Searle's Sophistry: Examining the Foundation of the Chinese Room

Christopher Watkins

John Searle's infamous CRA has aroused much controversy in the research fields of contemporary philosophy, cognitive science and artificial intelligence. While one must recognize the import of the cross-disciplinary conversation initiated in response to Searle's proposition, his actual argument and thought experiment cannot be similarly praised. There is ample literature picking out and igniting the myriad argumentative matchsticks that lie latent in Searle's reasoning, however, much of the opaque, ad hoc counter argumentation can be traced to a fundamental issue pertaining to the assumption made underneath Searle's argument and thought experiment; namely, that the question of whether Searle "understands" Chinese while in the room is a coherent and answerable yes-no question. This essay aims to bring clarity to the argument's scope, intended target, and reason for baselessness. I will assume the reader is familiar with Searle's original paper. The thesis, put simply, is that Searle's hypothetical question of "do I understand Chinese" is meaningless when stripped of appropriate context, and as such fails to justify the truth of his needed premise in his overall argument. As we shall see, Searle's use of loaded language exploits the linguistic and conceptual ambiguity present in his ill-formed thought experiment. This essay begins by clarifying the intended target and scope of the argument, followed by evidence and discussion of where and how Searle's argument goes wrong, and concludes with a discussion of the meaning and implications of the thesis for the larger research community.

Searle has, on repeated occasion, identified his target clearly: Strong AI. The trouble has been in figuring out exactly what is meant by this term. Searle explains 'My discussion here will be directed at the claims I have made defined as those of strong AI, specifically the claim that the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition' (Searle 1980: 417). Larry Hauser has correctly denoted this as the conjunction of the metaphysical claim 'the appropriately programmed computer literally has cognitive states' and the methodological one 'that the programs thereby explain human cognition' (Preston & Bishop 2002: 123). Part of the reason Searle's argument has aroused so much controversy is that the metaphysical claim seems to be the affirmative answer to Turing's original (1950) question: 'Can a machine think?' Yet, Searle admonishes, 'There are several misunderstandings of the Chinese room argument and of its significance. Many people suppose that it proves that 'computers cannot think'. But that is a misstatement' (Searle 1994: 546)[1]. The misunderstandings are not without justification, as we shall see. What then does the argument target? Searle concedes: 'this is the point of the parable—if I don't understand Chinese on the basis of implementing the program for understanding Chinese, then neither does any digital computer solely on that basis because no digital computer has anything that I do not have [in the example]' (Preston & Bishop 2002: 124) Hence, the focus is not whether computers *can* think, instead the argument is centered on *programs*—specifically the inability of programs, in and of themselves, to be responsible for (artificial) intelligence. This position is none other than *computationalism*, which holds that computation is essentially what thought is, that is, computation is both necessary and sufficient for thought (mind). The computationalist viewpoint, unlike 'strong AI', is one that is indeed held by researchers and motivates many research programs in cognitive science and

[1] Reprinted in (Preston & Bishop 2002: 124)

therefore positions the argument to undermine these activities, if sound. With these details clarified, it would seem natural to now assess the argument's soundness, as formally laid out, and examine whether the reasoning truly refutes the computationalist position. This is exactly what has been attempted and the result remains controversial. Instead, we will examine the assumptions underlying the Chinese room and see how they contradict the possibility of the thought experiment doing the philosophical work Searle claims it does.

The Chinese room thought experiment is proposed as justification of premise 2 of the original presentation of his argument.[2] The force of his thought experiment is, of course, that the reader's intuitions verify that Searle indeed would not understand Chinese while following the program for symbol manipulation of the Chinese inputs, while successfully producing the correct Chinese output character strings. Searle contrasts this situation with that of speaking his native English: 'what is it that I have in the case of the English sentences that I do not have in the case of the Chinese sentences? The *obvious* answer is that I know what the former mean, while I haven't the faintest idea what the latter mean' (Preston & Bishop 2002: 81: emphasis added). Searle's use of loaded language (in this case 'obvious') throughout the thought experiment begs the reader to appeal to common-sense intuition, and covers up the real philosophical issue present: Searle assumes "understanding" is an observer-independent predicate. This is decidedly not the case. Terry Winograd illustrates the notion of observer-dependence in our language with the following example– Consider a victim of the rare condition of testicular feminization, whose external characteristics are those of a woman, but whose chromosomal pattern is male and whose reproductive organs are not developed for either gender (Preston & Bishop 2002: 81). Is this individual a man or a woman? This example shows the necessity of the predicate (man or woman)

[2] The argument is printed at the end of this essay.

to be grounded in a certain contextual framework to be appropriate. The social system employing the linguistic terms ‘man’ or ‘woman’ may decide, based on their appraisal of which concerns matter most, whether the label ‘man’ is correct or incorrect. Important here is the recognition that there is not an objectively correct answer to the question of the individual’s gender, the answer is relative. Hence, predicates like this, despite seeming intuitively decidable for things we commonly come into contact with in daily life, become ambiguous when viewed fundamentally. This is exactly the ambiguity that Searle exploits in the Chinese room argument.

Let us consider more specifically the case with ‘understanding’. In language one most often uses this term to describe a psychological state of a human and can be judged both internally (I understood the lecture today) and externally (he did not understand my point). This demonstrates the continuum on which we use the term when describing humans, but it is important to notice that the measure of whether someone understands or not is something that cannot be objectively verified in *any* case; the extent to which we say someone understands something is due to how well they stack up against the context specific measure of understanding which is seen as most coherent for that instance. For example, I may understand a written text without comprehending the literary significance of it, or may algorithmically complete a mathematical problem without proper conception of why I was doing the computations. In each case it is not coherent to say that I do not *objectively* understand- for this description is relative to the standard of the socially agreed upon measure for each context. However, I will, in everyday language be fine with labeling these cases as not understanding, and there is no problem with this. Winograd summarizes: ‘everyday common language urges us to believe that there must be a ‘right’ answer, not dependent on purposes, context and interpretation’ (Preston & Bishop 2002: 83). Searle begs that we recognize the obviousness of the everyday common-sense proposition that he does not understand Chinese.

The issue is not that this is a unique case, wherein the system with Searle and the program running in sync with each other is odd and doesn't neatly fit our language, the issue is that *every* case in which we employ the predicate 'understand' is observer-dependent and context relative. Hence, it is not that Searle does not have grounds to use one interpretation (the common-sense one) over another (possible, theoretical one?), but rather that the question of whether or not he understands Chinese is a relative one, not an objective one, and therefore will vary depending on the context and assumptions of the observers. No wonder that there is such disagreement as to whether the Chinese room understands – there cannot be an objective agreement! Even if all parties (Searle and interlocuters) were convinced one way or the other, it would be an illusionary objectivity – only induced by the seeming agreement of observers on the contextual appropriateness of the applied predicate.

In this way, Searle claims to have justified his premise (2) that the execution of a computer program is insufficient for understanding. As we have seen, what he really has done is posed an ill formed question that cannot be answered objectively, and answers it by appealing to common sense and the shared contextual judgement associated with understanding. The counter arguments to Searle assume a different linguistic orientation to what counts as understanding and what types of things it makes sense to ascribe that predicate to. For example, the systems reply, in which much derivative argumentation is grounded in, simply assumes understanding to be a more functionally justified, 3rd person relative predicate. This is not, again, correct or incorrect, the point is that it is relative and hence neither position can adequately prove the other wrong. A similar logic can be applied to each reply to Searle that has been given, and will result in the comprehension that they too brought different assumptions about what "understanding" is to the conversation.

It has been shown that recognizing the context dependency and relative nature of predicates such as ‘understanding’ as portrayed in the Chinese room experiment renders Searle’s question (and negative answer) insufficient to justify his premise used in his Chinese room argument. One might be satisfied with the clarity brought to Searle’s sophistry by this discussion but also feel unsettled about the larger philosophical debate surrounding computers and thinking. This allows us to see Searle’s question of whether the Chinese room understands as incoherent, yet productive. In a different context, one with the aim of providing more clarity to the already misconstrued (popular culture) portrayal of AI , the question is indeed useful, for it demonstrates where aspects of our language prove inadequate in examining and describing new types of systems that exhibit properties we seem to understand from a functional level (deciding, seeking, processing) but not from an intuitive, causal or experiential level. It is common in the current technologically advanced social climate to describe computers as thinking or deciding things, and understanding natural language. It is up to the linguistic community to decide the context with which we view these descriptions and that should be informed by the scientific evidence and other empirical and theoretical indications present, with the aid from philosophers who can fundamentally examine the nature of these systems and guide the larger culture in a way that most accurately reflects what the systems are doing, to the best of our knowledge. Research continues to push the frontier of what dynamical and computational systems can do and it is an important, even urgent, job of philosophy to clarify the best way to view such activities, and point out where language or other constraints on actual knowledge can impede on our reasoning. What things should we consider autonomous? How should we talk about computational systems preforming tasks that seem to require intelligence? These questions should be examined so as to reduce the uncertainty and confusion that rests around dynamical systems, the nature of mind, and the prospect of artificial intelligence,

providing a way to accurately understand new technologies and precisely communicate their relation to long standing philosophical ideas.

Searle's argument: (as presented in MBP)

Premises

1) Intentionality in human beings (and animals) is a product of causal features of the brain.
2) Instantiating a computer program is never by itself a sufficient condition of intentionality.

Conclusion

3) The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program.
4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain.
5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain.
    a. claim: this follows from 2 and 4

References:

SEARLE, J. R. (1980) ‘Minds, Brains, and Programs’, *Behavioral and Brain Sciences*, 3: 417-24.

PRESTON, J. M. & BISHOP, J. M. (eds.) (2002). *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence.* Oxford University Press.