The Chinese Room, and What It Takes to Build a Mind

In the past several decades, computer scientists and artificial intelligence researchers have made great progress in developing computer programs with the ability to mimic processes of human cognition. With increasing speed, complexity, and processing power, these programs show a striking resemblance to the human mind. Some even claim these technologies further "emphasize the analogies between the functioning of the human brain and the functioning of digital computers" and promote a functionalist conception of the mind[1]. Within philosophy of mind, functionalism is a doctrine generally stating that what makes something a mental state does not rely on its physical composition, but solely on its function, or the role it plays within a larger cognitive system. All forms of functionalism, in one way or another, claim that mental states are those states which are causally related to so-called inputs, outputs and internal roles – or in other words, sensory stimulations, behavior and other mental states. In his 1980 essay, "Minds, Brains and Programs," philosopher John Searle questions the functionalist conception of the mind by asking, "what psychological and philosophical significance should we attach to recent efforts at computer simulations of human cognitive capacities?"[2]. Generally speaking, functionalist theories of the mind hold that if two things such as human brains and digital computers are functionally similar, they are also psychologically similar. Searle rejects this claim by arguing appropriately programmed computers with the right inputs and outputs *are not* minds, and cannot be said to understand and have mental states in the same way humans can. Using what is called the "Chinese Room counterexample," Searle attempts to show the absurdity of functionalism by giving an example of a system that is functionally similar to the human mind, but different he believes psychologically. Searle hopes to describe a scenario in which all plausible versions of functionalism agree that understanding of Chinese – and hence mental states exist, but our intuitions prevent us from believing so[3]. Searle's counterexample has been extensively critiqued since it was described in 1980, and many replies to the Chinese Room are included in "Minds, Brains and Programs." I believe two of these replies, the "Systems Reply" and the "Robot Reply," provide the best forum for

[1] Searle, J.R. Minds, Brains, and Science. Harvard University Press, Cambridge: 1984. Page 28.
[2] Searle, J.R. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences*. Cambridge University Press. 1980. (3) Page 417.
[3] Braddon-Mitchell, D. and Jackson, F. Philosophy of Mind and Cognition. Blackwell Publishing, Malden: 1996. Page 111.

exploring the strengths and weaknesses of Searle's argument against functionalism. Through my analysis of the Systems Reply, I intend to show that no digital computer could ever understand any human language because doing so would require semantic ability beyond the capacity of the computer. I will also use the Systems Reply as a means to clarify which functional roles are important for understanding, and therefore important for having mental states in general. I believe that Searle's argument ultimately fails to invalidate functionalism because it does not describe a functional duplicate to the human mind. Using an embellished version of the Robot Reply, I will show that the Chinese Room as described by Searle does not have the kind of causally interactive internal roles we associate with minds, and therefore poses no threat to functionalism.

Searle begins "Minds, Brains and Programs" by describing a program, developed by Roger Schank, with the ability to *simulate* the human capacity to understand stories. A characteristic of this capacity lies in our ability to answer questions about information not explicitly stated in the stories. "Schank's machines can similarly answer questions in this fashion […] to do this they have a 'representation' of the sort of in formation that human beings have, which enables them to answer such questions given the stories"[4]. Using Schank's program as a model, Searle's Chinese Room counterexample to functionalism goes as follows. Searle asks us to imagine an English-speaking man, who I will call Hal, sitting alone in a room. Hal has absolutely no understanding of Chinese, and would be unable for that matter to distinguish Chinese characters from meaningless scribbles. Two slots, one where papers are fed into the room and one where papers can be sent out, are his only contact with the outside world. Hal is fed stories and questions in Chinese through the input slot. With an instruction manual written in English, he manipulates the meaningless characters he sees, makes various calculations, and ultimately writes down more meaningless characters. Once he is done writing, he feeds his work through the output slot. When read by a native-Chinese speaker, Hal's writings appear to be intelligent answers to questions about the stories written by someone who *understands* Chinese. The crux of Searle's argument is that to a bystander reading his work, Hal appears to understand Chinese, but we know well that Hal has no understanding of Chinese at all. If functionalism is true, and indeed the Chinese Room is functionally

[4] Searle, J.R. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences*. Cambridge University Press. 1980. (3) Page 417.

similar to our minds in the relevant ways, then Hal *must* understand Chinese. Therefore, since Hal clearly does not understand Chinese, functionalism must be false. Troubled by this conclusion, many philosophers have responded to Searle's argument in attempts to explain how understanding is in fact generated in the Chinese Room counterexample.

Two of the replies included in "Minds, Brains and Programs," the "Systems Reply" and the "Robot Reply," defend functionalism by showing us either that understanding does exist in the Chinese Room or what would be required for such to be the case. Though neither succeeds in completely undermining Searle's conclusions, I believe they indicate both what would be required for understanding and why Searle's argument against functionalism ultimately fails. The "Systems Reply" asserts "while it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand"[5]. In other words, Hal does not understand Chinese, but the entire system – encompassing Hal, the instruction manual, and any other materials used by Hal to produce his final responses – does understand Chinese. First of all, I find the idea absurd (as does Searle) that if Hal clearly does not understand Chinese, somehow the conjunction of Hal, the instruction manual, and any other materials in the room might understand Chinese[6]. Searle however, entertains the objection and responds by suggesting that Hal memorize the instruction manual and do all of the character manipulations in his head. Searle believes even in this new case, Hal encompasses the entire system and still does not understand Chinese. Many philosophers have argued that in this new version of the Chinese Room where Hal has internalized all elements of the system, the system (which does understand Chinese) would continue to exist – just as a component of Hal's mind. Analogies of the scenario have been given – either of two personalities inhabiting the same brain or of one computer emulating the behavior of a different machine. I do not find this strand of argument against Searle's response to be compelling. This is because these analogies are set up under fundamentally different circumstances; namely, not involving the creation of a new entity through memorizing a book. It goes against our intuitions that Hal's memorizing the instruction manual should create an extra entity with the capability to

[5] Searle, J.R. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences*. Cambridge University Press. 1980. (3) Page 419.
[6] *Ibid*.

understand Chinese. If the instruction manual and other elements in the Chinese Room do not constitute a mind with the ability to understand Chinese in the original case, it does not seem that they should then constitute a mind when internalized by Hal. Therefore, I believe that because the system lacks some fundamental property necessary for understanding, the Systems Reply cannot explain away the conclusions of Searle's argument.

Although the Systems Reply does not refute Searle's argument against functionalism, it does help clarify what is required both for understanding and having mental states. Understanding a language such as Chinese requires more than the ability to manipulate characters, even if those manipulations are able to produce intelligible answers to questions in Chinese. It seems, as illustrated by the Systems Reply, that Hal (or the entire system) will only understand Chinese when the characters are no longer recognized as mere scribbles; they must be formed into words that carry meaning. Imagine what would happen if Hal were to read a sentence in English. He would immediately see beyond the shape of the letters composing the words, he would comprehend their meaning. In this sense, Hal understands English because he has a mastery of English semantics. It is clear Hal does not have the same semantic mastery of Chinese. He is only able to recognize the structure of Chinese characters and follow instructions for manipulating these characters. Thus, it could be said Hal has some syntactic ability, but clearly not semantic ability. Searle's Chinese Room counterexample succeeds in showing that "understanding a language, or indeed, having mental states at all, involves more than just having a bunch of formal symbols. It involves having an interpretation, or a meaning attached to those symbols"[7]. In response to the Systems reply, it is clear that regardless of whether the system is external or internal to Hal, the fact still remains that predefined and purely abstract steps are used to generate his responses to the stories in Chinese. Therefore, the Chinese Room counterexample shows us that no computer program (by today's standards of programming) could ever understand language, because these systems are purely syntactical. They are based on a formal structure of rules, and lack the sort of semantic content contained within human minds. Even in light of the Systems Reply, it appears that the system contained in the

[7] Searle, J.R. Minds, Brains, and Science. Harvard University Press, Cambridge: 1984. Page 33.

Chinese Room does not understand Chinese. However, in order to show the absurdity of functionalism, Searle's argument must also establish that the Chinese Room counterexample describes an entity that is functionally identical to our minds. Thus, as we have seen, it must describe a program with mastery of semantics.

Our understanding of English semantics developed for each of us early in our childhood as we gained experience with the world. If mastery of semantics is necessary for understanding language, a system must be able to interact with the environment in order to acquire semantics through learning. The "Robot Reply" can be seen as an attempt to create a scenario where through interaction with the environment, a syntactical system such as Hal in the Chinese Room can gain mastery of semantics In this sense, the Robot Reply and its embellished forms transform the Chinese Room into an entity that is a true candidate for understanding and having mental states. Take for example, a fully embellished form of the Robot Reply presented by Braddon-Mitchell and Jackson in Philosophy of Mind and Cognition. Imagine a robot that transmits information about inputs from its visual, auditory, and other sensory components directly to Hal in the form of Chinese characters. Hal takes the characters from these inputs, and now working much quicker and using an even larger instruction manual, he performs the appropriate manipulations and calculations. Once Hal is finished, he transmits his response back to the robot, which cause the robot to interact with its environment in the same way we do. Imagine also that the instruction manual Hal uses is fashioned such that it takes into account all input and output characters, and how often they have occurred. This enormously complex instruction manual would include instructions on how Hal should alter it in response to previously received input, ensuring that novel responses are generated in each case[8]. With the increased complexity of this highly embellished case, we may begin to intuit that this robot is capable of understanding. After all, in addition to having functionally similar inputs and outputs, the robot's ability to produce novel behaviors in response to experience seems functionally similar to our capacity for memory. Indeed, this kind of "memory" would also seem to allow the robot to gain semantic mastery. Notice that this embellished example is so far removed from Searle's original Chinese Room scenario that it poses no threat to his argument. Without the

[8] Braddon-Mitchell, D. and Jackson, F. Philosophy of Mind and Cognition. Blackwell Publishing, Malden: 1996. Page 110.

embellishment, the Robot Reply would be no better off than the Systems Reply because as we have seen, the purely syntactic organization of the Chinese Room is incapable of generating understanding. However, with even more embellishment, the Robot Reply can describe an entity that all plausible forms of functionalism would agree understood and had mental states. Contrasting the entity in Searle's original counterexample to this embellished Chinese Room-Robot, we can see why Searle's argument against functionalism is unsuccessful.

Searle's Chinese Room counterexample to functionalism fails to describe an entity with functionally identical inputs, outputs and internal roles associated with our minds. Indeed, certain forms of the Robot Reply create unembellished examples of the Chinese Room with functionally identical inputs and outputs, including the ability to have sensory experiences and interact with the environment. However, these forms are unable to produce the sorts of functionally identical internal roles important for understanding of language and having mental states. The Systems and Robot Replies to Searle's argument have shown that in order for an entity to understand a language, it must master the language's semantics. This can only be accomplished through learning and memory, by which words become associated with meanings. The entity in Searle's original counterexample lacks the internal roles necessary for learning and memory. For instance, it is incapable of accounting for all previous inputs and outputs, and generating novel responses in each case. All plausible forms of functionalism and our intuitions agree that Searle's original Chinese Room is not a candidate for understanding and having mental states. Our intuition is that only the highly embellished forms of the counterexample (those much different from Searle's original counterexample) have these capacities, and are able to understand Chinese and have mental states. Only these highly embellished Chinese Room-Robots can begin to have the sort of functionally identical internal roles (such as those involved in learning, memory, and other processes) required by functionalist theories to be psychologically similar to us. Therefore, it is impossible for Searle's Chinese Room counterexample to prevent our intuitions from agreeing with functionalism.