Aporia vol. 17 no. 2–2007

Intensional States and Computer Programs

Joseph Sousek

We may define a computer program as a list of instructions which determines some specific output response on reception of some specific input. The means by which this output is achieved is purely formal. That is, both the input and output are symbols (for example, ones and zeros) which have no intrinsic meaning, though they can be differentiated from other symbols. A simple instruction is conceived here as a conditional rule that can be applied unthinkingly.

This description of a computer program is to be distinguished from a computer itself, which can be described as "the instantiation of a computer program" (Searle, "Minds, Brains and Programs" 347). This distinction between a computer and a computer program is important to my argument about John Searle's proposition. Searle does not suggest that computers have intensional states: surely those things we know to have intensional states (ourselves) can be described generally in terms of various computer programs. Rather, Searle argues that nothing can have intensional states "solely by virtue of being a computer with the right sort of program" ("Minds, Brains and Programs" 347).

"Intensional states," "meaning," and "understanding" will be taken here to mean a conscious understanding of what given symbols represent. While the symbols, in Searle's terminology, are merely *syntactical*, the meaning, intension, or understanding of what a symbol represents is *semantic*. Where this meaning comes from and what determines each symbol's specific meaning is something I will return to later in this essay.

Searle's position can be stated as follows:

> 1) Computer programs are syntactic, dealing only with the manipulation of intrinsically meaningless symbols.

Joseph Sousek is an undergraduate at Bristol University, majoring in philosophy. This essay was written while on exchange to the National University of Singapore.

> 2) Human minds have mental content; they have semantics.
>
> 3) Anything that is purely syntactic is not sufficient for semantics.
>
> 4) Computer programs are not sufficient for semantics; they do not have mental content.

Premise one is true by definition. Premise two has been questioned by philosophers such as Quine in *Word and Object*, but Searle takes premise two as self-evident even in the face of other philosophers' arguments. In fact, he sees the conclusion that we do not mean things by our words as a *reductio ad absurdum* of Quine's whole project, and I will accept this conclusion provisionally ("Indeterminacy, Empiricism, and the First Person" 137). All that is now required in order to derive the conclusion that computer programs alone cannot give rise to intensional states is the third premise, and this is the thesis of Searle's famous thought experiment, the Chinese room. Searle asks us to imagine a situation where a system runs a computer program that simulates the understanding of a foreign language, and he tries to show that such a system cannot rightly be said to experience intensional states or to understand any meaning in its operation.

The situation is as follows: A monolingual English-speaking man is locked in a room containing a large selection of Chinese symbols and a pictorial rulebook. Now and then a symbol is posted into the room through a slot in the wall. For every symbol or combination of symbols coming into the room, the rulebook provides simple instructions for determining a symbol or combination of symbols to post out of the room through a second slot. These instructions are entirely formal, giving no hint of the translated meaning of any of the symbols. Unknown to the man, the people outside the room are asking coherent questions in Chinese script, and the rulebook constitutes a program that provides coherent and informative responses to these questions, ones "indistinguishable from those of a native Chinese speaker" (Searle, *Minds, Brains and Science* 32).

To a Chinese-speaking observer, then, whatever is inside the room seems to understand Chinese and to mean things by its responses. However, says Searle, all the man is doing is following simple rules for the manipulation of what are to him meaningless images: "there is no way you could learn any Chinese simply by manipulating these formal symbols" (*Minds, Brains and Science* 32). The man inside does not understand the inputs or outputs and certainly does not assign the real Chinese meaning to the output symbols. The conclusion Searle draws from this

scenario is that the systematic manipulation of syntax is not sufficient for semantics. Therefore, the instantiation of a computer program alone is not sufficient to constitute mental content (meaning, understanding, intension, etc.).

Searle expresses some bewilderment with both the resistance to his argument and the persistence of the proponents of this resistance ("Is the Brain's Mind a Computer Program?" 270). Those who resist Searle's argument give objections that aim at finding an interpretation or stipulation of the Chinese room scenario where the system (the instantiation of the computer program) really does have the mental states that Searle denies it has. Although Searle has provided responses to many of these objections, I think two of them have proven particularly difficult to dispel adequately. For reasons I will outline in this paper, these objections seem to cast a considerable degree of doubt on Searle's conclusion. After evaluating the threat posed to Searle's position, I will examine whether there is any conclusive way to counter this threat or whether Searle's thesis must be rejected or at least suspended.

The *systems reply* makes the claim that, while it is certainly true that the man in the Chinese room does not understand Chinese, the system as a whole does. The system as a whole includes everything in the room: the man himself, the rulebook, the Chinese symbols, the input and output slots, and so on. Only a system that includes all of these elements is truly analogous to an instantiation of a computer program and to a human brain. This system altogether can converse in Chinese perfectly meaningfully.

Searle's response to this reply is twofold. First, he asserts that it is simply implausible to think that "somehow the *conjunction* of that person and some bits of paper might understand Chinese" ("Minds, Brains and Programs" 337). In fact, Searle says he feels "somewhat embarrassed" to be taking it seriously. Searle's response is unsatisfactory because it comes from the assumption that a system that is supposed to be analogous to a conventional computer cannot have intensional states, and this is the very claim that he set out to prove. It begs the question.

Second, Searle gives a more reasoned argument, asking us to imagine another situation in which the man takes the time to memorize the entire (untranslated) rulebook and all the symbols, doing the whole process from memory. In such a case the man *is* the whole system, and he still has no idea what any of the symbols mean. He still does not understand Chinese. If people claim he does, they are clearly using a very different conception of understanding from the one we are concerned with. The conception we are interested in is the sort of understanding that the same man has of English. We would not imagine that the man converses in Chinese with the same sort of semantic understanding as he has when he converses in English. His understanding of Chinese is limited to the reception and response of

symbols to which he ascribes no meaning. So even in this case, the system cannot be said to have the relevant understanding.

However, there are problems with this response. First, it might be argued that the example is now implausible—not in the sense that the example is logically impossible, for it seems not to be, nor in the sense that the practical impossibility of trying out the experiment is a problem, for this probably applies equally to the original Chinese room scenario. The problem is that we have reached so "deep into counterfactual territory" (Lowe 217) that it is impossible to convincingly forecast the consequences of such a scenario. If a person were able to memorize a rulebook enabling him to answer any question in the Chinese language sufficiently well to feign understanding, "who is to say whether or not he would as a consequence be able to understand Chinese?" (Lowe 217). I would be inclined to side with Searle in saying that understanding could no more arise from this scenario than from the standard Chinese room scenario, but the certainty of the conclusion that is present in the original argument has diminished here.

Second, and more importantly, one of the strengths of the Chinese room as originally conceived is that, when looked at in isolation, the man himself clearly did not understand Chinese. Rather, he relied on following instructions physically external to himself. Whether the man could understand Chinese in the same sense he understands English was a criterion for whether the system had any intensional states. If, however, all aspects of the system are internalized in the man, or if we are required to look for semantics in the Chinese room as a whole system, then we can no longer look at a distinct element of the system which clearly does or does not understand and whose lack of understanding constitutes an absence of semantics. If we must examine the entire system of the Chinese room, or one person who embodies the entire system, then it will of course always *appear* to understand Chinese, and the only reason we have to deny this conclusion is the pre-experimental, though plausible, intuition that something which merely manipulates syntax does not grasp any semantics. Once again, Searle is begging the question, using the conclusion of his argument as a premise. The absence of a distinct criterion for the system's having intensional states forces him to resort at least in part to reliance on the very principle he set out to prove.

The second objection I will discuss is the brain-simulator reply. This reply states that if the system on which the computer program was run simulated the way a human brain works, then the system would have intensional states. That is, it would have intensional states if it "simulated the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives

answers to them" (Searle, "Minds, Brains and Programs" 341). Searle gives two counterexamples to this objection.

The first counterexample is of the man in the room manipulating an intricate complex of water pipes and valves that match the layout of a Chinese speaker's brain. The Chinese symbols correspond to English instructions in the rulebook for which valves to turn on or off in order to produce the output symbol. He then claims that the described system does not produce understanding any more than the standard Chinese room does: "As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain: its ability to produce intensional states" (Searle, "Minds, Brains and Programs" 341).

The second counterexample, produced against Paul Churchland's objections, invokes the case of a large gym full of monolingual English-speakers who play the role of synapses and nodes in a Chinese-speaking brain, resulting in matching output to input symbols as before. Again, Searle claims that the lack of any understanding of Chinese existing anywhere in the room shows that the brain-simulation has failed.

Searle's defence can be criticized because it seems to take the idea of brain-form simulation a little too literally. The assertion that water pipes in the shape of a brain could never be enough to constitute understanding fails to engage with the point at issue. Since Searle offers little explanation of what it is about the brain that '"produces intensional states," (though he does elsewhere concede that a non-biological system could be capable of intensional states so long as it had "the relevant causal capacities equivalent to those of brains" ["Is the Brain's Mind a Computer Program?" 269]) one is led to understand that, in Searle's opinion, a system that is sufficiently similar in the right respects, whatever these are, might be capable of understanding. The right respects might be to do with formal structure, the means by which information is passed, the speed at which it is transferred, and so on. If, for example, the system running the program comprised not water pipes but a hugely complicated network of wires carrying electronic impulses in the layout of a Chinese speaker's brain, it might be deemed sufficiently similar to a brain to produce understanding. As in the case where a man internalizes the rule book in the systems reply, who is to say whether the system would have intensional states?

Again, Searle's response is subject to the serious flaw that it loses its criterion of the presence of intensional states since here they would take place in the simulated brain which, whether it meant anything or not, would of course always appear to. Searle is forced once again to call on our intuition that syntax alone is not sufficient for semantics in order to conclude that the brain has no semantics.

These two objections cast significant doubt on the certainty of Searle's conclusion, and his responses to both of them have not been decisive, resorting to the preconception that syntax is insufficient for semantics. One's position on the issue seems to depend simply on whether one finds Searle's conclusion intuitively plausible (as I must admit I do). In this context it seems unsurprising that the issue has lingered in stalemate. If we are to show that Searle is right—that instantiations of computer programs alone do not have intensional states—then some effort should be made to produce more effective arguments.

Implicit throughout the foregoing discussion is the idea that for a computer to mean something, to understand, or to have intensional states, it is necessary that the symbols involved represent something, refer to something, or at least connote other things. Obviously, for us to mean something by a word or symbol, the word or symbol must represent or refer to something other than itself. This principle implies that whenever the idea that a computer has intensional states is postulated, it is because the computer appears to understand inputs and produce outputs that use syntactic symbols to represent or refer to things beyond the symbols themselves. Thus, something that inputs and outputs symbols that have no representation or reference would certainly not be said to mean anything by the symbols.

In the case I have been discussing, the instantiation of a computer program tends to provide evidence that it understands Chinese and can converse in Chinese fluently and meaningfully. This proficiency in Chinese seems genuine because a Chinese observer will see that when the system is asked a question in symbols that represent some idea, it will respond with an answer in symbols that represent a related idea. So when given symbols meaning "how old are you?" rather than responding with some unrelated symbols, it responds with "I am thirty-seven." Such coherent answers are the reason for claiming that the system understands Chinese. Its syntactic answers represent something semantically, something coherent and appropriate to the semantics of the question.

Suppose we were to design another room similar to the Chinese room and put a man in that room. Instead of containing a large selection of Chinese symbols and a Chinese pictorial rulebook, this room contains an equally extensive selection of different Rorschach-style inkblot images. It also contains a Rorschach-style pictorial rulebook for matching input images to output images, and from time to time an image or combination of images are posted in the room through a slot, prompting the appropriate output to be posted back out. I will call this room the "Rorschach room."

Suppose we run this computer program and find that all is in order. The man in the room quite competently finds the correct output for every input, and the result is a steady flow of inkblot images coming out that

have some relation to the ones going in. However, the relation exists solely because the output is entailed by the rulebook's selection of the input. What would outside observers, unaware of the nature of the system inside the room, make of this? It seems obvious they would have to conclude one of the following: 1) many images are going in and out of the room at random; 2) the images are going in and coming out according to a system (not a language) the people don't know; 3) the images are questions and answers in a language they do not know.

Because there is a consistent system in place, we can reject option one. And because we know the images do not constitute the script of any language (rather, they are non-representative symbols that we happen to manipulate consistently with a computer program), we can eliminate option three. Thus, we are left with option two, which, because it does not employ language, suggests that a computer program is not sufficient for intensional states. However, so far, even accepting option two is unproblematic for opponents of Searle. Someone who believes that the instantiation of a computer program is sufficient for intensional states to occur can simply say that the given scenario is not an appropriate computer program. An appropriate computer program would have to recreate the use of a language, but this one does not.

Having tried out this computer program, let us imagine that we take the time to create an entirely new language which we will call *Rorschach-speak*. When we come to designing the writing of this language we use the previously mentioned inkblot images, and we carefully ensure that every possible question of this language is represented by a symbol or sequence of symbols in our previously designed computer program. We also ensure that every output symbol or combination of output symbols is given an appropriate corresponding meaning in relation to the meaning of every input symbol or combination of input symbols. This process might sound implausible, but even if it is rather complicated I see no reason why it should be logically impossible. Furthermore, even if designing a complete language in this way were far-fetched, it is not strictly necessary for this thought experiment. Languages can undoubtedly have meaning even when they are not holistic. There could be, for example, a unique language, the scope of which is solely to talk about events that routinely take place on a construction site. Presumably, in this case there is much less room for inconsistencies to arise between the language and the rulebook.

We now teach an unsuspecting man this language until he has mastered it and invite him to observe the Rorschach room from the outside. We run the computer program, giving exactly the same input, receiving exactly the same output as before. The participant will recognize the input as questions and the output as appropriate responses. On exactly the same basis

as the Chinese room case, we might have reason to say that the computer program understands a language. However, in this case someone making such a claim is faced with a dilemma. If the system means the Rorschach-speak meanings by its outputs, either the system has always understood these meanings of the symbols—even before we invented them—or it understands them now solely in virtue of someone else's understanding them.

The implications of either of these options will be very difficult to accommodate. We clearly have no grounds to suppose that the system has always had understanding of the syntax we gave it and somehow held each symbol to have the same meaning as those we later defined ourselves. On the other hand, to say that the system understands Rorschach-speak now simply because there is another speaker present would compromise our conception of what meaning is. Surely I would understand and mean things by my words in English even if I were the last English speaker alive. Our Rorschach-speaker's understanding the symbols to represent something does not entail that the source of the symbols means something by them.

Let us add another element to this story. Suppose we invent another language. We will call this language *Rorschach-talk*. This language, despite also using Rorschach-style inkblots as script, sounds completely different from Rorschach-speak. Again, any input-output pair of symbols from the original program will match up to a question-answer pair of sentences with appropriate corresponding meanings, but each symbol will have a different meaning in Rorschach-talk from what it has in Rorschach-speak. Making this language might also sound implausible, but I think two languages could share the same syntax. And they could do this while keeping a large range of question-answer pairs coherent within each language but differing in meaning between the two languages. As previously mentioned, it is certainly possible if they are specialized, non-holistic languages, and this approach would be adequate for our purposes.

The next step is to allow a different person to master Rorschach-talk and to allow them to observe the Rorschach room from the outside. Running the exact inputs as before on the computer program, we will find that the new observer recognizes the inputs and outputs as appropriate questions and answers, and we will have grounds to make the claim that the system of the Rorschach room understands Rorschach-talk. Here Searle's opponents certainly have a problem: they now are faced with the prospect that the system understands both languages at the same time. It experiences intensional states about both the Rorschach-speak meaning and the Rorschach-talk meaning of the syntactic answer that it gives in response to any input.

We have as good a reason for saying that the Rorschach room understands one or both of the languages as we had for saying that the Chinese

room understands Chinese. Surely to say that the room means two unrelated things by every symbol it outputs is absurd. Any attempt to argue that the system understands only one of the languages will be entirely arbitrary. There is no way to choose between them. On the other hand, if we concede that the system can mean more than one unrelated thing at a time by a symbol, what is to stop it from having other meanings?

As seen in the conclusions we began to draw from the previous scenario, people who think the system has intensional states are committed to the system's either having always associated these intensional states with the symbol or to the intensional states arising from other people who understand the meanings of the symbols. In the latter case, people who think the system has intensional states would be committed to meaning's being dependent on having understanding listeners. They would be, therefore, committed to thinking that any meaning ascribed to any of the symbols—by any observer—is somehow also found in the system.

If I thought that one inkblot meant "what do you want to do today?" and another meant "I think I'll go to the mall," then on this line of reasoning the system would also ascribe those meanings to the inkblots. We can view this either as an *a priori* proof of telepathy or a *reductio ad absurdum*. However, if we take the view that the system has always understood both languages then the absurdity will be all the more dramatic, since it follows that for any other meaning we might ascribe to each symbol in the future, the system will always have meant these simultaneously. This means that at any moment in time the system must mean every possible meaning by each symbol it uses. It is in the nature of a symbol that it can stand for more or less anything, and as a logical consequence, the system must mean *everything by every symbol it uses*. One can hardly imagine something more different from our conception of what it is to have intensional states.

It is not clear how anyone could argue that the Chinese room is significantly different from this example. Certainly, there is only one lexicon of the meanings of Chinese symbols, but it is clearly wrong to suggest that a computer program that manipulates syntax has intensional states because of popular understanding of the symbols we feed it. If it has intensions, the Chinese room system could mean anything. There is no reason to believe that the system actually means the Chinese meanings of each symbol. Thus, systems that merely instantiate computer programs do not have semantics or intensional states, at least not in the way we understand humans to have semantics and intensional states. These systems only mean something by their syntax either by an external observer's ascribing meaning to it (which hardly constitutes meaning on the part of the system) or by meaning absolutely everything (which is both absurd and degrades the whole value of having meaning; it fails to pick out a representation or referent).

Importantly, this conclusion is immune to the objections Searle failed to refute conclusively. Where his justification for concluding that syntax alone is not sufficient for semantics was ultimately dependent on his prior assumption that semantics could not arise from the syntactic operations alone, my argument takes a different line. Regardless of whether we look at the man in the room, or the room as a whole, or a brain-simulator, as long as the systems simply manipulate syntax in accordance with given instructions, the systems will either have no meaning or only have meaning of the absurd kind that we encountered in the culmination of the Rorschach room argument.

A concern for proponents of my position is that one may have no justification for the belief that anyone other than oneself has intensional states. The obvious reason for thinking that a computer program can mean things is that, based on our understanding of the language it is manipulating, it appears to. The same can be said of human beings. There is a possibility of a sort of meaning solipsism here, and like all forms of solipsism, it is very difficult to escape once explicitly entertained. Unlike a mere syntax manipulator, humans display great versatility of interaction with their environment in such a way that they can indicate what they mean by symbols or words. There are all sorts of ways we could determine what a person means by a symbol that would fail to derive any response from a syntax manipulator.[1] Equipping an instantiation of a computer program with the means to behave in a humanlike manner (as in a further objection to Searle, the *robot reply*) would do nothing to convince us that it has intensional states if at the heart of things it is merely a Chinese room attached to very sophisticated inputs and outputs, but equally we have no way of proving that other humans are anything more than functioning setups of this kind.

I think this can be overcome. I know I mean things by my words, and I suspect other people do too. My reasons for thinking so are that I seem to be the same kind of entity as everyone around me, functioning in the same way. Furthermore, I know that other people are built of the same organic stuff. In my case it seems to give rise to, or is at least in constant conjunction with the appearance of consciousness, sensory data, and of course, meaning, and it would take a truly sound argument to convince me that mine is the only case of meaning. I have not come across such a positive argument. But meaning and consciousness (and all the things caught up in solipsistic considerations) are incredibly mysterious. We should not so lightly attribute them to something as simple as a computer program, whatever the size and efficiency of that program. Whatever it is that generates meaning, it is

[1]For example, broadly speaking, we would understand that "gavagai" had something to do with rabbits even if we failed to identify the precise meaning.

not the manipulation of symbols, and consequently, so long as we maintain a reasonable notion of what it is to mean something, we can be sure that anything which merely manipulates symbols means nothing by them.

Works Cited

Jacquette, Dale. *Philosophy of Mind*. Englewood Cliffs, NJ: Prentice Hall, 1994.

Lowe, E. J. *An Introduction to the Philosophy of Mind*. New York: Cambridge UP, 2000.

Moody, Todd C. *Philosophy and Artificial Intelligence*. Englewood Cliffs, NJ: Prentice Hall, 1993.

Searle, John. *Minds, Brains and Science*. New York: Penguin Books, 1984.

——. "Indeterminacy, Empiricism, and the First Person." *The Journal of Philosophy* 84.3 (1987): 123–46.

——. "Is the Brain's Mind a Computer Program?" *Readings in Language and Mind*. Ed. Heimir Geirsson and Michael Losonsky. Cambridge, MA: Blackwell Publishers, 1996. 264–72.

——. "Minds, Brains and Programs." *Philosophy of Mind: Contemporary Readings*. Ed. Timothy O'Connor and David Robb. New York: Routledge, 2003. 332–52.