Minds & Machines (2009) 19:229–235
DOI 10.1007/s11023-009-9151-9

A Logical Hole in the Chinese Room

Michael John Shaffer

Received: 27 August 2007 / Accepted: 18 May 2009 / Published online: 12 June 2009

Abstract Searle's Chinese Room Argument (CRA) has been the object of great interest in the philosophy of mind, artificial intelligence and cognitive science since its initial presentation in 'Minds, Brains and Programs' in 1980. It is by no means an overstatement to assert that it has been a main focus of attention for philosophers and computer scientists of many stripes. It is then especially interesting to note that relatively little has been said about the detailed logic of the argument, whatever significance Searle intended CRA to have. The problem with the CRA is that it involves a very strong modal claim, the truth of which is both unproved and highly questionable. So it will be argued here that the CRA does not prove what it was intended to prove.

Keywords Chinese room · Computation · Mind · Artificial intelligence

Searle's Chinese Room Argument (CRA) has been the object of great interest in the philosophy of mind, artificial intelligence and cognitive science since its initial presentation in 'Minds, Brains and Programs' in 1980. It is by no means an overstatement to assert that it has been a main focus of attention for philosophers and computer scientists of many stripes. In fact, one recent book (Preston and Bishop 2002) is exclusively dedicated to the ongoing debate about that argument 20 some years since its introduction. In any case, the significance of the CRA is supposed to be clear. The CRA is supposed to scuttle the specific project known as Strong Artificial Intelligence (SAI) and "good old fashioned artificial intelligence" (GOFAI) in general, and so it has been thought to have important implications for how we ought to reorient the artificial intelligence community's attempts to create intelligent systems.

M. J. Shaffer (✉)
St. CloudState University, St. Cloud, MN, USA
e-mail: shaffermphil@hotmail.com

Recall that SAI is, more or less, just the view that mental content just is, or at least is determined by, the manipulation of purely formal symbols in accordance with syntactic rules. So SAI is often thought to be an extreme deflationary view of content. Given this view Searle himself is explicit about the significance of CRA in this respect:

> Because programs are defined purely formally or syntactically, and because minds have an intrinsic mental content, it follows immediately that the program itself cannot constitute the mind. The formal syntax of the program does not by itself guarantee the presence of mental contents. I showed this a decade ago in the Chinese room argument (Searle 1992, p. 200).

Of course, the CRA and what exactly it really implies about SAI and GOFAI has always been a matter of great controversy, and so it is then especially interesting to note that relatively little has been said about the detailed logic of the argument, whatever significance Searle intended CRA to have.[1]

Typically, the "argument" is presented in the form of a story, a kind of thought experiment. This is unfortunate for at least two reasons: (1) it makes it difficult to assess the significance of the CRA and (2) it obscures the fact that the CRA has not been shown to be sound and appears as if it might be straightforwardly unsound. Here we will be concerned with issue (1), but only in so far as it is necessary for the consideration of issue (2). The CRA arises out of the following, now familiar, story:

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch a "script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions," and the set of rules in English that they gave me, they call the "program." Now just to

[1] To my knowledge Copeland (1993) is the only specific and detailed treatment of the logic of the CRA, although Cole (2004) briefly addresses the issue. Copeland (2002) also takes issue with the CRA but in a different manner than I do.

> complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from tile point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view—from the point of view of someone reading my "answers"—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program (Searle 1980).

The conclusion that Searle believes we should draw based on this possibility so described is that SAI and GOFAI cannot ever succeed. Why can they never succeed? SAI in particular cannot ever succeed because the manipulation of formal, symbolic elements via the implementation of syntactical rules is supposed to be insufficient to generate mental content (Searle 2002) and this is supposed to be the case because, "[t]he argument rests on the simple logical truth that syntax is not the same as, nor is it sufficient for, semantics (Searle 1992, p. 200)." As such, the CRA is supposed to challenge a variety of related views about the relationship between computation and mental content, importantly including behavioral analyses of mental content based on the Turing test and the whole project of computer functionalism. But what exactly is the *argument* in the infamous passage cited above?

The CRA story appears to contain the following argument:

1. The room occupant knows no Chinese.
2. The room occupant knows English.
3. The room occupant is given sets of written strings of Chinese, $\{C_i, C_j, \ldots, C_n\}$
4. The room occupant is given formal instructions in English that correlate pairs of sets of Chinese strings, $\langle C_i, C_j \rangle$.
5. The room occupant is given formal instructions in English to output some particular C_i given a particular C_j.
6. The room occupant's skill at syntactically manipulating the strings of Chinese is behaviorally indistinguishable from that of a fully competent speaker of Chinese.
7. If 1–6 are jointly possible, then syntax is not sufficient for mental content.
8. 1–6 are jointly possible.
9. Therefore, syntax is not sufficient for mental content.

So, the basic notion behind Searle's CRA is that we can have a system that is input-, output- and even transition state-equivalent to a being with mental states but which does not have those very same correspondent mental states. The room occupant perfectly *mimics* a native Chinese speaker both internally and externally *qua* formal syntactic features while the room occupant, at least according to Searle, does not have the mental states present in the native Chinese speaker.

But, Searle cannot simply assert the logical truth of the claim that "syntax is not the same as, nor is it sufficient for, semantics." This would be to beg the very question at issue. It would simply be *to assert the conclusion of the CRA rather than to provide an argument for it and against SAI*. Moreover, Searle cannot legitimately replace premise 7 with the weakened premise:

7′. If 1–6 are jointly possible, then the room occupant does not have the mental states that a native Chinese speaker has

This would only allow Searle to conclude that the room occupant does not have the mental states that *a native Chinese speaker* has and this is a far cry from the level of generality necessary to threaten SAI and, more generally, GOFAI. Drawing the general conclusion against SAI requires greater generality and so the following additional premise would need to be added in addition to 7′:

(SM) If the room occupant does not have the mental states that a native Chinese speaker has, then syntax is not sufficient for mental content.

7′ and SM are jointly, however, equivalent to 7. Let us call the problem that arises here *Searle's mistake* for ease of reference. The problem involved in 7 is not a problem with 7′ but with rather with SM and the problem is that it is not all clear that it is true. We shall soon see why. In any case, the rendition of the argument given above then helps to reveal precisely what Searle would have to establish in order to yield the controversial conclusion that he endorses without begging the question of the truth of SAI and GOFAI.

So given this rendering of the CRA, we can then proceed to ask whether it is in fact sound. Searle himself remains explicitly convinced that it is, in fact, sound (Searle 2002, p. 51). The CRA appears to be obviously valid and so if we are to find a hole in the argument it must be a matter of challenging the truth of one or more premises of the CRA, *pace* Searle. However, premises 1–6 appear to be individually immune to challenge. They are rather like paradox-constituting propositions that are "part of the story," as in the case of, for example, the Barber Paradox (See Olin 2003, pp. 9–12). One cannot simply challenge individual components of the story itself in order to reject the conclusion as they are simply stipulated *ex hypothesi*. As a result, this leaves 7 and 8 as the only real potential targets if we are to take issue with the conclusion of the CRA.

There are in fact two fairly obvious ways in which one can challenge the CRA and both can be understood to be versions of what is known as the systems reply to the CRA. First, one might challenge premise 8 by asserting that the sub-set of premises of the CRA {1, 2, 3, 4, 5, 6} is inconsistent and so 8 cannot be true. Again, this is one interpretation of a fairly standard kind of computer functionalist inspired response. One simply bites the bullet and restores consistency to the story by rejecting premise 1 and concluding that, in some sense, the room occupant actually knows Chinese in some important sense when we consider the details of the story more carefully.[2] The second way one might interpret the systems reply would be to accept that the sub-set of premises of CRA {1, 2, 3, 4, 5, 6} is actually consistent, but that 7 is false because *the room* understands Chinese even if the occupant does not.[3] The problem with both of these versions of the systems response, however, is that they simply beg the question of the truth of SAI. In answering the CRA in *either manner* the functionalist, for example, simply *asserts* that the occupant knows Chinese or that the room knows Chinese because, respectively, the room occupant's formal manipulations or the operations of the occupant and the room as a whole are—from both the internal and external perspectives—structurally identical to a native Chinese speaker's formal manipulations. Of course, that kind of response simply won't do. We need at very least to have a better understanding of *why* it is reasonable to hold that syntax might be sufficient for semantics and this cannot be achieved by simply stating that the situation described by {1, 2, 3, 4, 5, 6} is such a case. What is then more intriguing is that in challenging CRA in either of these manners a more robust way of defusing the CRA has been overlooked that does not depend on simply asserting either that the room understands Chinese even if the occupant does not or that the room occupant understands Chinese in some important sense. Both of the versions of the systems response presented above are then apparently question-begging and so do not constitute an adequate response to the CRA, but examining them in light of having an explicit understanding of the structure of the CRA does point us in the right direction.

So, if we are to provide an adequate (i.e., non question-begging) response to the CRA we must pay more careful attention to the key premises of the CRA, especially with respect to the modal strength of those premises. The contention that will be made here then is that premise 7 seems to be false because SM seems to be false and so at very least the CRA has not been established as sound once properly rendered as an explicit argument in the manner we have done here and once we recognize the modal strength of these claims. This allows one to reject the CRA *without* begging the question of the truth of SAI against Searle. Thus, SAI is not necessarily imperiled by the CRA, but not for the precise reasons that most respondents have typically claimed. In effect, we can regard the point made here as a new and much more powerful modalized version of the systems reply.

Consider premise 7: If 1–6 are jointly possible, then syntax is not sufficient for mental content. What this premise asserts is essentially that the compossibility of facts about the room occupant imply that semantics, or intentionality, cannot arise

[2] This interpretation of the systems response assumes that 1 and 2–6 are inconsistent in some sense.

[3] I thank an anonymous referee for pointing out the second way of interpreting the systems response, and I suspect that the referee is correct in asserting that it is the more typical interpretation of the systems response. Nevertheless, the first approach is interesting in and of itself as a response to the CRA and so it is worthy of attention here. The real point is then that the CRA can be rebutted without either having to assert that *the room* understands Chinese even if the occupant does not or that the occupant understand Chinese.

out of *syntax alone*. But why should we accept premise 7? More specifically, why should we accept SM? Searle himself offers no substantive argument that either claim is true, let alone that they are logically true, and to be sure 7 is *the* crux of the CRA.[4] Moreover, we should be careful to note the modal character of premise 7 and how logically strong it actually makes premise 7 and thereby how strong it makes SM. What Searle asserts in endorsing premise 7 is nothing less than the claim that there is *no* possible world in which 1–6 are true and where the room occupant knows Chinese in the sense of having mental states corresponding to those possessed by a native speaker of Chinese.[5] SM is the claim that if the room occupant does not have the mental states that a native Chinese speaker has, then syntax is not sufficient for mental content. So, it is just the claim that it is not possible that the room occupant in the world described in the CRA fails to understand Chinese and that syntax is sufficient for semantics. This is the key to the CRA.

But this modal claim is surely false in a very straightforward way. Consider a possible world, w_1, described as follows. World w_1 is much like our own and so let us assume that is a close possible world in the sense that it differs in no other way from the actual world than in the following single respect.[6] In world w_1 let us assume that there is an additional *emergent property* that is a member of the set of properties permissible by the laws of nature of w_1, L. Let us then define this emergent property permitted by L simply as the property that from *sufficiently* rich systems of syntax, semantics properties (or intentional properties, or meanings, or mental states—pick your favorite) emerge. Nothing that Searle says in the CRA story precludes the existence of the causal emergence of semantic properties from syntactic systems, at least not without begging the very question at issue, and so we then have a clear counter-example to premise 7 by having a clear counter-example to SM. In w_1 all of 1–6 can be jointly true and premise 8 can be true, but premise 7 can false because while the room occupant knows no Chinese in the sense of having mental states corresponding to those possessed by a native speaker of Chinese, *if* the room occupant's system of syntax were sufficiently richer (and just rich enough for emergent semantic properties to arise as they do in w_1), then she *would* understand Chinese in the sense of having mental states corresponding to those possessed by a native speaker of Chinese.[7] In other words, the simple failure of *the room occupant* to understand Chinese in the CRA story is insufficient to validate the claim that syntax is insufficient for semantics, as SM would require.[8] In a sense then the systems reply

[4] Searle's only real reason for accepting this contention appears to be that there really is no syntax at all. Syntax is rather something that we impose on systems when we interpret their behaviors. This is however not an adequate response. What is important about syntax is just structure and structures—or structural properties—are as real as anything else. So the discussion could be formulated in terms of the causal power of structural properties to produce semantic properties and Searle offers no arguments against this possibility.

[5] The same point holds for the matter of whether the room understands Chinese.

[6] To be sure, this world may be the actual world.

[7] Again, the same point can be made with respect to the matter of the room's understanding Chinese.

[8] Again, I wish to thank an anonymous referee for pointing out that one might regard this as the proper way to understand the systems reply. I am somewhat unsure about this matter, as the exact nature of the systems reply is not entirely clear for reasons noted earlier. If my solution in fact agrees with the second

can be strengthened—whichever interpretation one favors—and functionalists can simply and completely defuse the CRA by asserting neither that the room nor that the room occupant understands Chinese, but by merely asserting the weaker claim that it is possible that syntax is sufficient for semantics because there possible worlds where syntax is sufficient for semantics. So the debate about functionalism and the CRA has apparently been predicated on the basis of a failure to take account of the modal aspects of Searle's claims which could perhaps have been avoided were the parties to the debate more explicit about the argument involved.

So, in any case, Searle has not shown that the CRA is sound and there appears to be some good reasons to believe that it is in fact unsound. Moreover, this result is resilient because CRA could only be repaired by showing that such emergent semantic properties are outright *impossibilities*. This would require showing that that *SM* is a logical truth. But this seems highly unlikely, as there is nothing in the least contradictory about the existence of such properties and no reason has been offered by Searle that would underwrite treating SM as a necessary truth. Moreover, it has become increasingly clear that there are many *actual* emergent properties (e.g., liquidity, chaotic phenomena, etc.) permitted by the laws of the actual world and so there is no special reason to suppose that the CRA threatens SAI in the way that Searle believes it does because there is no special reason to suppose that emergent semantic properties are especially strange or impossible given our knowledge of other well-known kinds of emergent properties. In short, the CRA depends on 7 and in so doing depends on SM. The problem is then that SM is a very strong modal claim, the truth of which is both unproved and highly questionable. At best what CRA then does is to reveal that the success of SAI and GOFAI may well depend on the details of the concept of emergence and its application to semantics, but this, I take it, is not necessarily entirely new news. This of course means that SAI and GOFAI are not as deflationary as the may initially appear to be. What is, however, abundantly clear is that the CRA does not seem to prove what it was intended to prove.

References

Cole, D. (2004). The Chinese room argument. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/fall2004/entries/chinese-room/.

Copeland, J. (1993). *Artificial intelligence*. Cambridge, MA: Blackwell.

Copeland, J. (2002). The Chinese room from a logical point of view. In: J. Preston & M. Bishop (Eds.). *Views into the Chinese room* (pp. 109–122). Oxford: Oxford University Press.

Olin, D. (2003). *Paradoxes*. Montreal: McGill-Queens University Press.

Preston, J., & Bishop, M. (2002). *Views into the Chinese room*. Oxford: Oxford University Press.

Searle, J. (1980). Minds, brains and programs. *The Behavioral and Brain Sciences, 3*, 417–424.

Searle, J. (1992). *The rediscovery of mind*. Cambridge, MA: MIT Press.

Searle, J. (2002). Twenty-one years in the Chinese room. In: J. Preston & M. Bishop (Eds.). *Views into the Chinese room* (pp. 51–69). Oxford: Oxford University Press.

Footnote 8 continued
interpretation of the systems reply, then the solution offered here can simply be regarded as a more well-defined way to see the modal error involved in the CRA. Again, I can remain neutral on this matter here.