thereafter, we approach the man himself (that is, we ask him to stop playing with the pieces of paper and talk to us directly) and ask him if he happens to know Chinese. He will of course deny such knowledge.

Searle's mistake of identifying the experiences of one system with those of its implementing system is one philosophers often make when referring to AI systems. For example, Searle says that the English subsystem knows that "hamburgers" refer to hamburgers, but that the Chinese subsystem knows only about formal symbols. But it is really the homunculus who is conscious of symbol manipulation, and has no idea what higher level task he is engaged in. The parasitic system is involved in this higher level task, and has no knowledge at all that he is implemented via symbol manipulation, anymore than we are aware of how our own cognitive processes are implemented.

What's unusual about this situation is not that one system is embedded in a weak one, but that the implementing system is so much more powerful than it need be. That is, the homunculus is a full-fledged understander, operating at a small percentage of its capacity to push around some symbols. If we replace the man by a device that is capable of performing only these opertaions, the temptation to view the systems as identical greatly diminishes.

It is important to point out, contrary to Searle's claim, that the systems position itself does not constitute a strong AI claim. It simply shows that *if* it is possible that a system other than a person functioning in the standard manner can understand, then the man-in-the-room argument is not at all problematic. If we deny this possibility to begin with, then the delicate man-in-the-room argument is unnecessary – a computer program is something other than a person functioning normally, and by assumption would not be capable of understanding.

Searle also puts forth an argument about simulation in general. He states that since a simulation of a storm won't leave us wet, why should we assume that a simulation of understanding should understand? Well, the reason is that while simulations don't necessarily preserve all the properties of what they simulate, they don't necessarliy violate particular properties either. I could simulate a storm in the lab by spraying water through a hose. If I'm interested in studying particular properties, I don't have to abandon simulations; I merely have to be careful about which properties the simulation I construct is likely to preserve.

So it all boils down to the question, what sort of thing is understanding? If it is an inherently physical thing, like fire or rain or digestion, then preserving the logical properties of understanding will in fact not preserve the essential nature of the phenomenon, and a computer simulation will not understand. If, on the other hand, understanding is essentially a logical or symbolic type of activity, then preserving its logical properties would be sufficient to have understanding, and a computer simulation will literally understand.

Searle's claim is that the term "understanding" refers to a physical phenomenon, much in the same way that the term "photosynthesis" does. His argument here is strictly an appeal to our intuitions about the meaning of this term. My own intuitions simply do not involve the causal properties of biological organisms (although they do involve their logical and behavioral properties). It seems to me that this must be true for most people, as most people could be fooled into thinking that a computer simulation really understands, but a simulation of photosynthesis would not fool anyone into thinking it had actually created sugar from water and carbon dioxide.

A major theme in Searle's paper is that intentionality is really at the bottom of the problem. Computers fail to meet the critieria of true understanding because they just don't have intentional states, with all that entails. This, according to Searle, is in fact what boggles one's intuitions in the man-in-the-room example.

However, it seems to me that Searle's argument has nothing to do with intentionality at all. What causes difficultly in attributing intentional states to machines is the fact that most of these states have a *subjective* nature as well. If this is the case, then Searle's man-in-the-room example could be used to simulate a person having some nonintentional but subjective state, and still have its desired effect. This is precisely what happens. For example, suppose we simulated someone undergoing undirected anxiety. It's hard to believe that anything – the man doing the simulation or the system he implements – is actually experiencing undirected anxiety, even though this is not an intentional state.

Futhermore, the experience of discomfort seems proportional to subjectivity, but independent of intentionality. It doesn't bother my intuitions much to hear that a computer can understand or know something; that it is believing something is a little harder to swallow, and that it has love, hate, rage, pain, and anxiety are much worse. Notice that the subjectivity seems to increase in each case, but the intentionality remains the same. The point is that Searle's argument has nothing to do with intentionality per se, and sheds no light on the nature of intentional states or on the kinds of mechanisms capable of having them.

I'd like to sum up by saying one last word on Searle's man-in-the-room experiment, as this forms the basis for most of his subsequent arguments. Woody Allen in *Without Feathers* describes a mythical beast called the Great Roe. The Great Roe has the head of a lion, and the body of a lion – but not the same lion. Searle's *Gedankenexperiment* is really a Great Roe – the head of an understander and the body of an understander, but not the same understander. Herein lies the difficulty.

Author's Response

by **John Searle**

Department of Philosophy, University of California, Berkeley, Calif. 94720

Intrinsic intentionality

I am pleased at the amount of interest my target article has aroused and grateful that such a high percentage of the commentaries are thoughtful and forcefully argued. In this response I am going to attempt to answer every major criticism directed at my argument. To do that, however, I need to make fully explicit some of the points that were implicit in the target article, as these points involve recurring themes in the commentaries.

Strong AI. One of the virtues of the commentaries is that they make clear the extreme character of the strong AI thesis. The thesis implies that of all known types of specifically biological processes, from mitosis and meiosis to photosynthesis, digestion, lactation, and the secretion of auxin, one and only one type is completely independent of the biochemistry of its origins, and that one is cognition. The reason it is independent is that cognition consists entirely of computational processes, and since those processes are purely formal, any substance whatever that is capable of instantiating the formalism is capable of cognition. Brains just happen to be one of the indefinite number of different types of computers capable of cognition, but computers made of water pipes, toilet paper and stones, electric wires – anything solid and enduring enough to carry the right program – will necessarily have thoughts, feelings, and the rest of the forms of intentionality, because that is all that intentionality consists in: instantiating the right programs. The point of strong AI is not that if we built a computer big enough or complex enought to carry the actual programs that brains presumably instantiate we would get intentionality as a byproduct (contra **Dennett**), but rather that there isn't anything to intentionality other than instantiating the right program.

Now I find the thesis of strong AI incredible in every sense of the word. But it is not enough to find a thesis incredible, one has to have an argument, and I offer an argument that is very simple: instantiating a program could not be constitutive of intentionality, because it would be possible for an agent to instantiate the program and still not have the right kind of

intentionality. That is the point of the Chinese room example. Much of what follows will concern the force of that argument.

Intuitions. Several commentators (**Block, Dennett, Pylyshyn, Marshall**) claim that the argument is just based on intuitions of mine, and that such intuitions, things we feel ourselves inclined to say, could never prove the sort of thing I am trying to prove (Block), or that equally valid contrary intuitions can be generated (Dennett), and that the history of human knowledge is full of the refutation of such intuitions as that the earth is flat or that the table is solid, so intuitions here are of no force.

But consider. When I now say that I at this moment do not understand Chinese, that claim does not merely record an intuition of mine, something I find myself inclined to say. It is a plain fact about me that I don't understand Chinese. Furthermore, in a situation in which I am given a set of rules for manipulating uninterpreted Chinese symbols, rules that allow no possibility of attaching any semantic content to these Chinese symbols, it is still a fact about me that I do not understand Chinese. Indeed, it is the very same fact as before. But, **Wilensky** suggests, suppose that among those rules for manipulating symbols are some that are Chinese for "Do you understand Chinese?," and in response to these I hand back the Chinese symbols for "Of course I understand Chinese." Does that show, as Wilensky implies, that there is a subsystem in me that understands Chinese? As long as there is no semantic content attaching to these symbols, the fact remains that there is no understanding.

The form of **Block**'s argument about intuition is that since there are allegedly empirical data to show that thinking is just formal symbol manipulation, we could not refute the thesis with untutored intuitions. One might as well try to refute the view that the earth is round by appealing to our intuition that it is flat. Now Block concedes that it is not a matter of intuition but a plain fact that our brains are "the seat" of our intentionality. I want to add that it is equally a plain fact that I don't understand Chinese. My paper is an attempt to explore the logical consequences of these and other such plain facts. Intuitions in his deprecatory sense have nothing to do with the argument. One consequence is that the formal symbol manipulations could not be constitutive of thinking. Block never comes to grips with the arguments for this consequence. He simply laments the feebleness of our intuitions.

Dennett thinks that he can generate counterintuitions. Suppose, in the "robot reply," that the robot is my very own body. What then? Wouldn't I understand Chinese then? Well, the trouble is that the case, as he gives it to us, is underdescribed, because we are never told what is going on in the mind of the agent. (Remember, in these discussions, always insist on the first person point of view. The first step in the operationalist sleight of hand occurs when we try to figure out how we would *know* what it would be like for others.) If we describe Dennett's case sufficiently explicitly it is not hard to see what the facts would be. Suppose that the program contains such instructions as the following: when somebody holds up the squiggle-squiggle sign, pass him the salt. With such instructions it wouldn't take one long to figure out that "squiggle squiggle" probably means pass the salt. But now the agent is starting to learn Chinese from following the program. But this "intuition" doesn't run counter to the facts I was pointing out, for what the agent is doing in such a case is attaching a semantic content to a formal symbol and thus taking a step toward language comprehension. It would be equally possible to describe a case in such a way that it was impossible to attach any semantic content, even though my own body was in question, and in such a case it would be impossible for me to learn Chinese from following the program. Dennett's examples do not generate counterintuitions, they are simply so inadequately described that we can't tell from his description what the facts would be.

At one point **Dennett** and I really do have contrary intuitions. He says "*I* understand English my brain doesn't." I think on the contrary that when I understand English; it is my brain that is doing the work. I find nothing at all odd about saying that my brain understands English, or indeed about saying that my brain is conscious. I find his claim as implausible as insisting, "I digest pizza; my stomach and digestive tract don't."

Marshall suggests that the claim that thermostats don't have beliefs is just as refutable by subsequent scientific discovery as the claim that tables are solid. But notice the difference. In the case of tables we discovered previously unknown facts about the microstructure of apparently solid objects. In the case of thermostats the relevant facts are all quite well known already. Of course such facts as that thermostats don't have beliefs and that I don't speak Chinese are, like all empirical facts, subject to disconfirmation. We might for example discover that, contrary to my deepest beliefs, I am a competent speaker of Mandarin. But think how we would establish such a thing. At a minimum we would have to establish that, quite unconsciously, I know the meanings of a large number of Chinese expressions; and to establish that thermostats had beliefs, in exactly the same sense that I do, we would have to establish, for example, that by some miracle thermostats had nervous systems capable of supporting mental states, and so on. In sum, though in some sense intuition figures in any argument, you will mistake the nature of the present dispute entirely if you think it is a matter of my intuitions against someone else's, or that some set of contrary intuitions has equal validity. The claim that I don't speak Chinese and that my thermostat lacks beliefs aren't just things that I somehow find myself mysteriously inclined to say.

Finally, in response to **Dennett** (and also **Pylyshyn**), I do not, of course, think that intentionality is a fluid. Nor does anything I say commit me to that view. I think, on the contrary, that intentional states, processes, and events are precisely that: states, processes, and events. The point is that they are both caused by and realized in the structure of the brain. Dennett assures me that such a view runs counter to "the prevailing winds of doctrine." So much the worse for the prevailing winds.

Intrinsic intentionality and observer-relative ascriptions of intentionality. Why then do people feel inclined to say that, in some sense at least, thermostats have beliefs? I think that in order to understand what is going on when people make such claims we need to distinguish carefully between cases of what I will call *intrinsic intentionality,* which are cases of actual mental states, and what I will call *observer-relative ascriptions of intentionality,* which are ways that people have of speaking about entities figuring in our activities but lacking intrinsic intentionality. We can illustrate this distinction with examples that are quite uncontroversial. If I say that I am hungry or that Carter believes he can win the election, the form of intentionality in question is intrinsic. I am discussing, truly or falsely, certain psychological facts about me and Carter. But if I say the word "Carter" refers to the present president, or the sentence "Es regnet" means it's raining, I am not ascribing any mental states to the word "Carter" or the sentence "Es regnet." These are ascriptions of intentionality made to entities that lack any mental states, but in which the ascription is a manner of speaking about the intentionality of the observers. It is a way of saying that people *use* the name Carter to refer, or that when people say literally "Es regnet" they *mean* it's raining.

Observer-relative ascriptions of intentionality are always

dependent on the intrinsic intentionality of the observers. There are not two kinds of intentional mental states; there is only one kind, those that have intrinsic intentionality; but there are ascriptions of intentionality in which the ascription does not ascribe intrinsic intentionality to the subject of the ascription. Now I believe that a great deal of the present dispute rests on a failure to appreciate this distinction. When **McCarthy** stoutly maintains that thermostats have beliefs, he is confusing observer-relative ascriptions of intentionality with ascriptions of intrinsic intentionality. To see this point, ask yourself why we make these attributions to thermostats and the like at all. It is not because we suppose they have a mental life very much like our own; on the contrary, we know that they have no mental life at all. Rather, it is because we have designed them (our intentionality) to serve certain of our purposes (more of our intentionality), to perform the sort of functions that we perform on the basis of our intentionality. I believe it is equally clear that our ascription of intentionality to cars, computers, and adding machines is observer relative.

Functionalism, by the way, is an entire system erected on the failure to see this distinction. Functional attributions are always observer relative. There is no such thing as an intrinsic function, in the way that there are intrinsic intentional states.

Natural kinds. This distinction between intrinsic intentionality and observer-relative ascriptions of intentionality might seem less important if we could, as several commentators (**Minsky, Block, Marshall**) suggest, assimilate intrinsic intentionality to some larger natural kind that would subsume both existing mental phenomena and other natural phenomena under a more general explanatory apparatus. Minsky says that "prescientific idea germs like 'believe' " have no place in the mind science of the future (presumably "mind" will also have no place in the "mind science" of the future). But even if this is true, it is really quite irrelevant to my argument, which is addressed to the mind science of the present. Even if, as Minsky suggests, we eventually come to talk of our present beliefs as if they were on a continuum with things that are not intentional states at all, this does not alter the fact that we do have intrinsic beliefs and computers and thermostats do not. That is, even if some future science comes up with a category that supersedes belief and thus enables us to place thermostats and people on a single continuum, this would not alter the fact that under our present concept of belief, people literally have beliefs and thermostats don't. Nor would it refute my diagnosis of the mistake of attributing intrinsic mental states to thermostats as based on a confusion between intrinsic intentionality and observer-relative ascriptions of intentionality.

Minsky further points out that our own mental operations are often split into parts that are not fully integrated by any "self" and only some of which carry on interpretation. And, he asks, if that is how it is in our own minds, why not in computers as well? The reply is that even if there are parts of our mental processes where processing takes place without any intentional content, there still have to be other parts that attach semantic content to syntactic elements if there is to be any understanding to all. The point of the Chinese room example is that the formal symbol manipulations never by themselves carry any semantic content, and thus instantiating a computer program is not by itself sufficient for understanding.

How the brain works. Several commentators take me to task because I don't explain how the brain works to produce intentionality, and at least two (**Dennett** and **Fodor**) object to my claim that where intentionality is concerned – as opposed to the conditions of satisfaction of the intentionality – what matters are the internal and not the external causes. Well I don't know *how* the brain produces mental phenomena, and apparently no one else does either, but *that* it produces mental phenomena and that the internal operations of the brain are causally sufficient for the phenomena is fairly evident from what we do know.

Consider the following case, in which we do know a little about how the brain works. From where I am seated, I can see a tree. Light reflected from the tree in the form of photons strikes my optical apparatus. This sets up a series of sequences of neural firings. Some of these neurons in the visual cortex are in fact remarkably specialized to respond to certain sorts of visual stimuli. When the whole set of sequences occurs, it causes a visual experience, and the visual experience has intentionality. It is a conscious mental event with an intentional content; that is, its conditions of satisfaction are internal to it. Now I could be having exactly that visual experience even if there were no tree there, provided only that something was going on in my brain sufficient to produce the experience. In such a case I would not *see* the tree but would be having a hallucination. In such a case, therefore, the intentionality is a matter of the *internal* causes; whether the intentionality is satisfied, that is, whether I actually see a tree as opposed to having a hallucination of the tree, is a matter of the *external* causes as well. If I were a brain in a vat I could have exactly the same mental states I have now; it is just that most of them would be false or otherwise unsatisfied. Now this simple example of visual experience is designed to make clear what I have in mind when I say that the operation of the brain is causally sufficient for intentionality, and that it is the operation of the brain and not the impact of the outside world that matters for the content of our intentional states, in at least one important sense of "content."

Some of the commentators seem to suppose that I take the causal powers of the brain by themselves to be an argument against strong AI. But that is a misunderstanding. It is an empirical question whether any given machine has causal powers equivalent to the brain. My argument against strong AI is that instantiating a program is not enough to guarantee that it has those causal powers.

Wait till next year. Many authors (**Block, Sloman & Croucher, Dennett, Lycan, Bridgeman, Schank**) claim that Schank's program is just not good enough but that newer and better programs will defeat my objection. I think this misses the point of the objection. My objection would hold against any program at all, qua formal computer program. Nor does it help the argument to add the causal theory of reference, for even if the formal tokens in the program have some causal connection to their alleged referents in the real world, as long as the agent has no way of knowing that, it adds no intentionality whatever to the formal tokens. Suppose, for example, that the symbol for egg foo yung in the Chinese room is actually causally connected to egg foo yung. Still, the man in the room has no way of knowing that. For him, it remains an uninterpreted formal symbol, with no semantic content whatever. I will return to this last point in the discussion of specific authors, especially **Fodor.**

Seriatim. I now turn, with the usual apologies for brevity, from these more general considerations to a series of specific arguments.

Haugeland has an argument that is genuinely original. Suppose a Chinese speaker has her neurons coated with a thin coating that prevents neuron firing. Suppose "Searle's demon" fills the gap by stimulating the neurons as if they had been fired. Then she will understand Chinese even though none of her neurons has the right causal powers; the demon has them, and he understands only English.

My objection is only to the last sentence. Her neurons still have the right causal powers; they just need some help from the demon. More generally if the stimulation of the causes is

at a low enough level to *reproduce* the causes and not merely *describe* them, the "simulation" will reproduce the effects. If what the demon does is reproduce the right causal phenomena, he will have reproduced the intentionality, which constitutes the effects of that phenomena. And it does not, for example, show that my brain lacks the capacity for consciousness if someone has to wake me up in the morning by massaging my head.

Haugeland's distinction between original and derivative intentionality is somewhat like mine between intrinsic intentionality and observer-relative ascriptions of intentionality. But he is mistaken in thinking that the only distinction is that original intentionality is "sufficiently rich" in its "semantic activity": the semantic activity in question is still observer-relative and hence not sufficient for intentionality. My car engine is, in his observer-relative sense, semantically active in all sorts of "rich" ways, but it has no intentionality. A human infant is semantically rather inactive, but it still has intentionality.

Rorty sets up an argument concerning transubstantiation that is formally parallel to mine concerning intrinsic and observer-relative attributions of intentionality. Since the premises of the transubstantiation argument are presumably false, the parallel is supposed to be an objection to my argument. But the parallel is totally irrelevant. Any valid argument whatever from true premises to true conclusions has exact formal analogues from false premises to false conclusions. Parallel to the familiar "Socrates is mortal" argument we have "Socrates is a dog. All dogs have three heads. Therefore Socrates has three heads." The possibility of such formal parallels does nothing to weaken the original arguments. To show that the parallel was insightful Rorty would have to show that my premises are as unfounded as the doctrine of transubstantiation. But what are my premises? They are such things as that people have mental states such as beliefs, desires, and visual experiences, that they also have brains, and that their mental states are causally the products of the operation of their brains. Rorty says nothing whatever to show that these propositions are false, and I frankly can't suppose that he doubts their truth. Would he like evidence for these three? He concludes by lamenting that if my views gain currency the "good work" of his favorite behaviorist and functionalist authors will be "undone." This is not a prospect I find at all distressing, since implicit in my whole account is the view that people really do have mental states, and to say so is not just to ascribe to them tendencies to behave, or to adopt a certain kind of stance toward them, or to suggest functional explanations of their behaviours. This does not give the mental a "numinous Cartesian glow," it just implies that mental processes are as real as any other biological processes.

McCarthy and **Wilensky** both endorse the "systems reply." The major addition made by Wilensky is to suppose that we ask the Chinese subsystem whether it speaks Chinese and it answers yes. I have already suggested that this adds no plausibility whatever to the claim that there is any Chinese understanding going on in the system. Both Wilensky and McCarthy fail to answer the three objections I made to the systems reply.

1. The Chinese subsystem still attaches no semantic content whatever to the formal tokens. The English subsystem knows that "hamburger" means hamburger. The Chinese subsystem knows only that squiggle squiggle is followed by squoggle squoggle.

2. The systems reply is totally unmotivated. Its only motivation is the Turing test, and to appeal to that is precisely to beg the question by assuming what is in dispute.

3. The systems reply has the consequence that all sorts of systematic input-output relations (for example, digestion) would have to count as understanding, since they warrant as much observer-relative ascription of intentionality as does the Chinese subsystem. (And it is, by the way, no answer to this point to appeal to the cognitive impenetrability of digestion, in Pylyshyn's [1980a] sense, since digestion is congnitively penetrable: the content of my beliefs can upset my digestion.)

Wilensky seems to think that it is an objection that other sorts of mental states besides intentional ones could have been made the subject of the argument. But I quite agree. I could have made the argument about pains, tickles, and anxiety, but these are (a) less interesting to me and (b) less discussed in the AI literature. I prefer to attack strong AI on what its proponents take to be their strongest ground.

Pylyshyn misstates my argument. I offer no a priori proof that a system of integrated circuit chips couldn't have intentionality. That is, as I say repeatedly, an empirical question. What I do argue is that in order to produce intentionality the system would have to duplicate the causal powers of the brain and that simply instantiating a formal program would not be sufficient for that. Pylyshyn offers no answer to the arguments I give for these conclusions.

Since **Pylyshyn** is not the only one who has this misunderstanding, it is perhaps worth emphasizing just what is at stake. The position of strong AI is that anything with the right program would have to have the relevant intentionality. The circuit chips in his example would necessarily have intentionality, and it wouldn't matter if they were circuit chips or water pipes or paper clips, provided they instantiated the program. Now I argue at some length that they couldn't have intentionality solely in virtue of instantiating the program. Once you see that the program doesn't necessarily add intentionality to a system, it then becomes an empirical question which kinds of systems really do have intentionality, and the condition necessary for that is that they must have causal powers equivalent to those of the brain. I think it is evident that all sorts of substances in the world, like water pipes and toilet paper, are going to lack those powers, but that is an empirical claim on my part. On my account it is a testable empirical claim whether in repairing a damaged brain we could duplicate the electrochemical basis of intentionality using some other substance, say silicon. On the position of strong AI there cannot be any empirical questions about the electrochemical bases necessary for intentionality since any substance whatever is sufficient for intentionality if it has the right program. I am simply trying to lay bare for all to see the full preposterousness of that view.

I believe that **Pylyshyn** also misunderstands the distinction between intrinsic and observer-relative ascriptions of intentionality. The relevant question is not how much latitude the observer has in making observer-relative ascriptions, but whether there is any intrinsic intentionality in the system to which the ascriptions could correspond.

Schank and I would appear to be in agreement on many issues, but there is at least one small misunderstanding. He thinks I want "to call into question the enterprise of AI." That is not true. I am all in favor of weak AI, at least as a research program. I entirely agree that if someone could write a program that would give the right input and output for Chinese stories it would be a "great achievement" requiring a "great understanding of the nature of language." I am not even sure it can be done. My point is that instantiating the program is not constitutive of understanding.

Abelson, like **Schank,** points out that it is no mean feat to program computers that can simulate story understanding. But, to repeat, that is an achievement of what I call weak AI, and I would enthusiastically applaud it. He mars this valid point by insisting that since our own understanding of most things, arithmetic for example, is very imperfect, "we might well be humble and give the computer the benefit of the doubt when and if it performs as well as we do." I am afraid that neither this nor his other points meets my arguments to

show that, humble as we would wish to be, there is no reason to suppose that instantiating a formal program in the way a computer does is any reason *at all* for ascribing intentionality to it.

Fodor agrees with my central thesis that instantiating a program is not a sufficient condition of intentionality. He thinks, however, that if we got the right causal links between the formal symbols and things in the world that would be sufficient. Now there is an obvious objection to this variant of the robot reply that I have made several times: the same thought experiment as before applies to this case. That is, no matter what outside causal impacts there are on the formal tokens, these are not by themselves sufficient to give the tokens any intentional content. No matter what caused the tokens, the agent still doesn't understand Chinese. Let the egg foo yung symbol be causally connected to egg foo yung in any way you like, that connection by itself will never enable the agent to interpret the symbol as meaning egg foo young. To do that he would have to have, for example, some *awareness* of the causal relation between the symbol and the referent; but now we are no longer explaining intentionality in terms of symbols and causes but in terms of symbols, causes, and intentionality, and we have abandoned both strong AI and the robot reply. Fodor's only answer to this is to say that it shows we haven't yet got the right kind of causal linkage. But what is the right kind, since the above argument applies to any kind? He says he can't tell us, but it is there all the same. Well I can tell him what it is: it is any form of causation sufficient to produce intentional content in the agent, sufficient to produce, for example, a visual experience, or a memory, or a belief, or a semantic interpretation of some word.

Fodor's variant of the robot reply is therefore confronted with a dilemma. If the causal linkages are just matters of fact about the relations between the symbols and the outside world, they will never by themselves give any interpretation to the symbols; they will carry by themselves no intentional content. If, on the other hand, the causal impact is sufficient to produce intentionality in the agent, it can only be because there is something more to the system than the *fact* of the causal impact and the *symbol*, namely the intentional content that the impact produces in the agent. Either the man in the room doesn't learn the meaning of the symbol from the causal impact, in which case the causal impact adds nothing to the interpretation, or the causal impact teaches him the meaning of the word, in which case the cause is relevant only because it produces a form of intentionality that is something in addition to itself and the symbol. In neither case is symbol, or cause and symbol, constitutive of intentionality.

This is not the place to discuss the general role of formal processes in mental processes, but I cannot resist calling attention to one massive use-mention confusion implicit in **Fodor's** account. From the fact that, for example, syntactical rules concern formal objects, it does not follow that they are formal rules. Like other rules affecting human behavior they are defined by their content, not their form. It just so happens that in this case their content concerns forms.

In what is perhaps his crucial point, **Fodor** suggests that we should think of the brain or the computer as performing formal operations only on *interpreted* and not just on *formal* symbols. But who does the interpreting? And what is an interpretation? If he is saying that for intentionality there must be intentional content in addition to the formal symbols, then I of course agree. Indeed, two of the main points of my argument are that in our own case we have the "interpretation," that is, we have intrinsic intentionality, and that the computer program could never by itself be sufficient for that. In the case of the computer we make observer-relative ascriptions of intentionality, but that should not be mistaken for the real thing since the computer program by itself has no intrinsic intentionality.

Sloman & Croucher claim that the problem in my thought experiment is that the system isn't big enough. To Schank's story understander they would add all sorts of other operations, but they emphasize that these operations are computational and not physical. The obvious objection to their proposal is one they anticipate: I can still repeat my thought experiment with their system no matter how big it is. To this, they reply that I assume "without argument, that it is impossible for another mind to be based on his [my] mental process without his [my] knowing." But that is not what I assume. For all I know, that may be false. Rather, what I assume is that you can't understand Chinese if you don't know the meanings of any of the words in Chinese. More generally, unless a system can attach semantic content to a set of syntactic elements, the introduction of the elements in the system adds nothing by way of intentionality. That goes for me and for all the little subsystems that are being postulated inside me.

Eccles points out quite correctly that I never undertake to refute the dualist-interaction position held by him and Popper. Instead, I argue against strong AI on the basis of what might be called a monist interactionist position. My only excuse for not attacking his form of dualism head-on is that this paper really had other aims. I am concerned directly with strong AI and only incidentally with the "mind-brain problem." He is quite right in thinking that my arguments against strong AI are not by themselves inconsistent with his version of dualist interactionism, and I am pleased to see that we share the belief that "it is high time that strong AI was discredited."

I fear I have nothing original to say about **Rachlin's** behaviorist response, and if I discussed it I would make only the usual objections to extreme behaviorism. In my own case I have an extra difficulty with behaviorism and functionalism because I cannot imagine anybody actually believing these views. I know that poeple say they do, but what am I to make of it when Rachlin says that there are no "mental states underlying . . . behavior" and "the pattern of the behavior *is* the mental state"? Are there no pains underlying Rachlin's pain behavior? For my own case I must confess that there unfortunately often are pains underlying my pain behavior, and I therefore conclude that Rachlin's form of behaviorism is not generally true.

Lycan tells us that my counterexamples are not counterexamples to a functionalist theory of language understanding, because the man in my counterexample would be using the wrong programs. Fine. Then tell us what the right programs are, and we will program the man with those programs and still produce a counterexample. He also tells us that the right causal connections will determine the appropriate content to attach to the formal symbols. I believe my reply to **Fodor** and other versions of the causal or robot reply is relevant to his argument as well, and so I will not repeat it.

Hofstadter cheerfully describes my target article as "one of the wrongest, most infuriating articles I have ever read in my life." I believe that he would have been less (or perhaps more?) infuriated if he had troubled to read the article at all carefully. His general strategy appears to be that whenever I assert p, he says that I assert not p. For example, I reject dualism, so he says I believe in the soul. I think it is a plain fact of nature that mental phenomena are caused by neurophysiological phenomena, so he says I have "deep difficulty" in accepting any such view. The whole tone of my article is one of treating the mind as part of the (physical) world like anything else, so he says I have an "instinctive horror" of any such reductionism. He misrepresents my views at almost every point, and in consequence I find it difficult to take his commentary seriously. If my text is too difficult I suggest Hofstadter read **Eccles** who correctly perceives my rejection of dualism.

Furthermore, **Hofstadter's** commentary contains the

following non sequitur. From the fact that intentionality "springs from" the brain, together with the extra premise that "physical processes are formal, that is, rule governed" he infers that formal processes are constitutive of the mental, that we are "at bottom, formal systems." But that conclusion simply does not follow from the two premises. It does not even follow given his weird interpretation of the second premise: "To put it another way, the extra premise is that there is no intentionality at the level of particles." I can accept all these premises, but they just do not entail the conclusion. They do entail that intentionality is an "*outcome* of formal processess" in the trivial sense that it is an outcome of processes that have a level of description at which they are the instantiation of a computer program, but the same is true of milk and sugar and countless other "outcomes of formal processes."

Hofstadter also hypothesizes that perhaps a few trillion water pipes might work to produce consciousness, but he fails to come to grips with the crucial element of my argument, which is that even if this were the case it would have to be because the water-pipe system was duplicating the causal powers of the brain and not simply instantiating a formal program.

I think I agree with **Smythe's** subtle commentary except perhaps on one point. He seems to suppose that to the extent that the program is instantiated by "primitive hardware operations" my objections would not apply. But why? Let the man in my example have the program mapped into his hardware. He still doesn't thereby understand Chinese. Suppose he is so "hard wired" that he automatically comes out with uninterpreted Chinese sentences in response to uninterpreted Chinese stimuli. The case is still the same except that he is no longer acting voluntarily.

Side issues. I felt that some of the commentators missed the point or concentrated on peripheral issues, so my remarks about them will be even briefer.

I believe that **Bridgeman** has missed the point of my argument when he claims that though the homunculus in my example might not know what was going on, it could soon learn, and that it would simply need more information, specifically "information with a known relationship to the outside world." I quite agree. To the extent that the homunculus has such information it is more than a mere instantiation of a computer program, and thus it is irrelevant to my dispute with strong AI. According to strong AI, if the homunculus has the right program it must already have the information. But I disagree with Bridgeman's claim that the only properties of the brain are the properties it has at the level of neurons. I think all sides to the present dispute would agree that the brain has all sorts of properties that are not ascribable at the level of individual neurons – for example, causal properties (such as the brain's control of breathing).

Similar misgivings apply to the remarks of **Marshall.** He stoutly denounces the idea that there is anything weak about the great achievements of weak AI, and concludes "Clearly, there must be some radical misunderstanding here." The only misunderstanding was in his supposing that in contrasting weak with strong AI, I was in some way disparaging the former.

Marshall finds it strange that anyone should think that a program could be a theory. But the word program is used ambiguously. Sometimes "program" refers to the pile of punch cards, sometimes to a set of statements. It is in the latter sense that the programs are sometimes supposed to be theories. If Marshall objects to that sense, the dispute is still merely verbal and can be resolved by saying not that the program is a theory, but that the program is an embodiment of a theory. And the idea that programs could be theories is not something I invented. Consider the following. "Occasionally after seeing what a program can do, someone will ask for a specification of the theory behind it. Often the correct response is that the program *is* the theory" (Winston 1977, p. 259).

Ringle also missed my point. He says I take refuge in mysticism by arguing that "the physical properties of neuronal systems are such that they cannot *in principle* be simulated by a nonprotoplasmic computer." But that is not even remotely close to my claim. I think that anything can be given a formal simulation, and it is an empirical question in each case whether the simulation duplicated the causal features. The question is whether the formal simulation by *itself*, without any further causal elements, is sufficient to reproduce the mental. And the answer to that question is no, because of the arguments I have stated repeatedly, and which Ringle does not answer. It is just a fallacy to suppose that because the brain has a program and because the computer could have the same program, that what the brain does is nothing more than what the computer does. It is for each case an empirical question whether a rival system duplicates the causal powers of the brain, but it is a quite different question whether instantiating a formal program is by itself constitutive of the mental.

I also have the feeling, perhaps based on a misunderstanding, that **Menzel's** discussion is based on a confusion between *how one knows* that some system has mental states and *what it is* to have a mental state. He assumes that I am looking for a criterion for the mental, and he cannot see the point in my saying such vague things about the brain. But I am not in any sense looking for a criterion for the mental. I know what mental states are, at least in part, by myself being a system of mental states. My objection to strong AI is not, as Menzel claims, that it might fail in a single possible instance, but rather that in the instance in which it fails, it possesses no more resources than in any other instance; hence if it fails in that instance it fails in every instance.

I fail to detect any arguments in **Walter's** paper, only a few weak analogies. He laments my failure to make my views on intentionality more explicit. They are so made in the three papers cited by **Natsoulas** (Searle 1979a; 1979b; 1979c).

Further implications. I can only express my appreciation for the contributions of **Danto, Libet, Maxwell, Puccetti,** and **Natsoulas.** In various ways, they each add supporting arguments and commentary to the main thesis. Both Natsoulas and Maxwell challenge me to provide some answers to questions about the relevance of the discussion to the traditional ontological and mind-body issues. I try to avoid as much as possible the traditional vocabulary and categories, and my own – very tentative – picture is this. Mental states are as real as any other biological phenomena. They are both caused by and realized in the brain. That is no more mysterious than the fact that such properties as the elasticity and puncture resistance of an inflated car tire are both caused by and realized in its microstructure. Of course, this does not imply that mental states are ascribable to individual neurons, any more than the properties at the level of the tire are ascribable to individual electrons. To pursue the analogy: the brain operates causally both at the level of the neurons and at the level of the mental states, in the same sense that the tire operates causally both at the level of particles and at the level of its overall properties. Mental states are no more epiphenomenal than are the elasticity and puncture resistance of an inflated tire, and interactions can be described both at the higher and lower levels, just as in the analogous case of the tire.

Some, but not all, mental states are conscious, and the intentional-nonintentional distinction cuts across the conscious-unconscious distinction. At every level the phenomena are causal. I suppose this is "interactionism," and I guess it is

also, in some sense, "monism," but I would prefer to avoid this myth-eaten vocabulary altogether.

Conclusion. I conclude that the Chinese room has survived the assaults of its critics. The remaining puzzle to me is this: why do so many workers in AI still want to adhere to strong AI? Surely weak AI is challenging, interesting, and difficult enough.

ACKNOWLEDGMENT

I am indebted to Paul Kube for discussion of these issues.

References

Anderson, J. (1980) Cognitive units. Paper presented at the Society for Philosophy and Psychology, Ann Arbor, Mich. [RCS]

Block, N. J. (1978) Troubles with functionalism. In: *Minnesota studies in the philosophy of science, vol. 9*, ed. C. W. Savage, Minneapolis: University of Minnesota Press. [NB, WGL]

(forthcoming) Psychologism and behaviorism. *Philosophical Review*. [NB, WGL]

Bower, G. H.; Black, J. B., & Turner, T. J. (1979) Scripts in text comprehension and memory. *Cognitive Psychology* 11: 177–220. [RCS]

Carroll, C. W. (1975) *The great chess automaton*. New York: Dover. [RP]

Cummins, R. (1977) Programs in the explanation of behavior. *Philosophy of Science* 44: 269–87. [JCM]

Dennett, D. C. (1969) *Content and consciousness*. London: Routledge & Kegan Paul. [DD,TN]

(1971) Intentional systems. *Journal of Philosophy* 68: 87–106. [TN]

(1972) Reply to Arbib and Gunderson. Paper presented at the Eastern Division meeting of the American Philosophical Association. Boston, Mass. [TN]

(1975) Why the law of effect won't go away. *Journal for the Theory of Social Behavior* 5: 169–87. [NB]

(1978) *Brainstorms*. Montgomery, Vt.: Bradford Books. [DD, AS]

Eccles, J. C. (1978) A critical appraisal of brain-mind theories. In: *Cerebral correlates of conscious experiences*, ed. P. A. Buser and A. Rougeul-Buser, pp. 347–55. Amsterdam: North Holland. [JCE]

(1979) *The human mystery*. Heidelberg: Springer Verlag. [JCE]

Fodor, J. A. (1968) The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy* 65: 627–40. [NB]

(1980) Methodological solopsism considered as a research strategy in cognitive psychology. *The Behavioral and Brain Sciences* 3:1. [NB, WGL, WES]

Freud, S. (1895) Project for a scientific psychology. In: *The standard edition of the complete psychological works of Sigmund Freud*, vol. 1, ed. J. Strachey. London: Hogarth Press, 1966. [JCM]

Frey, P. W. (1977) An introduction to computer chess. In: *Chess skill in man and machine*, ed. P. W. Frey. New York, Heidelberg, Berlin: Springer-Verlag. [RP]

Fryer, D. M. & Marshall, J. C. (1979) The motives of Jacques de Vaucanson. *Technology and Culture* 20: 257–69. [JCM]

Gibson, J. J. (1966) *The senses considered as perceptual systems*. Boston: Houghton Mifflin. [TN]

(1967) New reasons for realism. *Synthese* 17: 162–72. [TN]

(1972) A theory of direct visual perception. In: *The psychology of knowing* ed. S. R. Royce & W. W. Rozeboom. New York: Gordon & Breach. [TN]

Graesser, A. C.; Gordon, S. E.; & Sawyer, J. D. (1979) Recognition memory for typical and atypical actions in scripted activities: tests for a script pointer and tag hypotheses. *Journal of Verbal Learning and Verbal Behavior* 1: 319–32. [RCS]

Gruendel, J. (1980). Scripts and stories: a study of children's event narratives. Ph.D. dissertation, Yale University. [RCS]

Hanson, N. R. (1969) *Perception and discovery*. San Francisco: Freeman, Cooper. [DOW]

Hayes, P. J. (1977) In defence of logic. In: *Proceedings of the 5th international joint conference on artificial intelligence*, ed. R. Reddy. Cambridge, Mass.: M.I.T. Press. [WES]

Hobbes, T. (1651) *Leviathan*. London: Willis. [JCM]

Hofstadter, D. R. (1979) *Gödel, Escher, Bach*. New York: Basic Books. [DOW]

Householder, F. W. (1962) On the uniqueness of semantic mapping. *Word* 18: 173–85. [JCM]

Huxley, T. H. (1874) On the hypothesis that animals are automata and its history. In: *Collected Essays*, vol. 1. London: Macmillan, 1893. [JCM]

Kolers, P. A. & Smythe, W. E. (1979) Images, symbols, and skills. *Canadian Journal of Psychology* 33: 158–84. [WES]

Kosslyn, S. M. & Shwartz, S. P. (1977) A simulation of visual imagery. *Cognitive Science* 1: 265–95. [WES]

Lenneberg, E. H. (1975) A neuropsychological comparison between man, chimpanzee and monkey. *Neuropsychologia* 13: 125. [JCE]

Libet, B. (1973) Electrical stimulation of cortex in human subjects and conscious sensory aspects. In: *Handbook of sensory physiology*, vol. II, ed. A. Iggo, pp. 743–90. New York: Springer-Verlag. [BL]

Libet, B., Wright, E. W., Jr., Feinstein, B., and Pearl, D. K. (1979) Subjective referral of the timing for a conscious sensory experience: a functional role for the somatosensory specific projection system in man. *Brain* 102:191–222. [BL]

Longuet-Higgins, H. C. (1979) The perception of music. *Proceedings of the Royal Society of London* B 205:307–22. [JCM]

Lucas, J. R. (1961) Minds, machines, and Gödel. *Philosophy* 36:112–127. [DRH]

Lycan, W. G. (forthcoming) Form, function, and feel. *Journal of Philosophy*. [NB, WGL]

McCarthy, J. (1979) Ascribing mental qualities to machines. In: *Philosophical perspectives in artificial intelligence*, ed. M. Ringle. Atlantic Highlands, N.J.: Humanities Press. [JM, JRS]

Marr, D. & Poggio, T. (1979) A computational theory of human stereo vision. *Proceedings of the Royal Society of London* B 204:301–28. [JCM]

Marshall, J. C. (1971) Can humans talk? In: *Biological and social factors in psycholinguistics*, ed. J. Morton. London: Logos Press. [JCM]

(1977) Minds, machines and metaphors. *Social Studies of Science* 7:475–88. [JCM]

Maxwell, G. (1976) Scientific results and the mind-brain issue. In: *Consciousness and the brain*, ed. G. G. Globus, G. Maxwell, & I. Savodnik. New York: Plenum Press. [GM]

(1978) Rigid designators and mind-brain identity. In: *Perception and cognition: Issues in the foundations of psychology*, Minnesota Studies in the Philosophy of Science, vol. 9, ed. C. W. Savage. Minneapolis: University of Minnesota Press. [GM]

Mersenne, M. (1636) *Harmonie universelle*. Paris: Le Gras. [JCM]

Moor, J. H. (1978) Three myths of computer science. *British Journal of the Philosophy of Science* 29:213–22. [JCM]

Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83:435–50. [GM]

Natsoulas, T. (1974) The subjective, experiential element in perception. *Psychological Bulletin* 81:611–31. [TN]

(1977) On perceptual aboutness. *Behaviorism* 5:75–97. [TN]

(1978a) Haugeland's first hurdle. *Behavioral and Brain Sciences* 1:243. [TN]

(1979b) Residual subjectivity. *American Psychologist* 33:269–83. [TN]

(1980) Dimensions of perceptual awareness. Psychology Department, University of California, Davis. Unpublished manuscript. [TN]

Nelson, K. & Gruendel, J. (1978) From person episode to social script: two dimensions in the development of event knowledge. Paper presented at the biennial meeting of the Society for Research in Child Development, San Francisco. [RCS]

Newell, A. (1973) Production systems: models of control structures. In: *Visual information processing*, ed. W. C. Chase. New York: Academic Press. [WES]

(1979) Physical symbol systems. Lecture at the La Jolla Conference on Cognitive Science. [JRS]

(1980) Harpy, production systems, and human cognition. In: *Perception and production of fluent speech*, ed. R. Cole. Hillsdale, N.J.: Erlbaum Press. [WES]

Newell, A. & Simon, H. A. (1963) GPS, a program that simulates human thought. In: *Computers and thought*, ed. A. Feigenbaum & V. Feldman, pp. 279–93. New York: McGraw Hill. [JRS]

Panofsky, E. (1954) *Galileo as a critic of the arts*. The Hague: Martinus Nijhoff. [JCM]

Popper, K. R. & Eccles, J. C. (1977) *The self and its brain*. Heidelberg: Springer-Verlag. [JCE, GM]

Putnam, H. (1960) Minds and machines. In: *Dimensions of mind*, ed. S. Hook, pp. 138–64. New York: Collier. [MR, RR]

(1975a) The meaning of "meaning." In: *Mind, language and reality*. Cambridge University Press. [NB, WGL]

(1975b) The nature of mental states. In: *Mind, language and reality*. Cambridge: Cambridge University Press. [NB]

(1975c) Philosophy and our mental life. In: *Mind, language and reality*. Cambridge: Cambridge University Press. [MM]

Pylyshyn, Z. W. (1980a) Computation and cognition: issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3. [JRS, WES]

(1980b) Cognitive representation and the process-architecture distinction. *Behavioral and Brain Sciences*. [ZWP]

Russell, B. (1948) *Human knowledge: its scope and limits*. New York: Simon and Schuster. [GM]

Schank, R. C. & Abelson, R. P. (1977) *Scripts, plans, goals, and understanding*. Hillsdale, N.J.: Lawrence Erlbaum Press. [RCS, JRS]

Searle, J. R. (1979a) Intentionality and the use of language. In: *Meaning and use*, ed. A. Margalit. Dordrecht: Reidel. [TN, JRS]

(1979b) The intentionality of intention and action. *Inquiry* 22:253-80. [TN, JRS]

(1979c) What is an intentional state? *Mind* 88:74–92. [JH, GM, TN, JRS]

Sherrington, C. S. (1950) Introductory. In: *The physical basis of mind*, ed. P. Laslett, Oxford: Basil Blackwell. [JCE]

Slate, J. S. & Atkin, L. R. (1977) CHESS 4.5 – the Northwestern University chess program. In: *Chess skill in man and machine*, ed. P. W. Frey. New York, Heidelberg, Berlin: Springer Verlag.

Sloman, A. (1978) *The computer revolution in philosophy*. Harvester Press and Humanities Press. [AS]

(1979) The primacy of non-communicative language. In: *The analysis of meaning (informatics 5)*, ed. M. McCafferty & K. Gray. London: ASLIB and British Computer Society. [AS]

Smith, E. E.; Adams, N.; & Schorr, D. (1978) Fact retrieval and the paradox of interference. *Cognitive Psychology* 10:438–64. [RCS]

Smythe, W. E. (1979) *The analogical/propositional debate about mental representation: a Goodmanian analysis*. Paper presented at the 5th annual meeting of the Society for Philosophy and Psychology, New York City. [WES]

Sperry, R. W. (1969) A modified concept of consciousness. *Psychological Review* 76:532–36. [TN]

(1970) An objective approach to subjective experience: further explanation of a hypothesis. *Psychological Review* 77:585–90. [TN]

(1976) Mental phenomena as causal determinants in brain function. In: *Consciousness and the brain*, ed. G. G. Globus, G. Maxwell, & I. Savodnik. New York: Plenum Press. [TN]

Stich, S. P. (in preparation) On the ascription of content. In: *Entertaining thoughts*, ed. A. Woodfield. [WGL]

Thorne, J. P. (1968) A computer model for the perception of syntactic structure. *Proceedings of the Royal Society of London* B 171:377–86. [JCM]

Turing, A. M. (1964) Computing machinery and intelligence. In: *Minds and machines*, ed. A. R. Anderson, pp.4–30. Englewood Cliffs, N.J.: Prentice-Hall. [MR]

Weizenbaum, J. (1965) Eliza – a computer program for the study of natural language communication between man and machine. *Communication of the Association for Computing Machinery* 9:36–45. [JRS]

(1976) *Computer power and human reason*. San Francisco: W. H. Freeman. [JRS]

Winograd, T. (1973) A procedural model of language understanding. In: *Computer models of thought and language*, ed. R. Schank & K. Colby. San Francisco: W. H. Freeman. [JRS]

Winston, P. H. (1977) *Artificial intelligence*. Reading, Mass. Addison-Wesley; [JRS]

Woodruff, G. & Premack, D. (1979) Intentional communication in the chimpanzee: the development of deception. *Cognition* 7:333–62. [JCM]