational program—e.g. the right functional organisation with the right input-output relation—are wrong since the Chinese Room shows that such an implementation does not suffice for Searle to know Chinese nor would it suffice for a computer working with a formal program to have understanding. The argument is summed up in the following way (Searle, 1994, pp. 39):

> Premise 1. Syntax is not sufficient for semantics.
>
> Premise 2. Computer programs are entirely defined by their formal, or syntactical, structure.
>
> Premise 3. Minds have mental contents; specifically, they have semantic contents.
>
> Conclusion: No computer program by itself is sufficient to give a system a mind. Programs, in short, are not minds, and they are not by themselves sufficient for having minds.

What it is to have a mind is not captured by mere extrinsic features like functional input-output relations or performance. Such machines would lack what Searle calls *intrinsic intentionality* and thus have no minds.

But what is intrinsic intentionality and how can we tell when someone has it? On the former question Searle answers that it is what makes brains think while mere computer simulations of our neural brain structure fail to do so.

> So I am just stipulating that by "intrinsic intentionality" I mean the real thing as
>
> opposed to the mere appearances of the thing (*as-if*), and opposed to derived forms of intentionality such as sentences, pictures, etc. (Searle, 1995a, p. 80).

There is no need to fault Searle for just stipulating this. Even though the notion of "intrinsic intentionality" is rather vague, we all have a certain intuitive grasp of the distinction he is making. I—and hopefully you—have intrinsic intentionality, but when I describe my car as wanting more gasoline I speak merely figuratively. My car has nothing but as-if intentionality and the fuel gauge that tells me how much gasoline there is in the tank can only be ascribed derived intentionality. The fuel gauge's intentionality derives from our usage of such devices to tell us something about the world. Elsewhere Searle gets somewhat more specific about intrinsic intentionality.

> My own view is that *only* a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains (Searle, 1980, p. 424).

The key word is "same causal powers" and not functional organisation or behavioural input-output relation. What proponents of strong AI ignore, according to the Searlean picture, is that the matter matters when it comes to the question of having a mind.

Searle is right, I believe, in pushing the point that for a system to have a mind it must have intrinsic intentionality—whatever that is. Our concept of what it is to have a mind entails among other things that the system has intrinsic intentionality. If we knew that the implementation of a computer program did not cause intrinsic intentionality, then we would know that that machine did not think. But this is the bone of contention.

Recall Searle's argument that because computer programs are defined by their formal or syntactical structure they do not have minds—e.g. intrinsic intentionality. But we are not mainly interested in the program itself but rather the implementation of the program in some material. It does not follow by necessity from the fact that computer programs are defined syntactically that the implementation of them cannot suffice for semantics. Consider David Chalmers' parody of Searle's argument (Chalmers, p. 327):

> Premise 1. Recipes are syntactic.
>
> Premise 2. Syntax is not sufficient for crumbliness.
>
> Premise 3. Cakes are crumbly.
>
> Conclusion. Therefore, implementing a recipe is insufficient for a cake.

The conclusion, of course, is false. For Chalmers' parody to work, however, it would have to concede the Searlean point that matter matters or that matter might matter. If a recipe for a cake was just defined formally – that is, if the recipe did not prescribe what materials one should use in implementing the recipe—then we could easily imagine an attempt to implement a recipe with materials that were not suited to create a cake.

We know that if a formal program is implemented in the right material for having a mind, then it has intrinsic intentionality. But what is the scope of "right material" for having a mind? We do not even know what particular features of our brains make this material adequate for implementing programs that yield minds. We have, furthermore, no argument for the impossibility of other causal structures than those found in human brains producing minds. We know trivially that if some causal structure in some material is capable of producing a mind, then it has the same causal powers as our brains. Presently, however, the question of causal powers in regard to the question of having a mind only comes into focus by looking at the systems' performance. We judge that something most likely has the right causal powers because it performs as if it had such powers.

The emphasis on the implementation of a program takes the wind out of the Chinese Room Argument. Granted, the scenario described in the argument is indeed conceivable and thus logically possible, but that is not enough for Searle to make his case against strong AI. If strong AI is taken to be a metaphysical thesis claiming that in no possible world could the implementation of a certain formal program that yields a system behaviourally (and functionally) indistinguishable from creatures like us fail to have intrinsic intentionality, then the Chinese Room Argument retains its bite.

If, however, the thesis of strong AI is taken to be an empirical hypothesis about our world—as I think it should since, for example, Searle original critique was aimed at the research program of strong AI—then Searle's argument can no longer have the same force. Perhaps a Chinese Room of the Searlean kind is nomically impossible—e.g. it is just an empirical fact about our world and the natural causal laws pertaining to it that we could not build a mindless Chinese Room. Perhaps this is a world where anything—or a wide variety of different material constitutions including that of computers—that implements a formal program of a certain complexity and which passes an appropriately sophisticated Turing test will inevitably have intrinsic intentionality. The metaphysical question now turns out to be a question of what kind of world we are in—what our modal address is. It does not seem to me that any of the sides enjoys an upper hand. Consider the following argument.

Modal Address Argument, Part I:

> Premise 1. If I know that an actual implementation of the right kind of formal program in a computer—e.g. a program that enables the system to pass a sophisticated Turing test—also, inevitably, gives the system intrinsic intentionality, then I know that John Searle's Chinese Room Scenario is empirically (nomically) impossible.
>
> Premise 2. I do not know that John Searle's Chinese Room Scenario is empirically (nomically) impossible.
>
> Conclusion. I do not know that an actual implementation of the right kind of formal program in a computer—e.g. a program that enables the system to pass a sophisticated Turing test—also, inevitably, gives the system intrinsic intentionality.

Perhaps Searle is right when he claims that strong AI is misguided. The following argument, however, can be made in favour of the research program of strong AI.

Modal Address Argument, Part II:

> Premise 1. If I know that an actual implementation of the right kind of formal program in a computer – e.g. a program that enables the system to pass a sophisticated Turing test – could fail to give the system intrinsic intentionality, then I know that John Searle's Chinese Room Scenario is empirically (nomically) possible.
>
> Premise 2. I do not know that John Searle's Chinese Room Scenario is empirically (nomically) possible.
>
> Conclusion. I do not know that an actual implementation of the right kind of formal program in a computer – e.g. a program that enables the system to pass a sophisticated Turing test – could fail to give the system intrinsic intentionality.

The net effect of the Modal Address Argument is to show us that we do not know what our modal address (i.e. our world location) is in regard to the empirical possibility of a Chinese Room Scenario. Since we are unaware of our world-location, strong AI is a perfectly legitimate research program to pursue. The argument also shows us that we currently need to rely on some operational test as our guide for mind ascription and subsequently for giving an answer to whether computers in our world could think, and that is, I believe, a vindication of Turing, if not in letter, then in spirit.

Steffen Borge, Department of Philosophy, University of Tromsø, 9037 Tromsø, Norway. E email: steffenborge@yahoo.com

REFERENCES

Chalmers, David (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Searle, John (1980). "Mind, Brains and Programs". *Behavioral and Brain Sciences* 3.

Searle, John (1994). *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Searle, John (1995). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Turing, Alan (1950). "Computer Machinery and Intelligence". *Mind* 59.