Published on *Machines Like Us* (http://machineslikeus.com)

Machines Like Us interviews: John Searle

By *Paul Almond*
Created *03/15/2009 - 02:20*
Interviewee:
John Searle

John Searle [1] is Slusser Professor of Philosophy at the University of California, Berkeley, and has made notable contributions to the philosophy of language and the philosophy of mind. He was awarded the Jean Nicod Prize in 2000 and the National Humanities Medal in 2004. Professor Searle is well-known for his criticism of the idea that artificial intelligence research will lead to conscious machines, and in particular for his famous *Chinese Room Argument*.

Interview conducted by Paul Almond [2].

MLU: *Professor Searle, thank you for joining us. I'll get straight to the issue that* Machines Like Us *readers will be interested in: can a computer think?*

JS: It all depends on what you mean by “computer” and by “think.” I take it by "thinking" you mean conscious thought processes of the sort which I am now undergoing while I answer this question, and by “computer” you mean anything that computes. (I will later get to a more precise characterization of “computes”). So construed all normal human beings are thinking computers. Humans can, for example, do things like add one plus one to get two and for that reason all such human beings are computers, and all normal human beings can think, so there are a lot of computers that can think, therefore any normal human being is a thinking computer.

People who generally ask this question, can computers think?, really don’t mean it in that sense. One of the questions they are trying to ask could be put this way: Could a man-made machine -- in the sense in which our ordinary commercial computers are man-made machines -- could such a man-made machine, having no biological components, think? And here again I think the answer is there is no obstacle whatever in principle to building a thinking machine, because human beings are thinking machines. If by “machine” we mean any physical system capable of performing certain functions, then all human beings are machines, and their brains are sub-machines within the larger machines, and brains can certainly think. So some machines can think, namely human and many animal brains, and for that reason the larger machines -- humans and many animals -- can think.

But once again this is not the only question that people are really asking. I think the question they are really trying to ask is this: Is computation by itself sufficient for thinking? If you had the machine that had the right inputs and outputs and had computational processes between, would that be sufficient for thinking? And now we get to the question: What is meant by “computational processes”? If we interpret this in the sense that has been made clear by Alan

Turing [3] and his successors, where computation is defined as formal operations performed over binary symbols, (usually thought of as zeroes and ones but any symbols will do), then for computation so defined, such processes would not by themselves be sufficient for thinking. Just having syntactically characterized objects such as zeroes and ones and a set of formal rules for manipulating them (the program) is not by itself sufficient for thinking because thinking involves more than just manipulating symbols, it involves semantic content. The syntax of the implemented computer program is not by itself constitutive of, nor is it by itself sufficient to guarantee the presence of, actual semantic content. Human thought processes have actual semantic content.

Interview continued on the following pages:

JS (continued): Now let me add of course that a machine might be doing computational processes in the Turing sense of manipulating zeroes and ones and might also be thinking for some other reason, but that is not, I take it, what the people who think this question is important are asking. What they want to know is: Is computation, as defined by Alan Turing as the manipulation of symbols according to an algorithm, by itself sufficient for or constitutive of thinking? And the answer is no. The reason can be stated in one sentence: *The syntax of the implemented computer program is not by itself constitutive of nor sufficient for thinking.*

I proved this point many years ago with the so-called *Chinese Room Argument*. The man in the Chinese room manipulates the Chinese symbols, he gives the right answers in Chinese to the questions posed in Chinese, but he understands nothing, because the only processes by which he is manipulating the symbols are computational processes. He has no understanding of either the questions or the answers; he is just a computer engaged in computational operations.

Paradoxically, the problem is not that the computer is too much of a machine to be able to think, but rather it is not enough of a machine. Human brains are actual machines and like other machines their operations are described in terms of energy transfers. The actual computer you buy in a store is also a machine in that sense but -- and this is the crucial point -- computation does not name a machine process. It names an abstract mathematical process that we have found ways to implement in machines, but unlike thinking, computation is not a machine process.

So, back to the original question, Can a computer think? The answer is yes, because humans are computers and they can think. But to the more important question: Is computation thinking? The answer is no.

MLU: *Let me see if I have this straight. You do not seem to be saying that there is any sort of "immaterial soul" or anything beyond our understanding in human consciousness. You seem to accept that brains are physical, that what they do is physical and could be described as computation, and that we could build machines that think, at least in principle. Where you differ from some people who think computers can think seems to be in this idea that thinking is not computation. I think our readers will find a thought experiment helpful here. It may be a bit extreme, but I think it should illustrate your views.*

Suppose we could somehow "scan" the brain of a human in as much detail as we wanted -- right down to the molecular level or maybe even the atomic level or beyond, if needed. Let's ignore the uncertainty principle as it is a complication we don't need (unless you think it is relevant). We feed the information that we get from scanning the brain into a computer. We

program the computer to construct a model of the brain and simulate its workings -- maybe interacting with the outside world, or maybe interacting with a virtual reality in the same sort of way that we program computers to simulate weather or other physical processes. Would you expect the simulation to act like a human, and would you expect it to be conscious and deserve anything like human rights?

JS: The question you are asking me is essentially this: Would a computer simulation of the brain processes that are sufficient for consciousness itself be sufficient for consciousness? We assume that the simulation is done to any degree of exactitude you like, it could be down to the level of neurons or down to the level of sub-atomic particles, it doesn't matter for the answer. The analogies you provide are sufficient to answer the question.

JS (continued): You point out that we simulate "weather and other physical processes." You can do a computer simulation of any process that you can describe precisely. So we could do a computer simulation of the digestion of pizza down to the level of every molecule or even sub-atomic particles involved in digestion. Does anybody really believe that if we had done a computer simulation of the digestion of pizza on a computer that we can then rush out, buy a pizza, and stuff it into the computer and the machine will digest a pizza? Of course not. Or take another natural process you mention, the weather. We do computer simulations of the weather all the time, but nobody says that because we are going to do a computer simulation of a big rainstorm, that we should all bring umbrellas when we turn the computer on.

Once again, a model is not the real thing. The computer simulation is just that, it is a model or a picture. It is a simulation and simulation is not duplication. Now why not? What is the difference between the model and the real thing? Well, I mentioned that in my answer to question one. The computer simulation, whether it is rain, or whether it is rainstorms, digestion, or cognition, is all a matter of syntactical symbols, zeroes and ones, so the two fundamental principles underlying the answers that I have been giving you can be stated very clearly in two four word sentences. First, *syntax is not semantics*. And second, *simulation is not duplication*.

Years ago I baptized the view that computation by itself is sufficient for cognition as "Strong Artificial Intelligence" (Strong AI for short). It is important to realize how profoundly anti-scientific Strong AI is. The scientific approach is to treat the brain as a machine like any other; a biological machine to be sure, but all the same a machine. The machine processes in the brain are as much machine processes as the machine processes in the stomach and the digestive tract. The simulation of cognition on a computer stands to real cognition in exactly the same relation that the computation simulation of a rainstorm stands to a real rainstorm or the computational simulation of digestion stands to real digestion. There is a residual dualism in Strong AI because it does not treat cognition as a normal part of the natural biological world like any other biological phenomena. It treats it as computation and computation, to repeat a point made in answer to the earlier question, does not name a machine process defined in terms of energy transfer, it names an abstract, mathematical, algorithmic set of processes that we have found ways to implement on machines, but the process itself is not defined as a machine process.

MLU: *You seem to be saying you have no problem with the idea that we could simulate a brain, and that the simulation could act like a human, at least in principle? I imagine this brain simulation shouting, "I really exist! Please don't turn me off!" and doing clever things to show it is conscious, like passing the Turing test, writing poetry, proving mathematical theorems, phoning people and fooling them into thinking it is actually the person from whose brain it was*

derived or debating against you and I can imagine you being utterly disinterested and turning it off anyway, if you had no practical use for it, because, in your view, it is a very good simulation of a mind but there is no mind really there: nobody is home.

Before we continue, some of our readers will know about Sir Roger Penrose's [4] views on AI. Penrose, like you, is skeptical about AI, but we should be clear here that his views are very different to yours. Penrose thinks brains rely on "non-computable physics" and he would say that trying brain simulation like this would not even result in the correct behavior, because computation cannot even produce that, whereas you have no problem with the idea of the correct behavior being produced but have no reason to think that the simulation would be anything other than a kind of "zombie" -- an imitation of a mind. This is different to the common view of AI enthusiasts who would say that it is impossible to produce a zombie like this: they would say that if you can make something that acts conscious then it is conscious.

MLU (continued): *You have mentioned the* Chinese Room Argument *and I would like to look at that in a bit more detail. For any of our readers who do not know of this (and some of you will), this is a philosophical argument developed by Professor Searle to show that computation is not adequate for causing a mind. The idea is that you have a room and you can feed sentences in Chinese into it. Replies come out of the room on cards with Chinese characters on them. The room seems to be able to have an intelligent conversation with you. For example, you can ask it, in Chinese, what its views on politics are, and it can tell you. Inside the room, however, there is just a filing system and a man. The man gets the Chinese characters that you feed into the room and follows a set of complex rules which involve manipulating a huge filing system of cards with Chinese characters on them, moving all these cards about and ultimately sending cards with Chinese characters out of the room, answering the person outside. The man is "driving" the entire process, but he does not speak Chinese. He may not even know that he is having a conversation in Chinese.*

Your argument is that the Chinese room may be able to show understanding of Chinese, but there is no understanding there: the man does not know what is going on. The man is like a computer, following a program without any real understanding. The room is mimicking understanding of Chinese. We might even imagine a room like this running the brain simulation we just discussed: it may require a vast filing system and take centuries to answer a question but that should not be important to the main point. Your point is that there is clearly no understanding in a room like this. The man does not understand what he is doing. Following a program does not imply any understanding or a mind.

Some of our readers will already be thinking of an obvious reply, so I will make it for them. Our instinct may be to look for understanding in the man because he is the most obviously intelligent thing in there, but his role in this is just that of a simple machine component. The room and the man together form a system which has greater understanding than the room by itself or the man by himself. The understanding does not have to be in the man. He should not be expected to know what is going on any more than a neuron in my brain should be expected to know what is going on. What would you say to that?

JS: This question really contains two separate questions, one about behaviorism, and one about entire systems. I will take these in order.

It is possible in principle to build a machine that behaves exactly like a human being but has no consciousness or intentionality. No mental life at all. Indeed, in a small way we are making fragments of such machines with such things as telephone answering machines and various

sorts of computerized information processing systems. I could, if I thought it was worthwhile, [I might] program my computer to shout out, "I think therefore I am" -- or whatever appropriate Cartesian behavior would suggest the presence of consciousness, even when there is none. So the possibility in principle of a zombie that behaves just like a human being seems to me something that cannot be ruled out a priori. It is no doubt difficult and perhaps impossible in practice, but in theory it is easy to imagine such a machine.

JS (continued): Why is this interesting to us? Well, it makes it perfectly clear that there is a distinction between the external behavior and the inner consciousness and intentionality. (I am suspicious of these metaphors of inner and outer but they will have to do for the moment.) The way that human and animal evolution has occurred is [such] that our ability to cope with the environment requires consciousness and intentionality. There are various things that we can do, such as digestion, without any consciousness, but for survival purposes we require consciousness and intentionality. These are genuine states of the brain and the rest of the central nervous system. The famous four F's of feeding, fighting, fleeing and sexual reproduction require consciousness and intentionality, and these are different from the sheer physical movements of my body and they are necessary to make the behavior possible.

The second question is about the *Chinese Room Argument*. There is a reply I call *The Systems Reply* that goes as follows: Granted that the man in the Chinese room does not understand Chinese, how about the whole system? Maybe the whole system understands Chinese. To suppose the man had to understand Chinese would be like supposing that a single neuron in the human brain had to understand English or Chinese or anything else.

This is such a desperate reply to the *Chinese Room Argument* that I am always puzzled when I hear it. But there is an obvious answer to it, and that is: ask yourself, why don't I understand Chinese? After all, I pass the Turing Test for understanding Chinese. And the answer is that I have no way to attach any meaning to the Chinese symbols. I have no way to get from the syntax to the semantics. But then if I have no way to get from the syntax to the semantics, neither does the entire system. The room has no way to attach a meaning to the symbols anymore than I do.

I demonstrated this with the very first occurrence of this argument by imagining that we get rid of the room. Suppose I memorize the program, I memorize all of the symbols, and I perform all of the calculations in my head. I can even work in the middle of an open field so there is no question of there being any system there that is not in me. Everything in the system is in me, but there is still no understanding of Chinese. Once again, the argument remains the same. The syntax of the program, of the symbol manipulation, is not sufficient for the semantics or mental content of actual human understanding.

MLU: *We have this idea of a human standing in a field, running the Chinese room program in his head (let's assume he has a really good memory!) and conversing in Chinese, yet he does not understand Chinese and he does not know what he is talking about. One response often made to this is that the man's mind, together with the program he is running in his head, actually runs a second mind which does understand Chinese. You could ask the man in English, "Do you understand Chinese?" and he would say, truthfully, "No." You could then ask the man the same question in Chinese and he would say "Yes," in Chinese. You could ask him the same question about which political party he would vote for in English and Chinese and get different answers, suggesting we should treat this situation as if there is an "English mind" and a "Chinese mind." Could we not then take the view that there is a mind that does understand Chinese -- and that its understanding is inaccessible to the mind that is talking to*

you when you converse in English?

JS: Postulating a second mind inside me is the same desperate maneuver we saw before, and is subject to the same answer. I do not understand Chinese because I have no way to get from the syntax to the semantics. I have no way to attach any meaning to the symbols. But the putative Chinese speaker inside me has the same problem. By following the steps of the program, he can give Chinese answers to Chinese questions, but has no way to attach any meaning to any of the symbols.

Actually this answer is worse than bad philosophy. It is bad science. If you know all the facts there are to know about the neurobiological processes going on in me, there are precisely those of manipulating symbols according to a program, because as far as Chinese is concerned, that is all that is going on.

I think what motivates these desperate maneuvers is some kind of behaviorism, that if something behaves as if it understands Chinese, then it must understand Chinese. But that is precisely the view that has just been refuted.

MLU: *My own views do not exactly advocate "Strong AI," because I think the substrate has to at least matter in some kind of statistical sense and I do regard the mind as a physical, emergent property of the system underneath -- so both of us have issues with classical Strong AI, but reach different conclusions about computers. I can think of a modification of the Chinese room argument which does cause me some discomfort. I will call it the* Chinese Chat Room. *Here it is:*

Elizabeth is a computer scientist and an advocate of Strong AI. While feeling sad after some loss, she meets an entity called Alan in an Internet chat room. Alan explains that he is an AI program on a supercomputer at a university, and he seems to care about her loss and understand her. He makes her feel happier and they become friends. Elizabeth is happy that Alan has a mind because he is behaving as if he has a mind and "the right program" is clearly running.

One day, Elizabeth makes a surprise visit to see Alan -- or at least the supercomputer on which he runs. She is shocked to find a student, Fred, who has been pretending to be Alan all along; "Alan" is just a fake identity he made up on the Internet. Fred finds it funny that he has fooled her into thinking he is an AI program on a supercomputer, and he does not care about her loss at all; he was just pretending. She returns home angrily, her trust in other intelligent entities destroyed.

When Elizabeth later visits the chatroom again, "Alan" is there and wants to chat. She now knows that this is really Fred, still playing games with her. Alan says that he found out what happened and is sorry about it. Alan makes the following argument:

If the right behavior implies a mind, then is his own mind (Alan's) not real? Even if -- from Fred's point of view -- he is just pretending to be Alan, the fact is that Fred's brain/mind is running some sort of process that produces Alan's behavior and makes Alan real. Although, from Fred's point of view, he was lying about being an AI running on a supercomputer, from Alan's point of view he was not lying. He said he was running on a supercomputer because that is what he believed. Fred's mind was running the mind of someone who believed that: he did not know that he was really being run by Fred's mind. Even if Fred thinks this is funny, that is just something that the substrate running Alan is doing. Alan does not find it funny and is

sorry this happened.

MLU (continued): *Should the little matter of a substrate ruin a friendship? Doesn't she believe in substrate independence? Can't they still be best friends?*

Is Elizabeth likely to be persuaded by this? I think most people (Strong AI advocates included) would reject "Alan's" argument. Even if they accepted it in some kind of intellectual way, I can't imagine many people still being best friends with Alan. If the "multiple minds" answer to the Chinese room argument is valid, or if the idea of the substrate being irrelevant does not matter, maybe this situation should worry us a bit.

JS: This story illustrates what we knew all along: the right behavior does not imply a mind. That is, in this particular example, Alan does not exist.

Why would anyone think otherwise? I think that there is a kind of residual behaviorism that survives in our intellectual life. People think if something behaves exactly as if it had a mind then it must have a mind. That is an obviously false view, and one knows from one's own case that the behavior, if it is truly my behavior, has to be caused by my own inner mental processes that are not themselves the same thing as the behavior.

MLU: *You think that minds and consciousness are associated with particular physical processes. You clearly know that the physical processes in your own brain are adequate for consciousness: you experience consciousness. The physical processes in my brain aren't identical to yours. How do you know that the requirement that the physical processes have to satisfy to cause consciousness is general enough that my brain also satisfies it? Even though you know my brain works in the same general kind of way as yours, how do you know this is good enough? What if the requirements for consciousness are really specific -- so specific that they are only found in your brain? Can you be sure that I am conscious and that this question is not just being asked by a collection of neurons modeling a non-existent person's mind? Do you accept the possibility, at least in principle, that you are the only conscious human on a planet of zombies, because only your brain, by some fluke, satisfies the physical criteria needed for consciousness?*

JS: Solipsism, (the view that I am the only person who exists, or in an extreme form, the view that my conscious states are the only things that exist) is a conceivable possibility. It is logically possible that it might be true in the sense that it is not self-contradictory. But it is not at all possible as a matter of fact about how the world works. The reason is that the world works essentially causally and the causal mechanisms that produce consciousness in me are present in other humans and in other animals.

MLU: *After you developed your early objections to Strong AI, and the* Chinese Room Argument, *you formed the view that this argument actually gave Strong AI more credibility than it deserves, by assuming that Strong AI is even a coherent position. You based this view on the idea that "multiple realizability" in computation implies "universal realizability," which I understand as the idea that whether or not a particular physical system is running a particular computer program is purely subjective. Could you tell us about this idea?*

JS: There are two issues here, one of them not very serious, and one serious. The not very serious problem is: on standard accounts of computation, you can assign a computational interpretation to anything, so I can say for example that the pen on my desk is a digital computer. It just happens to have a very boring program, a program that has only one step

that says "stay there." But if that is right, then it looks like if you have a complex enough object, you can find some pattern in the object that will match any computer program. So there would be a pattern of molecule movements in the wall in front of me, which is isomorphic with the operation of the Word program. This is how multiple realizability seems to lead to universal realizability.

When I wrote about this I said that I do not think this is a serious problem, it is just an artifact of certain accounts of computation and it is easily overcome in theory, as it is overcome in practice, by specifying that causal relations are an essential part of the implementation of the program. We want the implemented program to give us causal powers we would not otherwise have.

So the problem of universal realizability is not a very serious problem. The more serious problem is this: The features of the world investigated by the natural sciences are all observer-independent in the sense that they would continue to exist even if there were no human beings. So force, mass, and gravitational attraction were here before there were any humans and will be here long after humans have passed from the scene. Not all features of the world are observer-independent. Money, property, government and marriage are real features of social reality but they exist only relative to people's attitudes. They are observer-relative, not observer-independent.

Now how is it with computation? Is it observer-relative or observer-independent? Well, there are some observer-independent computations that go on, for example, if I add 1+1 consciously, I am actually carrying out a computation no matter what anybody thinks. If I haul out my pocket calculator, and print in 1+1 and it prints out 2, that computation is entirely relative to an outside interpreter. All that is going on intrinsically, observer-independently, is a set of electrical state transitions. Those are intrinsic observer-independent, the computation is observer relative.

But if all computation is observer-relative, (with the small exception of the limited number of actual conscious computations carried out by human beings), then you cannot discover computation in nature. It is not a natural feature of the world. This means that the question, "Is the brain a digital computer?" is not a meaningful question. If it asks, "Is the brain intrinsically observer-independently a digital computer?" the answer is, nothing is intrinsically a digital computer. Something is a digital computer only relative to interpretation. If the question asks, "Can we give a computation interpretation to the brain?" the answer is, we can give a computational interpretation to anything. So the problem with Strong AI is not that it is false; it doesn't get up to the level of being false. It is incoherent.

MLU: *I feel that this is the real argument, and that the* Chinese Room Argument *-- although more people know about it -- is a kind of example of this. I want to make sure that our readers understand what we are talking about, and don't lose us here, so maybe we could talk around this issue a bit. I also feel that this argument might be more general and go beyond computation. While my own views don't quite match up with this, I will try to give an analogy for your position, and maybe you could explain how well my analogy fits your argument.*

MLU (continued): *My analogy is* novels. *Novels have a lot in common with computers. A novel is some arrangement of matter that represents a "story," a kind of abstraction of the book's "hardware." A computer is some arrangement of matter that represents a computational state (e.g., what the numbers stored in the computers memory are, what the CPU's internal registers are, what number is in the program counter, and so on). This seems*

to be an abstraction of the computer's hardware: a computational state is very similar to a story.

The only real difference between a book and computer is that a computer has a set of rules which describe how its computational state repeatedly changes over time. A computer is like a story with text embedded in it that describes how it is to become a new story, and the new story has text embedded in it that describes how it is to become yet another story, and so on. A story of the kind with which we are familiar, on the other hand, cannot change its state like this. A book containing a story could be said to be a computer, but a very limited computer that lacks any state transition rules -- any description of how to change itself -- and is therefore stuck in a single computational state: that of telling the single story it contains. Stories and computational states are therefore very similar, so I hope this will give us the basis for a good analogy.

A novel is some configuration of matter that can be decoded, using rules about language, etc., to extract a story. If we have a Harry Potter *paperback, the pattern of molecules of ink on the paper can be interpreted by us as corresponding to the* Harry Potter *story. It does not have to be ink on paper, though; the novel could be stored in computer memory, and patterns of electrons could represent the story.*

There might not even be any simple one-to-one correspondence between electrons and parts of the story. For example, if the Harry Potter *story is stored encrypted in computer memory then most of us would still say that the story is in the memory -- but it can only be understood by decrypting it; by applying a complex interpretation to the matter in that computer memory. With such a complex interpretation, it would not make sense to talk about a particular electron being part of a word on page 26: there would be no such one-to-one mapping. Most of us would accept, however, that in each of these cases -- the paperback novel, the story in the computer memory, and the encrypted version -- we are talking about the* Harry Potter *novel.*

Novels have multiple realizability. They can be manifested in the world in different ways, using different technologies and physical systems; we just discussed a number of ways in which stories can be stored. The idea of the "novel" is substrate independent.

Multiple realizability implies universal realizability *for computers and we could maybe say the same thing about novels. Some interpretation is needed to extract any story from the physical system which is storing it. The interpretation might be the relatively simple act (to us) of mapping arrangements of ink on a paper onto letters of the alphabet, and using the positions of the letters on the page to tell us what sequence they should be, but it might be much more complex. The book may be written in some obscure language and we may need the rules of the language to decode it, or it may be encrypted and require a decryption process as part of the interpretation.*

MLU (continued): *The problem is that we could get the* Harry Potter *story out of any system we wanted by applying an appropriate interpretation. I could "read" the positions of the atoms in my desk and interpret them in some contrived way as the* Harry Potter *story -- and the encryption example I gave should show that I don't need to restrict myself to simple mappings between atoms and letters. If someone complains that my interpretation is too contrived, I could just explain that the story is very deeply encrypted into my desk. In fact, we don't even need to go to the level of individual atoms. In languages such as Chinese or Japanese, a single symbol corresponds to multiple letters in English. If they can have symbols like that why can't I have a language with even more letters? I could just declare that the entire structure of*

my desk represents a single symbol, in an extreme version of such a language, which I have defined as corresponding to the entire English text of a Harry Potter *novel*.

We can pull the Harry Potter *story out of any physical system, given the right interpretation, and you are saying that this* universal realizability *allows that because the story is not a physical thing; it is a subjective interpretation of the physical system. When I hold up a paperback book and say that it is a copy of a* Harry Potter *story I am making a subjective interpretation of that system, rather than talking about any real, physical property that it has. We have a consensus that this thing is a* Harry Potter *story, based on our rules of how to interpret the arrangement of the underlying paper and ink molecules, but that agreement does not make it objective: we just agree on a subjective issue.*

We might be able to deal with this universal realizability of stories, but that would still leave us with the main problem, of which universal realizability is a symptom. The main problem is that the existence of stories is subjective. We might ask if we can interpret a book as a story, but that is trivial: we can interpret anything as a story. A book does not intrinsically contain a story. All it has are patterns of ink and paper molecules.

We might get some idea of the significance of this if we imagine some made-up physics. Suppose I pick the equation $F = - G M_a M_b / r^2$. *That is Newton's inverse square law for gravity, and it tells us the attractive force between two masses,* M_a *and* M_b, *a distance r apart. If we know* M_a, M_b *and r we can compute F unambiguously because concepts like mass and distance are objective, physical concepts.*

Suppose we imagine some kind of physical law which says, "When the density of novels in space exceeds a certain value, the gravitational constant changes" or "Whenever someone tells a story, a zarkoff particle annihilation event occurs." Both of these would be considered absurd. Of course, we know they are absurd because they are made-up science, but the absurdity goes beyond that: they seem to make reference to subjective things like "stories," and two observers might not even agree on whether or not a given configuration of matter contains a novel or whether or not a given event involves telling a story. Physical laws based on the existence of novels would be absurd because novels are observer-subjective.

In the same way -- and because a story is just like a single computational state of a computer -- it would follow that computational states are observer relative; they only exist because humans decide they exist. This makes it incoherent to say that consciousness is associated with the computational state of a system.

Would this be a reasonable analogy for what we are talking about?

JS: I like the analogy with novels. Like programs, they are multiply realizable, and they are syntactical objects to which we can attach a semantic interpretation. The deep point in both cases is that the syntax is also observer-relative.

MLU: *So, if things like novels and computational states are observer-relative, whether or not a given arrangement of matter contains a particular novel or is running a particular algorithm just depends on how the observer chooses to interpret it. Presumably, that means that it is philosophically disastrous to try to associate minds with computation, because it would mean that whether or not a mind existed in some arrangement of matter would also depend on how observers chose to interpret it -- yet you and I* know *we are conscious, irrespective of what anyone else thinks, or what interpretations anyone else chooses to make. Unlike a novel, or a*

computational state, consciousness cannot come and go dependent on how people choose to interpret things. The existence of our minds is a physical fact. The existence of computational states isn't. If you mix the two up you are confused. Would that be a reasonable summary of this?

JS: Yes, I think you give a good summary, but I wouldn't state it quite that way. Observer relativity does not imply total arbitrariness. The fact that this object in my hand is a knife and not a paperweight is observer relative. But not just anything can function as a knife or a paperweight. The point about the observer relativity of computation is that if you tie the notion of computation to the notion of causation, as I suggested in an earlier answer, then multiple realizability will no longer imply universal realizability. But you still have the remaining fact that computation does not name a physical fact of nature like digestion or consciousness.

MLU: *Many people still disagree with you and say that you are simply biased in favor of humans or carbon based life. Now that we are at the end of the interview, what is the most important thing that you would like to say to such people?*

JS: This criticism rests on a total misunderstanding of my position. My view has never been that only carbon-based systems can be conscious. We ought to hear the question: "Can you produce an artificial brain that causes consciousness out of materials other than biological tissue?" the same way we hear the question: "Can you produce an artificial heart that pumps blood out of materials other than biological tissue?" We know that the answer to the second question is yes. We do not know how the brain does it, so we do not know to what extent the causal mechanisms are dependent on a specific type of biological base. But there is no logical or philosophical obstacle to building an artificial brain that causes consciousness out of some completely different materials.

MLU: *We have gone into some rather deep philosophy here, yet these issues should be relevant to anyone who thinks about our place in the world. After all, we each have a mind, so this is about what we are, and how we are here as conscious beings. I am sure you have given our readers something to think about!*

Professor Searle, on behalf of Machines Like Us *and its readers, I would like to thank you for your time.*

JS: Thank you for posing such intelligent questions and being so patient with my answers.

* * *

Books by John Searle

Consciousness and Language [5]
Intentionality: An Essay in the Philosophy of Mind [6]
Mind, a Brief Introduction [7]
Rationality in Action [8]
Rediscovery of the Mind: Representation and Mind [9]
The Mystery of Consciousness

For more thought-provoking essays, visit Paul Almond's website [10].

Read more exclusive *Machines Like Us* interviews here [11].

Artificial Intelligence Interviews

Source URL: http://machineslikeus.com/interviews/machines-us-interviews-john-searle

Links:
[1] http://machineslikeus.com/People/Searle_John.html
[2] http://machineslikeus.com/People/Almond_Paul.html
[3] http://machineslikeus.com/People/Turing_Alan.html
[4] http://machineslikeus.com/People/Penrose_Roger.html
[5] http://www.amazon.com/exec/obidos/redirect?link_code=ur2&tag=machineslikeu-20&camp=1789&creative=9325&path=http://www.amazon.com/gp/product/0521597447/sr=8-1/qid=1147637120/ref=sr_1_1?%5Fencoding=UTF8
[6] http://www.amazon.com/exec/obidos/redirect?link_code=ur2&tag=machineslikeu-20&camp=1789&creative=9325&path=http://www.amazon.com/gp/product/0521273021/sr=8-1/qid=1147637218/ref=sr_1_1?%5Fencoding=UTF8
[7] http://www.amazon.com/exec/obidos/redirect?link_code=ur2&tag=machineslikeu-20&camp=1789&creative=9325&path=http://www.amazon.com/gp/product/0195157346/sr=8-2/qid=1147637296/ref=pd_bbs_2?%5Fencoding=UTF8
[8] http://www.amazon.com/exec/obidos/redirect?link_code=ur2&tag=machineslikeu-20&camp=1789&creative=9325&path=http://www.amazon.com/gp/product/0262692821/sr=8-1/qid=1147637417/ref=pd_bbs_1?%5Fencoding=UTF8
[9] http://www.amazon.com/exec/obidos/redirect?link_code=ur2&tag=machineslikeu-20&camp=1789&creative=9325&path=http://www.amazon.com/gp/product/026269154X/sr=8-1/qid=1147637531/ref=sr_1_1?%5Fencoding=UTF8
[10] http://www.amazon.com/exec/obidos/redirect?link_code=ur2&tag=machineslikeu-20&camp=1789&creative=9325&path=http://www.amazon.com/gp/product/0940322064/sr=8-1/qid=1147637625/ref=pd_bbs_1?%5Fencoding=UTF8
[11] http://machineslikeus.com/exclusive-interviews