Macalester Journal of Philosophy

Volume 6
Issue 1 *Spring, 1996*

Article 7

2-17-2011

Searle, A Brain, A Mind, and Two Thought Experiments

Bonnie Plottner
Macalester College

Bonnie Plottner

"Searle, a Brain, a Mind, and Two Thought Experiments"

In Philosophy of Mind, it seems that thought experiments are as popular as empirical experiments. Various authors try to tweak one's intuitions with their imagined stories of operations, computer programs, and mechanical cats. They hope that the experiment is open to only one interpretation--the one that supports their argument. Searle's Chinese Room experiment is very much in this vein. Although many parts of the thought experiment are not explained fully, Searle attempts to jump right into his conclusion. Although I agree with his conclusion to an extent, that a very high level functionalism will not work, I do not think this is shown by the Chinese room example. To this end, I will expose its problems and explore its fuzzy parts. My positive argument is not as complete. Nonetheless, I hope to show that a low-level functionalism, along the lines of Douglas Hofstadter's "A Conversation with Einstein's Brain," takes the best of Searle and functionalism and combines them into a more likely brand of artificial intelligence.

In the Chinese room experiment, Searle imagines himself, as an English speaker, sitting in a room. With him are many, many bits of paper that have Chinese characters on them, none of which he can understand. Also, he has an algorithm for taking in some symbols from the outside, putting together symbols in the room, and then sending them out. Unbeknownst to Searle, these packs of symbols are meaningful sentences. The room seems to be carrying on a conversation in Chinese, although Searle speaks not a word of it.

Searle believes that this experiment shows that although the syntax of Chinese is given, and completely executed by the algorithm (that is, outsiders are fooled by the room and think it speaks Chinese), there is no semantics. The room cannot be intentional, since Searle is the one doing all the work and he does not have any idea what he is saying with all those symbols. Searle has no understanding, conscious or unconscious, of Chinese; hence, the room cannot either. No one on the outside can discover this, since the room will have all the right Chinese-speaking behaviors. Searle's conclusion is that language, and the mind as a whole, cannot be simply a syntactic algorithm. Since we, as humans, have understanding and intentionality, our minds cannot just be programs running on the brain. Rather, the mind must somehow be a biological result of the mechanics of the brain.

This thought experiment is not as clear-cut as it may seem, however. It leaves several questions unanswered. The first is, what is Searle's importance in the Chinese Room? This experiment is supposed to show that a computer could not think, or be conscious. In a computer, the program is run on hardware; here, the program is run by a conscious human being. It seems central to the experiment that Searle is conscious and yet claims to not know Chinese. In a computer, the hardware could not be asked such questions. Although Searle understands the instructions in English, but not any of the Chinese parts, a computer is not the same. It does not "understand" machine language or assembly language, or any upper-level programming language, any more than Searle understands Chinese. Although this may seem ever more damning to the

computers, it in fact shows a problem with the thought experiment. In fact, it demonstrates that a brain is more like a computer than a full-blown person. In a human, the brain itself is not an independent conscious entity. The brain, in isolation, does not know English, or any language, any more than the computer hardware knows a computer language. The brain of a human is basically useless without the rest of the person there. We do not care what it is the brain has to say; we want to know what the person has to say. In the Chinese Room, Searle tries to convince us that since he cannot understand Chinese, the room cannot. But it is his very presence in the room that is misleading.

Although the brain carries on other activities while one is speaking (regulating breathing and heartbeat, etc.) and computers carry on unseen activities (memory allocation, incrementing the address register, etc.), these all have to do with the system as a whole. In the Chinese Room, Searle carries out the algorithm, but he also amuses himself, thinks about his wife, eats dinner, etc. These activities are unrelated to his function as Chinese room worker. Hence, he has posited a homunculus (himself) that has understanding in the Chinese room. He hopes that we will be swayed by the point of view of this homunculus into believing that the Chinese room, and hence any computer-type mechanism, cannot be thinking or understanding. However, such a conclusion depends on the existence of this homunculus which is not posited, nor indeed present, in computers themselves. His conclusion rests on us identifying with the homunculus, instead of the room. But it is the room which is our behavioral twin. Hence, such a characterization as follows is misleading:

> Now the point of the story is simply this: by virtue of implementing a formal computer program from the point of view of an outside observer, you behave exactly as if you understood Chinese, but all the same you don't understand a word of Chinese (Searle 1984, 32-3).

Note that Searle thinks the important part of the Chinese room is the homunculus inside. If we rethink the above statement as "the Chinese room behaves exactly as if it understood Chinese, but all the same the brain inside does not understand a word of Chinese," it no longer seems so shocking and conclusive.

This argument against Searle has been given many times before, although I hope that my presentation of the reply is somewhat original. The "systems reply" suggests that although Searle cannot understand Chinese, the system of which he is only a part can. Searle mentions and dismisses such an argument in everything he has written about the Chinese room experiment. However, the difference is that Searle abbreviates this idea into some form of: "I, Searle, am in the machine and do not understand a word of Chinese, but the room, plus a bunch of meaningless squiggle-squoggles, plus a book of rules can understand Chinese." While such a characterization indeed seems easily dismissed, a more complete exploration of both the experiment itself and what the systems reply consists in makes Searle's argument less and less acceptable.

There are other problems with this experiment that are associated with the systems response. For instance, Searle does not explore in depth the computer program that he has posited. He describes it as:

> ...a program that will enable a computer to simulate the understanding of Chinese. So, for example, if the computer is given a question in Chinese, it will match the questions against its memory, or data base, and produce appropriate answers to the questions in Chinese (Searle 1984, 32).

As described, this program sounds simple. Any program that could actually fool all Chinese speakers would have to do more than just check answers against a data base. Although this type of approach is one in Artificial Intelligence (AI), such as in expert systems, most people do not have huge banks of information saved and ready to spew forth when asked the right question. A computer program that truly matched human behavior would have to be able to make up new ideas, think to itself, make statements that come out of nowhere, initiate conversations, and more. If we change the Chinese Room by describing the program as one that sometimes sits idly, sometimes asks people questions about the universe or God, comes up with a new way of presenting information, and all the other things that people can do, what we ought conclude is not as obvious. Although Searle claims that the room could pass the Turing test, he then describes a program which just answers questions. When we think about all the kinds of questions the room could answer, dismissing it is not as easy.

However, I am not sure that such a complex program can be written. Here is where I actually agree with Searle. A person has many interrelated activities and thoughts that do not seem clearly separated into functions. Writing a program that understands Chinese, and then adding on a bunch more that describe the room's opinion of politics, religion, art, and sports, a few more that describe its hobbies, some to regulate when it gets tired and when it is sad, seems to be an almost endless task and very ad hoc. Hence, although dismissing such a program is not as easy, I think one can still be justified in doing so.

The final problem I have with this experiment is Searle's contrast between the computer as not understanding and the human as full of understanding. He claims that the computer is only syntax, no semantics. As a regular person, though, "You understand the questions in English because they are expressed in symbols whose meanings are known to you" (Searle 1984, 33-4). This idea of "meaning" is taken as a primitive. It is assumed that we all mean something, and that this act of meaning something does not merit explanation. This seems to be the biggest question of all, however. Where does this meaning come from? How is it that for English the symbols are known to us? Searle claims that if the homunculus does not understand the semantics, then there is no way for the system to understand them either (Searle 1984, 34). However, this does not seem clear at all.

Searle has claimed that "Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain" (Searle 1992, 1). In relation to understanding, this suggests two interpretations: 1) the brain does not have semantic knowledge, but the mind does as some emergent feature of the brain, or

2) the brain has semantics. Although the first seems more in keeping with Searle's philosophy, this would actually support (to an extent) an AI program. That is, since semantics can emerge from the natural working of the brain, it could probably also arise from the working of a computer. At least, the argument given above will not be as decisive. Therefore, Searle must support some such interpretation as number 2. Yet, how can the brain have semantics? What could it mean for the brain to "mean" things?

The first problem with the brain meaning things concerns how meaning could be a part of biology. Searle believes that everything about the brain is explicable by biology. Yet, what kind of biological law will explain semantics? How will a biologist studying the brain discover semantics? Will meaning be like a Krebs Cycle, or the regulation of hormones in the body? Will there be some "semantic" neurotransmitter or process? These seem unbelievable, if not impossible, yet Searle wants us to believe that understanding and other "mental events and processes are as much a part of our biological natural history as digestion, mitosis, meiosis, or enzyme secretion" (Searle 1992, 1). Yet, how will we know when the researchers have discovered semantics in the brain? "Meaning" does not seem to be a brain or mind process in the sense that "seeing" or "feeling" may be. Such an approach seems to lead to two possible conclusions. One, along the lines of Roger Penrose, suggests that we need to find a new science to account for things like consciousness. The other, more pessimistic, follows Colin McGinn and suggests that whatever the connection between mind and brain, it is not to be discovered by humans. Let me reiterate, however, that we arrive at such conclusions only if we assume that semantics are to be found in the brain.

The other problem is more insidious: how are semantics possible at all in a materialist world? That is, there seems to be a difference between a computer printing out, "Have a nice day," on the screen whenever you boot it up and me saying, "Have a nice day," to a friend. Or (borrowing from Putnam 1993), if someone who has never seen trees, has no idea of trees, has no plants at all or anything that looks like trees, sees a picture of a tree, it has no meaning for her. If she drew an abstract picture (for them) that looked like a tree to us, it still would not be a picture of a tree. Yet, if we drew that picture, it would be of a tree. Can this all be in the head--some inherent feature in the brain? It seems the answer must be no.

Consider Putnam's objection to the "meaning in the head" idea (Putnam 1986, 110-11). He proposes two children, Oscar and Elmer, who grow up in Ruritania, the first in the south and the other in the north. In both areas, "grog" refers to the metal that makes up pots and pans. In the north, that metal is silver and in the south, it is aluminum. Their beliefs concerning grog, especially when they are young and not acquainted with chemistry, are likely the same, such as: "My mother has grog pots and pans," "Grog is gray and shiny," or "Grog is a metal." What makes the meaning of such a statement about grog different, as we can see from the outside it must be (since the words translated into English would be different)? There is no reason to assume that the state of the brain is any different for the two beliefs. Especially for Searle, there cannot be deeply unconscious facts that make the meaning different. It seems that the difference is that one lives in the north, the other in the south; that is, the context is different, although the speakers do not know that this matters. Their

beliefs about their living conditions do not include that grog is different in the north and south; hence, even all their beliefs together underdetermine the meaning of "grog." Yet, since we can tell the meaning is different, there must be some way to account for this.

The idea of there being a "something" in the head that determines meaning is untenable. As Wittgenstein explained in his various works, meaning cannot be something assigned privately. Rather, meaning is a social construction. The word "grog" has meaning for the northern Ruritanians from how it is used, not from pictures inside their heads, and the same for the southern Ruritanians. As Putnam explains:

> But in my theory of meaning...what you do is you look at the whole community, and you look at the environment, and you regard differences in reference in the two communities as infecting the speech of individual speakers (Putnam 1986, 111).

This is similar to Wittgenstein's idea, although for Putnam the experts determine meaning, while for Wittgenstein the use of the general public seems more pertinent, and gives a possible interpretation for meaning that can stand in a materialist world. However, such an explanation has nothing special to do with biology, except that in general it is people who participate in society.

Another explanation for how meaning is possible is given by John Haugeland (1986). He claims that the difference between sentences on paper and sentences in the mind is the way they interact with each other. For example, if you write the premises of a syllogism out on paper, nothing happens. But, if you tell a person these same premises, the person is likely to tell you the conclusion. In this way, the semantics are a kind of interactive property. Haugeland presents this as an argument in support of a computational view of the mind, claiming that "it is the causal relations which must be present 'materially,' and not just formally, to breathe genuine semantic life into a structure of formal tokens" (Haugeland, 90-2). This view is somewhat different from the Wittgenstein/Putnam one given earlier. However, it also emphasizes context. The earlier view showed the importance of the community of the speaker/understander, and this view asserts the importance of the context of the other sentences. The causal power of semantic activity can also be realized in a computer or other machine, since it does not depend on anything special about the humanity of the person. In the syllogism example a computer could also "figure out" the conclusions of syllogisms, since semantics is a emergent feature of syntax working together.

These last two discussions suggest that meaning is not an intrinsic feature of humans, and in fact, could not be. Certainly, a perfect explanation of semantics is still needed, but even a rough sketch makes Searle's Chinese room argument a bit shaky. Searle's conclusion rested on the reader believing semantics to be intrinsic to humans, yet lacking in the computer. However, if we find a computer developed enough to participate in the community and use sentences in the context of groups, it seems that they would have as much semantics as we have.

Oliver Sacks discusses something like this community in his essay on Temple Grandin, an autistic professor. She has trouble in social settings, by her own admission, since she only hears what is said and cannot ascertain the importance of what is not said.

> It has to do, she has inferred, with an implicit knowledge of social conventions and code, of cultural presuppositions of every sort. This implicit knowledge, which every normal person accumulates and generates throughout life on the basis of experience and encounters with other, Temple seems to be largely devoid of (Sacks, 270).

This sounds a lot like Wittgenstein's idea of forms of life. The things we say depend on the experiences we have had and shared with others; they cannot be understood in isolation. Grandin reads as many books as she can to try to figure out why people behave the way they do. Then, in social situations she applies her formal knowledge to the people around her. However, this never completely works. She still feels uncomfortable around people and prefers the company of animals. "Regular" people are able to pick up on social cues since they are somehow primed to notice the underlying emotions of other people. By participating in such behaviors themselves, these people have shared frames of reference that Grandin can never have. The words they speak can only be fully understood against this background (or, form of life). Hence, Grandin will always be at such a disadvantage.

In all fairness, Searle almost gives such a characterization of meaning by positing a Background of capacities that are non-intentional (kind of like the Wittgenstein/Putnam community) and a Network of other beliefs and desires (almost like the Haugeland interactive semantics). However, he wants to claim that both of these are features of the brain and cannot be had by computers. In fact, he expressly argues against the idea of "scripts" in AI, which was an attempt to incorporate just such ideas as Background and Network into a program that heard stories and answered comprehension questions. That is, a given story could only be understood against a script of the typical ideas and relations behind similar stories. Searle dismisses such an idea, also with the Chinese Room, claiming that the script is yet another bunch of meaningless symbols (Searle 1981, 355). Hence, as close as Searle was, his seemingly groundless insistence that there is something biologically special and essential about humans keeps him from accepting AI.

Searle does offer another problem for AI, besides the Chinese room. He claims that computers cannot even have syntax, since syntax is not intrinsic to physics, and needs a homunculus to ascribe it. I have several problems with this argument. The first concerns natural laws. Indeed, they describe (rather than regulating in a direct sense) natural activities, such as gravity. A simulation of gravity indeed depends on such a law, but the law still describes what is happening. Also, gravity is not quite the same as the workings of the brain. Gravity seems to be intrinsic to material objects, since by definition every material object is subject to gravity's force, whereas the brain is a result of natural selection. So, we have (or will eventually, according to Searle) biological laws that describe the workings of the brain. Here, he

asserts that if we created a computer simulation of the brain, it would follow those same laws, rather than be described by them. I would characterize it differently. Both the brain and the program have come to follow the same laws, the former by evolution and the latter by the quicker human programmer, and they both can be described as following the same laws. In fact, in some ways the neurons seem to be like a program running on top of chemical properties of neurotransmitter chemicals, and electrical properties. (I will return to such an idea in a little bit.)

Searle also decries the idea of finding complicated patterns everywhere, but Hofstadter has the perfect answer to this complaint. He writes, "The problem is, in all these cases, that of specifying the code without knowing in advance what you want to read" (Hofstadter, 382). The computer does not fall to this problem, as we do not find it running a program and then arbitrarily attribute the mind pattern to that program. Rather, we try to write a program that does all the same things as the mind. Searle also objects that the computer depends on a human to interpret what it is doing, while the mind somehow self-interprets. As we saw in the earlier discussions on meaning, however, it is unclear how much the mind comes pre-interpreted and how much we depend on a society surrounding us to interpret us, and us, them. That is, if there were an immortal human that lived alone, he would not have evolved as far as a society of people that work together and change as a species has evolved. The idea of species is much more important to the whole concept of consciousness than Searle recognizes.

As mentioned above, I would like to explore the idea of neurons as program a bit more. Evolution teaches us that neurons were not a building block of the world. They are not even a part of the most primitive biological entities. Rather, neurons evolved as biological entities evolved into more complex forms. In this way, chemistry is the primitive level, and neurons are a special kind of object that run on top of these chemicals in the brain, e.g., serotonin. Neurons may be completely explainable by biology, yet also be something that could be described by a functional state. Neurons can be characterized by what action potentials they have; which other neurons they have synapses with; which neurotransmitters they emit, when, and to which neurons; and to which neurotransmitters they respond. They also have a set way in which any of these limits can be changed over time. This description in many ways seems to be a functional one. As long as the other functionally described neurons interact with the first neuron, it seems that we have a description of the brain that would satisfy a more limited functionalism (neuron functionalism instead of mind functionalism), yet also satisfying a more limited Searle view (special causal relations between the neurons are still important, but the chemical realization of the neurons is inessential). In fact, a thought experiment along these lines is offered by Hofstadter in "A Conversation with Einstein's Brain."

In this thought experiment, we imagine a huge book, each page of which corresponds to a neuron in Einstein's brain. The page lists the functional description of the neuron, giving the limits and changes described above. He also posits an algorithm for transcribing words that a person using the book may want to ask Einstein into neuronal impulses. That person could then follow each neuron and note what neurotransmitters it sends out and to whom, and any changes that occur, then follow each consecutive neuron, etc., until there are output impulses to "mouth"

neurons, when we can use a reverse algorithm on to find out Einstein's answers to our questions.

Hofstadter and Dennett suggest that this experiment is very similar to the Chinese room experiment, just a different setting of the knobs on our "intuition pump," but I disagree. This experiment is very different in that the program is really split up among billions of neurons. Each neuron runs by a fairly simple algorithm: Update the numbers; plug new numbers into connected neurons; Repeat. The bulk of the important information is in the functional characterization of each of the neurons. This is very different from the Chinese Room, in which most of the complexity arose out of the program, while I see the many Chinese characters as somewhat unimportant (other than keeping track of them, they did not figure significantly into the complexity of the algorithm). Also, we can assume that everything about Einstein's brain must be carried out (emotions, the signals that control the beating of the heart, etc.). Perhaps the person talking to Einstein would even have to enter some digestive inputs, sleep cycles, etc., so that the book did not go crazy from lack of sleep (or food). In contrast, Searle's program seemed to just talk and answer questions. In many ways, this program seems perfect.

Yet, neither Hofstadter in the article itself, nor Dennett in the comments following the article, weigh in as to whether the book is conscious or not. In fact, the book intuitively seems a lot less interesting a conversation partner than the Chinese Room. Hofstadter raises a few questions, too, including whether closing the book and leaving it on the shelf is like killing a person. There are also the Parfitian issues of having several duplicate (at least, to begin with) books in circulation, or of whether we should be disappointed if we are about to die, yet our brain will be immortalized in such a book. None of the answers seem clear.

One problem could be that it is much harder to imagine a person carrying out all the tasks that this book involves than it was to imagine a person carrying out all the tasks in the Chinese Room. It seems that after several steps in the neuronal chain, so many neurons would be affected, changed, reaffected, and dependent on other neurons that it would be nearly impossible for a human to carry out the necessary moves in sequence. Perhaps, then, such a program should be implemented on a computer, or even better, some sort of connectionist machine. It could carry out all these tasks in parallel, and internalize the algorithms for "hearing" and "speaking."

In fact, implementing such an idea on a computer would seem to add the causal mechanisms that seem necessary for a real person. Some of these were discussed in the section on Haugeland and semantics. There is also the idea that one neuron affects other neurons unless actively stopped by outside forces, e.g., a stroke. With Einstein as a book, these mechanisms did not seem very strong, as a person could just get bored, or miss a calculation somewhere and mess up the entire program. Yet, once it was implemented on a computer, such mistakes would not be a problem, and since machines are inherently causal, this would transfer into Einstein as a connectionist machine.

This seems to be a real way in which brains are computational. That is, the neuron gets some inputs and "algorithmically" (rather than randomly) emits some outputs. The computational method argued against by Searle in the Chinese room was one in which all information that comes into the brain is treated computationally as a

whole. A picture was treated as a symbol and information to be processed, whereas now a picture is broken into pieces (light/dark contrasts, etc.) and processed in a much more simplistic sense. Also, the perfection of the computer handling the neuronal inputs and outputs seem to be more like natural law. That is, human neurons stick to their action potentials and reactions, unless they are changed by an equally law-like mechanism.

In the Chinese Room, we had to add on a lot of upper-level features ad hoc in an attempt to make the program more and more human. These included talking to itself, initiating conversations, making mistakes or slips of the tongue, and the like. With this new idea, these activities should appear naturally as a result of the implementation of all the neuronal functions. This idea takes better advantage of the perfection of the working of the computer without putting it at a disadvantage. Unfortunately, it is no simple task to find and program billions of microprocessors into individual neurons. As a theory, though, I think it offers hope and help to biologists and AI folk alike.

In fact, Searle almost accepts this idea of connectionism. He writes:

> Among their merits, at least some connectionist models show how a system might convert a meaningful input into a meaningful output without any rules, principles, inferences, or other sorts of meaningful phenomena in between.... [T]hey are not all obviously false or incoherent in the way that the traditional cognitivist models that violate the connection principle are (Searle 1992, 246-47).

Not exactly rousing support for connectionism, but if Searle is to maintain strong objections to traditional AI (such as scripts or knowledge data bases), yet account for the biological problems with semantics, it seems he may be forced into such a position.

Although it is doubtful than anyone can force Searle into any position he does not care to take.

Bibliography

Dennett, Daniel, "Reflections [on Hofstadter]," in *The Mind's I,* edited by Douglas Hofstadter and Daniel Dennett. New York: Bantam Books, 1981.

Haugeland, John, "How Can a Symbol 'Mean' Anything?", in *Meaning and Cognitive Structure,* edited by Zenon Plyshyn and William Demopoulos. Norwood, NJ: Ablex Publishing Corp., 1986.

Hofstadter, Douglas, "A Conversation with Einstein's Brain," in Hofstadter and Dennett, *op. cit.,* 1981.

----------, "Reflections [on Searle]," in Hofstadter and Dennett, *op. cit.,* 1981.

Putnam, Hilary. "Brains in a Vat," in *Introduction to Philosophy,* edited by John Perry and Michael Bratman. New York: Oxford Univ. Press, 1993.

----------, "Computational Psychology and Interpretation Theory," in Plyshyn and Demopoulos, *op. cit.,* 1986.

Sacks, Oliver, *An Anthropologist on Mars.* New York: Alfred A. Knopf, 1995.
Searle, John, "Minds, Brains, and Programs," in Hofstadter and Dennett, *op. cit.*, 1981.
----------, *Minds, Brains, and Science.* Cambridge, MA: Harvard Univ. Press, 1984.
----------, *The Rediscovery of the Mind.* Cambridge, MA: The MIT Press, 1992.