Weak Strong AI:
An elaboration of the English Reply to the Chinese Room

Ronald L. Chrisley
School of Cognitive & Computing Sciences
University of Sussex
ronc@cogs.susx.ac.uk

June 21, 1995

Abstract

Searle (1980) constructed the Chinese Room (CR) to argue against what he called "Strong AI": the claim that a computer can understand by virtue of running a program of the right sort. Margaret Boden (1990), in giving the English Reply to the Chinese Room argument, has pointed out that there *is* understanding in the Chinese Room: the understanding required to recognize the symbols, the understanding of English required to read the rulebook, etc. I elaborate on and defend this response to Searle. In particular, I use the insight of the English Reply to contend that Searle's Chinese Room cannot argue against what I call the claim of "Weak Strong AI": there are some cases of understanding that a computer can achieve solely by virtue of that computer running a program. I refute several objections to my defense of the Weak Strong AI thesis.

Searle constructed the Chinese Room (CR) to argue against what he called "Strong AI" (SAI): the claim that a computer can understand by virtue of running a program of the right sort. The argument was: Searle can execute any program you like and not thereby understand Chinese, so therefore there is no program the execution of which is sufficient for understanding Chinese. Since understanding Chinese is an arbitrary case of understanding, the argument generalizes to any kind of understanding: a computer cannot understand anything solely by virtue of running a program.

Margaret Boden (1988), in giving the English Reply to the Chinese Room argument, has pointed out that there *is* understanding in the Chinese Room: the understanding required to recognize the symbols, the understanding of English required to read the rulebook, etc. I would like to elaborate on and defend this response to Searle. In particular, I will use the insight of the English Reply to contend that Searle's Chinese Room cannot argue against the following claim:

> There are some cases of understanding that a computer can achieve solely by virtue of that computer running a program.

Although as worded, Searle's characterization of Strong AI could be construed to be something this weak, it is clear that he actually takes Strong AI to be stronger than the above. For example, it seems to be a consequence of Strong AI that one could understand Chinese solely by virtue of running a program, yet this is not an implication of the above claim.Thus I call the above the claim of "Weak Strong AI", or WSAI.

Here's why the CR does not argue against WSAI. Suppose Searle implements any program you like, in a manner similar to the way he implemented the program in the Chinese room: he is passed symbols, he consults a rulebook, he outputs symbols according to his understanding of the rulebook. (We shouldn't call this room the Chinese Room, since the symbols need not have anything to do with Chinese, syntactically or semantically. So let "CR" now stand for the Computational Room.) Now in order to reject WSAI, Searle has to show that for all programs, there will not be any understanding *at all* in the CR. For if there was ever a case of understanding, in the running of even one of the programs, then WSAI would be true. But the problem comes from Boden's observation: Searle *does* understand while implementing a program in the CR. Although Searle might deny, as before, that there is not this kind or that kind of understanding going on in the CR, he cannot deny that there is *some* understanding going on. So WSAI stands.

Objections:

> **Irrelevance objection:** "Searle's target was SAI, not WSAI, so the failure, if any, of the CR argument to refute WSAI is irrelevant."

Since I am granting that WSAI is not the same as SAI, it is open to Searle to agree with me, and concede that the Chinese Room does not argue against WSAI; he could claim that his target was only SAI, and against that position, his argument still holds. I don't think that Searle would want to take this route, though. I think that at heart he is opposed to the possibility that a computer could *ever* understand merely by virtue running a program. So even if denying SAI is not the same as denying WSAI, I do think that Searle is in fact committed to the rejection of WSAI. So he cannot dismiss the failure of the CR to reject WSAI.

> **Direct objection:** "We can just apply the original CR argument to the case of whatever understanding Searle has (understanding English, the rulebook, whatever), and show that even if such understanding were possessed by the room, it is not possessed solely by virtue of implementing a program."

Searle argued that the CR doesn't understand Chinese, on the grounds that if the CR understood Chinese, then Searle would understand Chinese, but he doesn't. But note that Searle cannot argue the same way concerning whatever understanding he has (understanding the rulebook, for instance), for example "if the CR understood the rulebook, then I, Searle, would understand the rulebook". For Searle does, ex hypothesi, understand the rulebook. No contradiction is generated, so no argument against WSAI is made.

> **Behavioural objection:** "A necessary condition for having understanding is having behaviour appropriate for understanding. In the case of the Chinese room this condition was met, but it was shown that this was not sufficient for understanding. The computational room does not even meet the necessary behavioural condition for understanding the rulebook. Thus we can rule out understanding of the rulebook and defeat WSAI."

Of course, this is based on a simple error of quantification. To defeat WSAI, the CR has to show that *no* program is sufficient for understanding. For example, Searle would have to show that a program that passes the Turing Test with respect to understanding the rulebook does not actually understand the rulebook. But he has not done so, and I see no way his Chinese Room argument can be modified to do so.

> **Module objection:** "Searle is not the CR; he is only a module of the CR. It is the not the CR that is doing the understanding, but only the Searle module. Thus, it has not been shown that the computer (the CR as a whole) understands."

This reply makes two errors. First, it misunderstands the dialectical role of this paper. I am not trying to argue directly for the truth of WSAI; I am only trying to show that Searle's CR does not refute it. I have not shown that the CR understands; but equally Searle has not shown that it doesn't understand. The second mistake: in order to make the original CR argument, Searle must *not* allow a split between what the CR understands and what Searle understands. If he does, his argument can be refuted by the Systems Reply. Making the Module reply to my argument would allow one to make the Systems Reply to Searle's.

> **Overkill objection:** "It is Searle's understanding that is getting in the way here. Searle and his understanding are not required in order to run a program in the CR. He could be replaced by something that doesn't understand, a simple mechanism to which no one would ascribe understanding. In that case there would be no reason to believe that the CR had understanding."

Again, this misunderstands the argument. It is Searle who has to show that WSAI is false. The entire trick of the CR argument is that Searle is the one executing the program. In order to argue against SAI or WSAI, the CR needs to have a homunculus that is sophisticated enough to a) be the obvious locus of the computational action in the CR, and b) report back that there is no understanding there. If Searle is replaced with a mere simple mechanism, then the first constraint is not met. The homunculus will just be an insignificant, even trivial part of the entire CR. Then we will not find compelling a key implicit premise in Searle's argument: that the homunculus would know if there were any understanding going on. The Systems Reply becomes irresistible. If the homunculus is so simple, why should it have any understanding, or be aware of any understanding any more than our hypothalamus is? Furthermore, a simple mechanism would violate b): it would not be able to tell us that there is no understanding. Not that communication is essential: in the original CR argument, each of us can imagine being Searle in the room, incorrigibly knowing directly that no understanding of Chinese is going on. But we cannot imagine ourselves being a non-understanding mechanism. Searle must be the homunculus for the CR argument to even get going.

> **Unconscious objection:** "OK then, how about the same objection as above, but with an unconscious Searle as the homunculus? In such a case, Searle could report back to us later, informing us that there was no understanding at all going on. Surely in other contexts we acknowledge cases in which a subject performs a task, but later reports that there was no awareness at all, even less an understanding of what was going on at the time. We take the subject to be authoritative concerning the absence of understanding in such cases, so we must take Searle to be authoritative in this case: the program is implemented, but there is no understanding."

Although there is no awareness in the case of the unconscious Searle, there is still understanding. The CR argument is about understanding, not awareness. One might ask: why think that non-aware subjects are really understanding anything? Because there are cases where the notion of understanding is crucial to explaining their unconscious behaviour. For example, if an unconscious subject responds to someone saying "go back to bed now", then one can only explain their subsequent bed-directed motion if one postulates that they, while unconscious, understood English in general and the utterance in particular.

> **Conscious objection:** "Alright, how about insisting only that running a program does not ensure *conscious* understanding? Then an unconscious Searle could report back that there was no unconscious understanding of any kind going on, for any program you like."

Even if we grant that Searle can authoritatively rule that there was no conscious understanding (in him or the CR as a whole) at an earlier time, there still remains a problem concerning the other requirement for a CR homunculus, as mentioned above. That is (and this applies to the previous objection as well), we can note again that as we simplify the cognitive capacity of the homunculus (in this case, by making Searle unconscious), we eliminate support for the crucial premise that if there is any understanding going on, the homunculus (unconscious Searle) would be aware of it.

> **Modal objection:** "The argument misunderstands what Searle must do in order to show WSAI false. He does not have to show that there is no understanding; he only has to show that the understanding, if any, is not had solely by virtue of running a program. And that is easy to do. That there is any understanding at all in the CR is the result of Searle's intrinsic properties, not by virtue of any particular program being run. The understanding is present no matter what the program. Vary the program, and you still have understanding; vary Searle's presence and you sometimes do, sometimes don't, have understanding. So it is not by virtue of running the program that there is understanding; it is an implementation detail (that Searle is involved) that is responsible for any understanding."

To defeat this objection I have to show that the following two conditions have not been shown to be false: that understanding in the CR could depend on the program being run; and that understanding in the CR need not depend on anything else (i.e., it need not depend on Searle being part of the implementation). For the first condition: it is true that no matter what program is being run, Searle will have some understanding. But it is not a consequence of WSAI that no matter what program is run, the CR will also thereby understand solely by virtue of that program running. Rather, there only has to be *one* case of a program whose running guarantees understanding. It could be that for every program but one, only Searle is doing any understanding at all. So it has not been shown that the presence of understanding does not depend on the program being run. For the second condition: the understanding of the CR would be shown to depend on Searle's presence only if it were shown that even though there is understanding when Searle is present, when Searle is not present, there is no understanding. But the CR is powerless to show this: it can only tell us about cases in which Searle is present. To assume that he is responsible for any understanding would be tantamount to assuming that there can be no understanding by virtue of running a program alone – which is to beg the question.

> **Panpsychism objection:** "Searle could run any program you like, and there would still be understanding of the rulebook on his part. Since WSAI claims that this understanding is by virtue of the program alone, and not Searle's particular implementation, it is committed to all sorts of absurdities concerning those very programs running without Searle as part of the implementation: that every program understands; that every program understands just what Searle understands, that every program understands the same things Searle understands, etc."

No. As said above, WSAI is not committed to claiming that whenever Searle understands something, there must be some understanding in the CR as a whole. (It is the converse of that claim that is being accepted here, as a concession to Searle: if there is understanding in the CR, then Searle would have it too. If Searle drops that claim, he is defenseless against the Systems Reply.) So it may only be that there is one program for which there is understanding on the part of the CR solely by virtue of the program being run. And in that case, Searle will not be able to say that there is no understanding going on.

> **Regress objection:** "If the CR with Searle does understand something, it must be something Searle understands – the rulebook, say. But then imagine that same program implemented in a different way: The bit that was done by Searle with the English rulebook is now computed by an *English* Room with a, say, Finnish person, Satu, inside it, who knows no English and is using a Finnish rulebook. Since it is the same program, according to WSAI it must have the same understanding (of the English rulebook) as the CR with Searle in it. Yet Satu would know if there was any understanding of English going on, and since she doesn't, there isn't." (Thanks to Mike Wheeler.)

This assumes that CR with Searle and the CR with the English Room (ER) and Satu are both running the same program. But they are formally and syntactically distinct. So the WSAI theorist is permitted to say that they have differences in understanding.

Another reply to this is to deny the implicit inductive step:

> If the homunculus in the CR understands everything the CR understands, then it must be that the homunculus in the ER (which is simulating the homunculus in the CR) understands everything the CR understands as well.

This has not been motivated. If we had a clear reason why Searle should be expected to understand everything the CR understands in the first place, then perhaps we would have some intuitions that could help us decide one way or another on the inductive step. But as things are, the inductive step is unjustified.

> **Consistency objection:** "Surely, Ron, you don't actually believe the Weak Strong AI thesis. So why are you supporting it?"

I suspect that Searle is right: a true understanding of understanding cannot be given in terms of formal, purely syntactic programs. But I don't think the Chinese Room shows this. If we understand why it does not, perhaps we will be in a better position to see the real reasons why formal cognitivism is doomed.

Bibliography:

Boden, M. (1990) Escaping from the Chinese room. In Boden, M. (ed.) The Philosophy of Artificial Intelligence. Oxford: OUP; pp 89-104.

Searle, J. Minds, Brain & Programs. (1980) Behavioral & Brain Sciences 3: 417-24.