compos mentis

Semantics, 'Strong' AI, and the Chinese Room Argument

Ameer Sarwar
University of Toronto, St. George

ABSTRACT
The main purpose of this paper is to defend Searle's (1980) classic Chinese room argument against a number of objections. Searle takes his argument to show that semantics do not inhere in formal symbols. Consequently, since 'strong' AI concerns itself solely with the implementation of formal symbols over recursive syntactical rules, its inability to account for inherent meaning precludes it from being established as a viable research program in cognitive sciences. Two major strands of objections and sub-objections are reviewed against Searle's argument, but it is ultimately concluded that they both fail. The 'disjoint personalities objection' fails primarily because there can be no change in the personality of the room without a change in the personality of the inhabiting symbol manipulator. The 'other minds objection' fails because it engages in reverse causality: it concludes from manifestations of intelligent behaviour that the thing behaving intelligently is thereby intelligent. My attempts at demonstrating the failure of the two objections rescue Searle's argument, and therefore, the problem of original meaning remains a thorn in the philosophical foundations of 'strong' AI.

KEYWORDS
Chinese Room Argument, 'Strong' AI, Turing Test, Semantics, Formal Systems, Joint Personalities, Other Minds, Original Meaning

INTRODUCTION

This essay attempts to establish that Searle is correct in arguing that semantics do not inhere in formal symbols, and so, there is no inherent understanding in formal computational systems, thereby brining into serious doubts the prospects of 'strong' AI as a research paradigm. I begin the paper by explaining what I mean by 'strong' AI, what the Chinese room argument is, and how the latter causes problems for the former. I then consider a series of objections—which are divided into the 'Disjoint Personalities Objection' and 'Other Minds Objection'—and try to diffuse them all in order to ultimately conclude that Searle's argument survives. The implication is that the prospects of 'strong' AI become doubtful; specifically, it cannot explain the problem of original meaning in terms of implementing formal programs.

"STRONG AI" AND THE CHINESE ROOM ARGUMENT

'Strong' artificial intelligence (AI) is the idea that an instantiation of a formal program is an instance of genuine intelligence. A formal program refers to a string of purely abstract symbols or tokens that are syntactically manipulated in a recursive or iterative fashion. The individuation of symbols occurs not via their immanent semantics but based on their orthographic characteristics or the functional roles they play within a formal system (Rescorla 2017). Hence, if a physical computer or a Turing machine can implement a formal system, then the implementation is taken to be a case of bona fide intelligence. Importantly, the physical substrate implementing the formal system is seen as secondary, if not altogether irrelevant. As long as the physical system is sufficiently complex to implement a given formal system, the latter can be realized by the former. Accordingly, formal systems are multiple realizable in a multitude of appropriate1 physical systems, and therefore, understanding the brain is secondary to understanding the computational mechanisms it implements. Proponents of this view think that implementations of formal systems constitute duplications, not mere simulations ('weak' AI), of real intelligence.

1. It is often argued that for some physical system to be an 'appropriate' realizer of a formal system, the physical system must play the same causal roles (i.e., have the same initial states, undergo the same state transitions, and produce the same output states) of the formal system in an isomorphic manner that is counterfactually-supported.

Alan Turing proposed in 1950 an 'imitation game' (or Turing test), which entails a human judge conversing with another human and a universal Turing machine.2 The judge's task is to correctly determine which of his interlocuters is a machine and which is a human. The probing conversation takes place over written text, so that voice, physical characteristics, gestures, and other non-linguistic elements do not tip the judge into finding the right answer. If the judge is fooled by the machine into thinking that it is a human, then the machine is said to have passed the test; it has successfully managed to imitate a human (Turing 2009). More often than not, the proponents of 'strong' AI take the passing of the Turing test as good reasons for believing in the existence of genuine machine intelligence (see, Oppy and Dowe 2019, for details).

Let me now explain the (in)famous Chinese room argument (cf. Cole 2019), followed by an explanation of its implications for 'strong' AI. John Searle (1980) invites us to imagine a monolingual Englishman situated in a room. He receives through an input slot a piece of text that is foreign to him; he then receives another piece of foreign text. Later, he finds a set of instructions, written in English language, that he can comprehend. These instructions tell him to place, say, symbols x after a, y after b, and so on. He diligently follows these instructions to string together sets of intricate symbols whose meanings he does not understand. Finally, the instructions tell him to insert in the output slot set p after set q and so forth. This is, in cursory terms, the experience of the man in the room: he is reading a book in English that tells him how to identify (based on physical characteristics) certain symbols and where, in the strings of symbols, each one belongs and when it is appropriate to put each symbol-string in the output slot. Despite becoming adept at following instructions, the Englishman has no idea about the meanings of the symbols.

Unbeknownst to him, the symbols actually belong to the Chinese language. Outside the room, there are native speakers of Chinese that are inserting the first batch of symbols, which may be understood as a story written in Chinese. Then,

2. Two results from mathematical logic lead to universal Turing machines. First is the Church-Turing thesis, which states that for any possible algorithm, there exists a Turing machine that can, at least in principle, implement it. Second, the Turing thesis states that a universal Turing machine can imitate any given Turing machine. Hence, a universal Turing machine is able to implement, at least in principle, any and all possible algorithms (Searle 1990). A universal Turing machine, then, is a good candidate for possessing domain-general cognitive capacities, because it can simultaneously imitate a number of different Turing machines with domain-specific capabilities.

they insert a second batch of symbols that is analogous to asking questions about the story they initially presented. The Englishman then manipulates the symbols in line with the English instructions, which are analogous to the abstract program the man is implementing. The strings of symbols, whose meaning I should emphasize he does not understand, that he places in the output slots of the room are interpreted by the outside native Chinese speakers as answers to the questions. Given how proficient the Englishman had become at manipulating symbols, from the perspective of the native speakers whatever is answering the questions inside the Chinese room 'black-box' understands Chinese very well.

However, as the experimental set-up makes clear, the man has no understanding of Chinese whatsoever. For him, the symbols may have belonged to Japanese, Dutch, C++, or no language at all. The important point that Searle takes the experiment to show is that since semantics do not inhere in the symbols, simply implementing a computational program, which is nothing more than a set of formally defined symbols that are syntactically manipulated, is not sufficient for understanding. The man clearly does not understand Chinese even though he can effectively perform syntactical manipulations with such adroitness that even the natives think that the 'processing' in the black-box (the Chinese room) is of the nature that there is understanding of the Chinese language.

The conclusion of the Chinese room argument—namely, that semantics do not inhere in symbols, which are defined formally and manipulated syntactically—is used as premise three in the following overarching argument that Searle (1984) makes against 'strong' AI: (P1) programs are defined purely formally or syntactically; (P2) human minds have mental content or semantics; (P3) syntax by itself is neither constitutive of nor sufficient for semantic content (Chinese room argument); (P4) so, programs are neither constitutive of nor sufficient for semantics; (P5) universal Turing machines implement abstract programs that are purely syntactical; (P6) thus, there are no inherent semantics in computers; (C) thus, 'strong' AI is not an instance of genuine cognition.3 As this deduction shows, the crucial premise in the argument is (P3), which is established by the Chinese room argument. I think it is imperative for the proponents of 'strong' AI to refute this premise in order to have a philosophically sound basis for procuring a computational research paradigm in cognitive science. Debates around the soundness of Searle's (1980)

3. The underlying, though plausible, assumption is that genuine cognitive agents have original intentionality. See the section on 'Other Minds Objection' for details.

argument will be the focus of the rest of this paper. I will review some major attempts at refuting (P3), and I shall respond on Searle's behalf to show that they all fail, thus preserving the argument laid out in this paragraph to conclude that 'strong' AI chronically suffers from the problem of original meaning.

DISJOINT PERSONALITIES OBJECTION

Searle's (1980) original paper anticipates a 'Systems Reply' according to which the room as a system understands Chinese even though the Englishman as its constituent does not. Searle simply replies that if the person memorizes the rule-book and all the information necessary and sufficient to effectively manipulate tokens, he would still have no understanding of Chinese even though he would have in his mind everything that the 'system' also has. Some interesting modifications were later made to the 'Systems Reply,' and I review and respond to them below.

Cole (1991) considers a thought-experiment in which the Englishman inhabits a joint Chinese-Korean room. In the morning, he may be implementing the program such that the 'answers' he gives are provided to Chinese speakers, and in the afternoon, he may be running the program in a way that the 'answers' are given to Korean speakers. Again, the Englishman is ignorant of the meanings of the symbols he is manipulating in accordance with a rule-book; what is more is that he does not know that he is manipulating two different types of linguistic tokens at different times of the day. Now, suppose that the 'answers' given to the two types of speakers display completely different psychological profiles: in the one case, the profile may be very amicable and polite, while in the other case, it may be aggressive and hostile. (Also suppose that the answers are given in such a way that an onlooker is convinced that the black-box does not understand a language other than that of the onlooker, e.g., by denying knowledge of the other language.) Suppose also that the Chinese and Koreans who attend this, say, 'festival' of sorts converse with each other later at night; they talk about their experiences of visiting the room and the attitude displayed by the 'answers' from the room. The behavioural evidence available to the speakers of both languages is markedly different, and so they conclude that there are two non-identical minds in the room. Since these minds have mutually exclusive psychological properties, they "cannot be identical [with each other], and ipso facto, [they] cannot be identical with the mind of the implementer in the room" (Cole 2019,

§ 4.1.1). Maudlin similarly observes that "Searle has done nothing to discount the possibility of simultaneously existing disjoint mentalities [that are different from each other and from that of the syntactical manipulator]" (1989, 414-15). This argument shows that since there can be psychological personalities different from that of the token manipulator, there is something in the room or the system as a whole that is not entailed by the psychological make-up of the Englishman inhabiting it. Accordingly, Searle is premature in asserting that the man's inability to understand Chinese constitutes that there is no understanding of Chinese.

I have three responses to this argument (presented in increasing degree of strength). The first is that the man in the room is manipulating symbols that he still has no understanding of. Instead of using tokens that were only in Chinese, he is now manipulating Korean tokens as well. This, no doubt, will produce different understandings in the native speakers who independently observe the room from the outside, but Searle's original claim that the man understands nothing still stands. It is similarly pertinent to observe that, rhetorically speaking, there is a certain element of magic associated with understanding being created in the "system as a whole." I am not sure where such understanding inheres if not in the mind of the only conscious and intentional entity present in the room (the Englishman).

The second reply I have asks the reader to imagine a trilingual man capable of speaking English, Chinese, and Korean. He meets the Chinese speakers in the morning and the Korean speakers in the afternoon, and just like the man in the joint-room, he exhibits (for whatever reason) different personality traits to the two types of speakers. When the Chinese and Koreans talk at night about meeting an Englishman that day, they do not think that they are talking about the same person (for how one person can be amicable in the morning and bellicose in the afternoon is difficult to comprehend) but about two people with different personalities. Given that they can erroneously and unknowingly think of the same person as having two different personalities, it does not follow that the Englishman indeed has two distinct personalities. If we are to consistently apply Cole's and Mauldin's arguments, it follows that the two personalities exhibited by the person are different from who he is. This strikes me as prima facie absurd to say that his two different personalities are not his own simply because the onlookers thought so (admittedly, based on evidence). Hence, I am inclined to reject their arguments.

My third reply rests on a distinction between a stronger and a weaker version of the Chinese room argument. The stronger version, which Searle proposes, maintains that understanding or personality of the room is constitutive of or identical with that of the symbol manipulator; understanding or personality of the room is nothing 'over and above' what the man in the room understands or exhibits, respectively.4 The weaker version, which we can consider here for the sake of argument, does not maintain a relationship of constitution or identity but that of supervenience. On this account, understanding in the room supervenes on the understanding of the Englishman. So, while the room may have a change in its understanding only if there is a change in man's understanding, it thereby does not mean that the two understandings are identical. So, the joint-room may exhibit personalities that are numerically non-identical from yet causally dependent on that of the person. Despite maintaining the weaker relationship of supervenience, it cannot be shown that the joint room has had any change in understanding or personality without a corresponding change in the Englishman. For the supervenience relationship to work, the lower-order organization (the symbol manipulator) must change in order for it to cause a change in the higher-order organization (the room as a whole) (McLaughlin and Bennett 2018)5; yet, even in the weaker version, the man not understanding the languages shows the untenability of Cole's and Mauldin's critiques, namely the room cannot have a change in its understanding or personality without a change in these characteristics with respect to the Englishman, thus vindicating Searle's arguments against these attacks.

OTHER MINDS OBJECTION

This objection essentially states that because we rely on behavioral information to attribute mentality or cognition to other people and animals, the native speakers observing the Chinese room or the human judge conversing with the Turing machine should likewise ascribe mentality to them due to their access to only the behavioral information. If we are to apply our epistemology

4. Indeed, this is the crux of Searle's (1980) response to the 'Systems Reply.'

5. A classical example is that of a painting, which has aesthetic properties organized at a higher-order and physical properties organized at a lower-order. The former properties supervene on the latter properties, so any change in the aesthetic qualities of a painting is brought about only through a change in its physical constituents.

consistently, the argument goes, then all entities, whether humans or machines, should be treated in the same manner—knowledge based on which we consider other humans as mental should similarly be sufficient to deem the Chinese room as cognitive. Otherwise, we are engaged in anthropocentric chauvinism.

My response to this objection is that there is first and foremost a difference in the degree/ amount of behavioral information available to the human being in charge of ascribing mentality. In the cases of the Chinese room and Turing test, the information that is available to us is in the form of language. While language is no doubt an important part of cognition, it should not be identified with it. When we ascribe mentality to other humans, however, we implicitly rely on a whole range of behavioral information at our disposal, including language, gestures, eye gazes, facial expressions, intonations, and so forth. The manifold data make it far easier to think that other people are conscious, but the machines may be pre-programmed to spit out certain linguistic phrases in light of questions. There is simply not a sufficient amount of behavioral evidence available that can justify the ascription of mentality to the Chinese room or the Turing machine.

The interlocuter may, however, rightly protest that it is 'just' a matter of time or technological advancement that we will create robots that are capable of producing behavioral outputs that are just as complex as those of humans. Films on artificial intelligence like Ex Machina already exploit our intuitions in this respect. So, a philosophical argument must rely on a difference in principle, not on a difference in degree, to explain why we cannot, assuming we have equal amounts of behavioral evidence for robots and humans, ascribe mentality to both. Furthermore, this argument about applying epistemology consistently can be extended from the pragmatic to the scientific domain. Dennett (1997) has argued that in the Turing test we should utilize what he calls the 'quick-probe' assumption. The idea is that since a machine must choose from a number of different possible responses that may be given to questions of an interrogative judge, it cannot utilize brute-force computation, because at each linguistic answer-node, a number of other topics are opened up that may need to be addressed, leading to combinatorial explosion. Thus, he argues that the act of providing an intelligent response without brute-force computation is good reason for thinking that the Turing machine has some kind of cognitive capacities. And from here it is not unreasonable to generalize that the machine may also be capable of exhibiting other mental abilities; this assumption, then, is used to quickly probe the mental

capacities of the Turing machine. Likewise, the native Chinese speakers looking at the room from the outside may construe it as a Turing machine answering questions; there is no reason for thinking that the room has no understanding, because the linguistic (or behavioral) outputs of the room are no different from those of other people. If our epistemology is applied consistently, then the Chinese room has internal meaning just as other people do (or else, other people are not cognitive either!).

I respond to this objection by first pointing out that there is a conflation between nomological or metaphysical facts with epistemological facts. When we are concerned with comprehending whether a system truly has understanding, we are interested in uncovering the mechanisms of its internal processing (Block 1981). The notion of ascribing mentality to a system as opposed to discovering it as inhering within it are two different things—only a behaviorist would be content with thinking that behavioral dispositions are all there is to having cognition. It is quite possible for things to behave intelligently (a parrot mimicking human language) without actually being intelligent (a parrot not understanding the words it uses). I suspect there is a fallacy of reverse causality underlying the interlocuters' claims: from the fact that intelligent things behave intelligently it does not follow that all those things that behave intelligently are thereby intelligent. The causality here is unidirectional: only intelligent things behave intelligently, not vice versa.6

6. An analogy is that a disease should not be confused with its symptom. Surely, the removal of a disease leads to the removal of its symptoms, as the former is the cause of the latter. However, just because the symptoms are removed, one cannot think that the disease is also gone (even though the symptoms may be used as indicators for the existence of the disease). There is no biconditional in this case. One may represent this formally as a modus ponens argument. Let the disease (or cognition) be p and the symptom (or intelligent behaviour) be q.

$p \rightarrow q$
p
$\therefore q$
The interlocuters are committing the fallacy of 'affirming the consequent':
$p \rightarrow q$
q
$\therefore p$

(Curiously, one sees in this modus ponens argument that the same symbols, p and q, can at one time stand for disease and symptoms, and they can at a different time stand for intelligence and intelligent behaviour. The rules of inference are the same irrespective of which semantics are ascribed to the symbols. So, even this argument shows the correctness of Searle's observation, namely that the meaning does not inhere in the symbols; it is simply ascribed to them.)

One may at this point say, "well, how do we then know that something really is intelligent if not due to the manifestations of intelligent behavior?" This leads me to the key point of the argument. We know for certain that what is going on in the Chinese room is nothing more than manipulations of formal symbols; we also know that the outsiders have access only to the linguistic outputs which they interpret as being a result of genuine cognitive activity. Since we, as philosophers thinking about the Chinese room thought-experiment, are already aware of what is taking place in the room (i.e., we have access to the matters of facts of the room's internal processing), we have no reason to believe that there is any understanding taking place in the mind of the Englishman or in the room as a whole.

Unlike in the case of humans, where we try to discover the nomological facts about the workings of the brain through neuroscience etc., in the case of computational machines implementing programs we are already aware of the principles underlying their workings. So, there is no reason to think that in the machine there is anything 'over and above' what we already understand. Searle (1984) is correct in pointing out that in the sciences we presuppose the existence explicandum. Here, in trying to understand the basis of cognition, we presuppose that human minds exist just as physicists presuppose the existence of physical things. The 'other minds' objections seem to place the cart before the horse. We already know how the computational systems work; we do not know how the human minds work, and this is something that needs to be explained. There is no reason to place two things on the same epistemic grounding when the metaphysics of both are asymmetrically known.

CONCLUSION

I have tried to show that Searle's argument that semantics do not inhere in formal symbols successfully survives the objections I have considered herein. Consequently, the plausibility of his argument is threatening to the prospects of 'strong' AI as a viable research programme, because one of the features that is crucial to human minds—and, therefore, something that any scientific theory of the mind must explain—is that they have inherent meaning.

REFERENCES

Block, Ned. 1981. "Psychologism and Behaviorism." *The Philosophical Review* 90 (1): 5–43.

Cole, David. 1991. "Artificial Minds: Cam on Searle." *Australasian Journal of Philosophy* 69 (3): 329–33.

———. 2019. "The Chinese Room Argument." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University.

Dennett, D. C. 1997. "Can Machines Think? Deep Blue and Beyond." *Icca Journal* 20 (4): 215–23.

Maudlin, Tim. 1989. "Computation and Consciousness." *The Journal of Philosophy* 86 (8): 407–32.

McLaughlin, Brian, and Karen Bennett. 2018. "Supervenience." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2018. Metaphysics Research Lab, Stanford University.

Oppy, Graham, and David Dowe. 2019. "The Turing Test." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University.

Rescorla, Michael. 2017. "The Computational Theory of Mind." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–24.

———. 1984. *Minds, Brains and Science*. Harvard University Press.

———. 1990. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64 (3): 21–37.

Turing, Alan M. 2009. "Computing Machinery and Intelligence." In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, edited by Robert Epstein, Gary Roberts, and Grace Beber, 23–65. Dordrecht: Springer Netherlands.