TOM CUDA

AGAINST NEURAL CHAUVINISM*

(Received 25 June, 1984)

John R. Searle ([2]) has argued that functional equivalence to a human being, even at the level of the formal structure of neuron firings, is not a sufficient condition for an organism's having conscious states. Speaking of a brain simulator (i.e., a device in which the functional role of each neuron of a normal human brain is realized by non biological hardware) Searle says: "The problem with the brain simulator is that it is simulating the wrong things about the brain. As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states" ([2], p. 421).

For Searle one thing that is necessary for an organism to have conscious states is that it be made of the proper material (neurons being the most obvious choice).[1] He argues for this mainly through thought experiments, in which creatures are described which are functionally equivalent to human beings (even at the fine level mentioned), but for which it seems counter-intuitive to believe that they have conscious states, because they do not seem to be made of the right kind of material (e.g., they are made of water pipes).

What follows is an argument against Searle's view. It is not an argument that functionalism is true. For example, I say nothing of any dualistic criticisms of functionalism. Rather the paper is directed only at those who, like Searle, are physicalists, but not functionalists on the grounds that functionalism attributes conscious states to things that don't have them (e.g., brain simulators made of water pipes).

I

To begin this argument we must imagine that we have access to a large pool of homunculi that know a great deal about neurophysiology, and that each

Philosophical Studies **48** (1985) 111–127. 0031–8116/85.10

homunculus is equipped with a tiny device that can both read the state of a neuron, and change the state of a neuron.

Now, one day we talk someone, call him Fred, into undergoing the following series of operations: During the first operation Fred's skull is opened up and one of his neurons, call it the A neuron, is removed. But right before the neuron is removed, a homunculus is placed in Fred's skull to take over its functional role.

To see that the homunculus can do this, list all of the neurons that were originally connected directly to A, as $B_1, \ldots, B_n$, and consider that the homunculus does the following: It continuously reads the state of each B_i with its device, and it takes note of the state of A right before it is removed. Say the state of A at this time is S_i. Since the homunculus is keeping track of each B_i, if any B_i is in such a state that it would have sent a message to A, that would have changed the state of A to S_j, had A not been removed, then the homunculus will know that A would have been in S_j. It will know this because it knows what the initial state of A would have been (it would have been S_i) and what the state of each B_i is, and since the homunculus knows a great deal about neurophysiology, it will therefore know any message that would have been sent to A, and how that message would have changed A.

In fact, since the homunculus knows the state of A when it is removed, and since it continuously reads the state of each B_i, it can always keep track of what state A would have been in. Hence the homunculus always knows whether or not A would have sent a signal to any B_i, and how that signal would have changed B_i. Consequently the homunculus can always change the state of each B_i (if a change is called for) to the state that the message from A would have changed it to. This is why A is no longer necessary to the functional organization of Fred.

After a while, there is a second operation, then a third, etc., and after each operation Fred is allowed to go about his business for a few days. Finally, after a trillion or so operations, there is nothing left of the original matter of Fred's brain. At this point, most of the homunculi don't do anything with neurons anymore and have put away their neuron manipulators. Instead, they operate only between themselves, calling out what would have been the state of the neuron that they replaced. (They can keep track of this by paying attention to what other homunculi are calling out.) The rest of the homunculi now adjust, as well as read and call out the state of the input and output neurons to Fred's 'brain'. They can adjust them because they always know what state

the neuron that they replaced would have been in, and hence any message that it would have sent to the input or output neuron that they adjust.

II

I have been calling what is left after each operation 'Fred'. However, someone might argue that after the m^{th} operation what we have is no longer Fred, but rather an unconscious robot, or some conscious thing different from Fred. Later I will argue that this kind of worry is unwarranted. But for right now, I don't want to beg any questions, so let us consider the series of operations to produce a sequence of Freds, and list them as Fred_1 ..., Fred_n, and leave it open as to whether or not they are all the same person. Let Fred_1 be the original normal Fred we began the operations with, and Fred_n that member who has had all of the brain's neurons replaced by homunculi.

Before going on, let me define some terms that will make things easier. Let us think not of the sequence of Freds actually produced, but another, and list them as Fred_1*, ..., Fred_n*. Let Fred_1 and Fred_1* be the same person (i.e., original Fred) and imagine that each Fred_i* has normal human neurophysiology. But also imagine that each Fred_i* goes through the same circumstances, other than neuron removal, that Fred_i goes through. For example, if Fred_i gets fired from his job, so does Fred_i*. They would even be put on the operating table, but no operations would be performed, however they would be told about the 'operations', whatever the Fred_i sequence was told. The following relationship will then hold: Each neuron in each Fred_i has a counterpart in Fred_i* and is in the same state as its counterpart. That this relationship holds is uncontroversial and does not depend on any theory of mind. It holds simply because the homunculi always adjust the remaining neurons in any circumstance, to the state they would have been in had no neurons been removed, and on the fact that the Fred_i* sequence goes through the same circumstances as the Fred_i sequence.[2]

Also, let us say of any Fred_i and Fred_i*, that they are mentally equivalent, just in case they have the same beliefs, desires, qualia, etc. Since Fred_1 and Fred_1* are both original Fred, we know that at least one pair is mentally equivalent.

III

Searle would, of course, argue that Fred_n is unconscious. But I am going to argue that each Fred_i (most importantly Fred_n) is fully conscious. Here is an outline of the argument: First I will show that anyone who maintains that Fred_n is unconscious, is committed to a certain proposition. After that, I will argue that this proposition is such, that being committed to it makes the view that Fred_n is unconscious, very unattractive. Here is a first approximation of the proposition:

(1) No matter in what order we remove the neurons from the Fred_i sequence, there will always be some m, such that Fred_m is mentally equivalent to Fred_m*, and hence fully conscious (since Fred_m* is a neurophysiologically normal human) and such that Fred_{m+1} is completely unconscious.

The paper will be clearer, if I state the refined version of the above proposition that one is actually committed to, if one maintains that Fred_n is unconscious, later, in the course of seeing how one who holds such a view might try to avoid commitment to (1).

IV

The only way that someone holding the view that Fred_n is unconscious could avoid commitment to (1), would be to claim that for some sequence of neuron removal, as we move from Fred_1 to Fred_n, the Fred_i sequence's consciousness fades. For example, one of them starts to lose his hearing, then another some thoughts, etc., until at some point in the sequence they are no longer conscious at all. To see what happens if one tries this, let us break up the view that the Fred_i sequence's consciousness fades, into two separate approaches.

One approach would be to claim that as the Fred_i sequence fades, *they notice* that they are losing their conscious capacities. The other approach would be to claim that as their consciousness fades, they *do not notice*.

As to the first approach, it is really a non starter because the claim that the Fred_i sequence notices that their consciousness is fading, is something that someone holding Searle's attitude about the mental cannot claim. I say this because for any Fred_i, his noticing that he has lost some conscious capacity

will itself be a conscious state, and hence, according to Searle, this state of noticing would have to be tokened in his remaining neurons. However, the neurons remaining in any Fred_i are always in the same state as their counterparts in Fred_i*, and consequently cannot be tokening the noticing of a loss of a part of consciousness, because Fred_i* will not have such a mental state tokened in him (since he is a neurophysiologically normal human, and hence has no such loss).

One might try to argue that the members of the Fred_i sequence could notice that their consciousness is fading because there is just no way that the neurons in each Fred_i could remain in the same state as their counterparts, and hence they could token mental states that are not tokened in Fred_i*. However, such a claim is incredibly *ad hoc*. It amounts to the claim that the homunculi can't do their jobs because it's not possible for anything but a neuron to change the state of a neuron, or read the state of a neuron, in the relevant ways. But nothing that we know about neurons warrants such a belief, and in fact there is strong evidence against it, since there are at present devices which are not neurons that manipulate and read the states of neurons to a certain extent.

The only other possibility open is that for each Fred_i, his neurons are in the same state as their counterparts, but that since the *total* brain state of each Fred_i where $i > 1$, is different than that of Fred_i* (i.e., thair 'brains' contain homunculi) the Fred_i sequence can token in their 'brains' conscious states that the Fred_i* sequence does not have, namely the noticing of a loss of some part of consciousness.

But, once again, there is simply no reason to believe such a thing, and it entails a thesis that should make anyone holding Searle's views very unhappy, namely that homunculi can play at least a partial role in tokening conscious states. This is because on such a view for each Fred_i his neurons are in the same state as their counterparts, hence, if some Fred_i has conscious states that Fred_i* does not have, then the homunculi must play some role in tokening them.[3] So one would have to wonder just how many neurons are necessary to token a conscious state. Might not the homunculi be able to do so on their own, or with the help of just one or two neurons? It would be wildly *ad hoc* to simply assert that homunculi can play a role in tokening conscious states, but that they cannot do so on their own. And the view is seen as even more *ad hoc* when we realize that the presence of the homunculi

is supposed to allow for some Fred_i, not the tokening of just any old conscious state that Fred_i* does not have, but that we have to accept without explanation, that their presence helps bring about the tokening of the proper conscious state, namely that noticing of the loss of some conscious capacity.

Furthermore, here is another unattractive thing that would be entailed: Each Fred_i would be functionally equivalent to Fred_i*, and hence would behave just like Fred_i*, in spite of the fact that he would have some conscious states that Fred_i* would not have, and lack others that Fred_i* would have. For example, Fred_i's sight would begin to fail (or he might go blind altogether) and *he would notice this,* yet he would behave just like Fred_i* (drive a car, say he can see fine, etc.) and not be able to express his loss of vision, not even his knowledge of the loss, in any way. And hence, we would have to accept without explanation, that conscious states which are in part tokened by homunculi are somehow different from normal mental states, in that they do not play a role in behavior.

V

It appears then, that we have eliminated the first approach as to how one might argue that the Fred_i sequence's consciousness fades. Hence, if one who holds that Fred_n is unconscious is goign to avoid being committed to (1), the only road open is to argue that as the Fred_i sequence loses consciousness, the members of the sequence *do not notice* that they are losing their conscious states. I would like to begin exploring this possibility by first showing that the sequence cannot fade by losing only one conscious ability at a time. Afterwards, we will see just what would have to happen.

First let's look at the case of qualia. What would it mean to say of some member of the Fred_i sequence, say Fred_m, that he is mentally equivalent to Fred_m* except that he lacks the ability to have some qualia, and does not notice this? For example, suppose that he no longer has the ability to have red qualia, but is otherwise mentally equivalent to Fred_m*. Fred_m would have to, at times, believe that he is having a red quale, behave just like he had them, desire to have red qualia and then believe that the desire has been fulfilled, etc. Would it make sense in such circumstances to say that Fred_m doesn't have any red qualia, that he is just mistaken in his belief that he does, mistaken in believing that his desire to have red qualia has been fulfilled, etc.?[4]

If it does, then we must wonder about ourselves. How do we know that we

really have red qualia? Maybe we are all just like Fred_m, that is we only believe that we have red qualia, behave like we do, believe that our desire to have them has been fulfilled, etc., but in fact we never have any red qualia. That all of our lives, and all of our ancestors lives, have been spent with the mistaken belief that we at times have red qualia.

Someone might say that we know we have red qualia because we are made of neurons. But this is to miss the point. If it makes sense to think that Fred_m could be mistaken in such a way, then it makes sense to think that we could be mistaken in such a way also. Hence we would have no reason to think that things made of neurons (i.e., ourselves) have red qualia. But clearly it makes no sense to think that we act like we have red qualia, believe that we have them, etc., but that we are all mistaken and really never have any red qualia. Therefore, it makes no sense to think that any Fred_i could be mistaken in this way either.

The same can be said of intentional states. For example, what would it mean to say of Fred_m that he is mentally equivalent to Fred_m*, except that he does not remember his childhood? Would it make sense to think that he could believe he remembers his childhood, act as if he did, desire to remember it and then have that desire seem to be fulfilled, etc., but that in reality he has no memory of his childhood? (When I say that he has no memory of his childhood, I do not mean that he has a conscious state which is like a childhood memory, but that happens to not truly represent his childhood. What is at issue is his not having any conscious state which serves as the memory of his childhood, whether correct or not.)

Again, if Fred_m could be mistaken in this way, then for all we know, we are mistaken in this way also. But clearly it does not make any sense to think that we are, hence it does not make any sense to think that any Fred_i could be mistaken in this way either.

Hence, if one is going to maintain that the sequence's consciousness fades, then one must avoid any view that entails the senseless claim that the Fred_i's are mistaken in the way described above, and there is only one way to do this. One would have to suppose that each time some Fred_i loses a conscious ability, he also loses all of the beliefs, desires, etc., that go along with that ability. For example, if some Fred_i loses his ability to have red qualia, then he also loses his ability to believe he has red qualia, his ability to desire to have or not have red qualia, etc. In this way we would not have to consider any Fred_i to be mistaken in a way that does not make sense. For no Fred_i

would have the mistaken belief that he at times has red qualia, or the mistaken belief that his desire to have them has been fulfilled, etc.

However, taking the stance described in the above paragraph commits one to a proposition which is similar to (1). To see this, consider some member of the Fred_i sequence, call him Fred_k, who is such that he has lost either the ability to have red qualia or childhood memories, and the ability to believe that he has red qualia, or childhood memories. Fred_k will at times, since he is functionally equivalent to Fred_k*, say things like, "I am having a red quale", or he might tell a story about his childhood. For Fred_k to go on in such circumstances not believing that he has any red qualia, or not believing that he remembers his childhood, he will have to be such, that he does not form beliefs about what he says concerning the part of consciousness the he has lost. In fact, he will not be able to form beliefs about what a good deal of what other people say, or what he reads, for these things might contain assertions about his childhood, or that he is now having a red quale. And of course, he cannot disbelieve what he hears, reads, etc. about such matters. For if he did, he would be noticing the loss of a part of consciousness, and as discussed earlier, that is not possible.

Also, Fred_k cannot believe that he believes what he hears, reads, etc. about these things either, otherwise we will have created the same problem we are trying to get out of. For example, he will believe that he believes some story about his childhood, behave like he does, derive conclusions from the story, etc., but not really believe the story. In other words, he would have to be mistaken about what he believes in a way that, as discussed earlier, does not make any sense. And he would have to lose many other conscious states as well. For example, he would have to lose the belief that such and such a feeling was like one that he experienced during his childhood, for how could he notice such a thing unless he remembered his childhood? And he will have to lose the ability to understand much of his behavior. For example, suppose that he is asked to pick the red card out of a pile of different color cards. Since he is functionally equivalent to Fred_k* he will be able to do this correctly. But he will not be able to understand how he did it, since he will not believe that he had any red qualia that enabled him to pick out the red card. He can say, "I was able to pick out the red card because it looked red", but this is something he will not be able to believe (otherwise he would be believing that he has a red qualia) and hence it cannot serve as an explanation. Furthermore, he will not be able to notice that he doesn't understand how he

picked out the red card. For to notice, he would have to notice that there is no qualia peculiar to the card he calls red that enabled him to pick it out, and to notice this, would be to notice that he no longer has red qualia, and as discussed earlier, that is impossible. But how could he not notice the he doesn't understand how he picked out the red card? It seems as if he would have to lose his ability to know what he is doing. For if he knew that he was picking a particular kind of card from the pile, how could he not notice that his ability to do so is something that he doesn't understand, unless he has lost so much mental ability that he can hardly be recognized as a person?

This list could go on and on, but what is important to recognize, is that if one adopts the view that as the Fred_i sequence loses their conscious abilities they also lose the beliefs and desires that go along with the abilities that they have lost, then one will be committed to the view that as soon as even one small ability is lost (e.g., the ability to have red qualia) then a great deal of other abilities must be lost also.

To sum it up then, we see that one cannot claim that the Fred_i sequence's consciousness fades, while they notice that they are losing their conscious abilities (Sec. IV). And one cannot claim that they do not notice, without being committed to the view that each time some Fred_i loses a conscious ability, he also loses the beliefs and desires that go with that ability, and hence many other conscious abilities as well (Sec. V). Hence, the claim that Fred_n is unconscious entails the following proposition:

> (2) No matter in what order we remove neurons from the Fred_i sequence, there will always be some m, such that Fred_m is mentally equivalent to Fred_m*, and hence fully conscious, and such that Fred_{m+1} is completely unconscious, or lacking a great deal of Fred_m's conscious abilities.

VI

I will soon argue that (2) has the following properties: (i) it is very counter-intuitive; (ii) there is evidence other than intuition, that it is false, and (iii) there is no evidence, nor could there ever be any evidence, that it is true. Hence, I feel that Searle's anti-functionalist thesis should not be accepted, since we should be very hesitant to accept any thesis that entails any proposition which has the three properties listed above. What follows is the argument that (2) has the properties listed.

Someone might argue that (2) is not counterintuitive by saying something like the following, 'Look, if we removed neurons from your brain one by one, there will come a point at which with the removal of one neuron, you will be rendered unconscious. So why think it is counterintuitive to believe such a thing about the sequence of Freds?'

It is true that if you were to remove neurons from my brain one by one, I would at some point be rendered unconscious with the removal of a single neuron. But, this has nothing to do with what is at issue. As you removed neurons from my brain, my ability to realize conscious states, would slowly fade away, for many, if not all sequences of neuron removal. For example, I might first have my hearing impaired, then my sight, then lose some beliefs, etc. At some later point, the amount of ability to realize conscious states that was left, would be destroyed by the removal of a single neuron. But (2) does not allow the possibility of such fading in the Fred_i's; it says that all of them up to a certain point in the sequence are fully conscious and that the very next member of the sequence is completely unconscious or very close to it.

Moreover, it says that this is so not merely for *some* sequence of Fred_i's, but for every such sequence: in other words, it says it has to be the case that, for *any* sequence of neuron removal from Fred, he will go from *normal*, to completely without conscious states, or close to it, with the removal of one neuron; that no matter in what order the neurons are removed, there cannot be *any* loss in conscious abilities, until the removal of one neuron causes total, or near total loss. By near total loss, I mean that Fred would have to lose, for example, the ability to realize certain qualitative states; the ability to either believe or disbelieve many things that he says, reads, or hears; the ability to recognize that he has lost any ability to realize conscious states (he would be unable, for example, to notice that he was blind); the ability to make many inferences; the ability to have certain memories that he previously had; the ability to understand his own behavior (i.e., the ability to understand, even in a crude way, how he performs certain tasks) etc. That there is no order in which we could remove neurons, so that some Fred_i with all the normal conscious abilities (Fred_1's (i.e., original Fred) being in the sequence assures us that there is such a Fred_i) is caused to lose only a slight bit of conscious ability with the removal of a single neuron, is something which is very difficult to believe, and is such that it would require a great deal of explanation before it became the least bit plausible.

One might argue that there is a lot of redundancy in our neural network

and that hence we could have many neurons removed with no effect on our conscious states. But then, once enough circuitry was destroyed, there would be no backups left and a dramatic change would occur, even with the removal of one neuron, and hence we do have reason to believe that (2) is plausible.

To see that such an argument cannot make (2) plausible, let us consider a member of the sequence of Freds, Fred_j, who has all the redundancy in his neural circuitry destroyed, but no more than the redundancy, so that his conscious life is still completely normal; that is, he has all the ability to realize conscious states, that he had before any neurons were removed. (2) entails, that no matter which neuron we now remove from Fred_j, he will become completely unconscious, or very close to it. (2) entails this, because there must be some loss in conscious ability in Fred_j, no matter which neuron we remove, because all the redundancy in his neural circuitry has already been destroyed. And since Fred_j has all his normal conscious abilities, (2) says that the loss must be huge.

I find this counterintuitive, because it is hard to believe that *every* single neuron left in Fred_j, would be so vitally important, it seems that some neurons would be such, that removing one of them, would cause Fred_j to lose only a little bit of conscious ability. For example, that there is some neuron which is only vital to the mechanisms which allow vision, and such that its removal would cause Fred_j to go blind, while he was still able to have other conscious states (e.g. to taste things; hear sounds, to believe or disbelieve statements that he hears, reads or speaks, to notice that he is blind, etc.).

That (2) is true, becomes even more implausible when we realize that Fred_j is an arbitrary example of a Fred who has had all of the redundancy in his neural circuitry destroyed. That is to say, that there may be more than one proper subset of the neurons of Fred_1 (i.e., of the original set of neurons in Fred before any operations were performed) which would constitute a person with all, but only, the redundancy destroyed. Since Fred_j is an arbitrary member of this class of Freds, (2) entails that for *any* of these Freds, no matter which neuron we remove from one of them, that one will suffer a huge (maybe total) loss in conscious abilities. To put it another way, (2) denies that their is *any* proper subset of the neurons of Fred_1 which constitute a person who is slightly impaired (e.g. blind, or unable to understand Godel's theorem, but otherwise normal) but asserts that some subsets of the neurons of Fred_1 constitute organisms which are so horribly impaired that they would not even qualify as persons, and that other subsets of those

neurons constitute persons that are completely normal. And that some subsets should constitute completely normal people, while others are hugely impaired, but that no subset can constitute a less radically impaired person, is extremely difficult to believe and would require a great deal of explanation before it became acceptable.

Also I claimed that there is evidence other than intuition that (2) is false. The kind evidence I have in mind is inductive evidence based on what happens in some cases of actual neuron removal from people's brains (e.g., the removal of some tumors, or the destruction of some neurons by a mild stroke).

In many cases the loss of conscious abilities is slight, and may even be hard to detect by anyone but an expert, and nothing like the dramatic changes that (2) postulates occurs; or consider someone who is caused to go blind due to brain damage, but is otherwise normal. Due to such cases, we know that some sequences of neuron removal cause people who are fully conscious, to lose only a slight bit of conscious ability, something that (2) denies.[5] (2) says that Fred *must* go from *normal* to unconscious, or very close to it, *regardless* of which neurons are removed in which order, it does not allow the possibility of Fred going from normal to slightly impaired, for any sequence of neuron removal. (2) does not even allow that there is any way that Fred can go from normal to slightly impaired, to slightly more impaired and then all of a sudden have a big change and blank out. It says that Fred must go completely unchanged in his ability to realize conscious states throughout the operations (regardless of which sequence we use to remove the neurons) and then lose all or almost all conscious abilities, with the removal of one neuron.

Of course in the case of the Fred_i sequence, things are a little different than with actual cases, since humunculi are added, and only one neuron is removed at a time (whereas in actual cases many neurons are removed, or destroyed at once). But, these differences should not block our ability to make an inductive inference because, in fact, the addition of humunculi and the fact that neurons are removed one at a time, should make us feel even stronger that the Fred_i sequenced would not lose huge chunks of consciousness at once for every sequence of neuron removal. This is because the humunculi at least take over the functional role of the neurons, whereas we know that in some actual cases even if nothing replaces the removed neurons, and many of them are removed at once, the patients still lose only a slight amount of conscious ability.

Another reason that one should count actual cases as a basis for an inductive inference about (2), is that if actual cases were different, they would in fact justify the conclusion that (2) is true. That is, if we found that all ways of removing neurons made it the case that patients only lost conscious abilities if they lost all of them, or a great deal of them, at once, then we would be justified in believing (2). Therefore actual cases serve as a real test, in that there are possible results which would lend evidence in support of (2), as well as possible results that would constitute evidence against (2).

So now I have argued that there is evidence against (2), and hence evidence against Searle's view since it entails (2). But don't actual cases of neuron removal also support the conclusion that Fred_n is unconscious, since actual patients do lose conscious abilities when their neurons are removed? If this were the case, we might conclude that Searle is correct even though (2) is fase, by arguing that there must be something wrong with my argument that Searle's view entails (2). In other words, one might argue that actual neuron removals provide evidence that (2) is false as well as evidence that Fred_n is unconscious, by giving evidence in support of the view that the Fred_i sequence's consciousness *fades*.

However, actual cases do not provide any evidence that Fred_n is unconscious. This is because if one tries to inductively infer from actual cases that the Fred_i sequence's consciousness fades, then the differences between the Fred_i sequence and actual cases become relevant, and the inference is blocked. The reason is that with the Fred_i sequence homunculi replace the removed neurons, so functional organization is preserved, whereas in actual cases of neuron removal functional organization is not preserved. Hence, there is an explanation as to why the Fred_i sequence would not lose consciousness, even though actual cases of neuron removal do result in losses of consciousness. To claim that actual cases of neuron removal are a basis from which to infer that the Fred_i sequence's consciousness fades, would be to do nothing more than beg the question. For one would have to already know that what is necessary for consciousness is the having of neurons, and that functional equivalence to a human is not sufficient.

I also claimed that there is no evidence, nor could there ever be any evidence that (2) is true. To see this, consider the following scenario:

Imagine that we have Fred_m standing before us. We then make a very tiny change (the removal of one neuron and the insertion of a homunculus) to transform Fred_m into Fred_{m+1}. Since by stipulation Fred_m is fully

conscious, what could possibly count as evidence that Fred_{m+1} is unconscious, or lacking a large number of Fred_m's conscious abilities?

Fred_{m+1} will behave just like Fred_{m+1}* (who is a normal human) hence it can't be his behavior that gives us any reason to believe that he has lost any conscious abilities. We can't appeal to the fact that Fred_{m+1} is composed of slightly different matter than Fred_m (i.e., he has one more homunculus and slightly less natural brain material) because whether or not a change in material that preserves functional organization makes a difference to an organisms conscious states, is precisely what is at issue. The only thing left would be to claim that there is some intuitive reason to believe that Fred_{m+1} is unconscious, or lacking a large number of Fred_m's conscious abilities. However, the difference between Fred_{m+1} and Fred_m is so slight, that an appeal to intuition is useless. How could it be intuitively appealing to think that Fred_m is fully conscious, but that Fred_{m+1} is unconscious or lacking a large number of Fred_m's conscious abilities? Would we glance within the skull of Fred_{m+1} notice that there is one less neuron than in Fred_m and exlciam, 'Aha! it's obvious that Fred_{m+1} is a mindless, or near mindless thing?' It is important to note here that thinking that Fred_{m+1} is not unconscious, but only lacking a great deal of Fred_m's conscious abilities, is not any more intuitively appealing than thinking that Fred_{m+1} is unconscious. Either way, the change from Fred_m to Fred_{m+1} would be too great. Intuition simply does not lead us to believe that as we remove a single neuron from a creature with all the normal conscious abilities, we simultaneously destroy his ability to be in any one of a large number of conscious states. Especially when we consider that it must be a set of *quite different* conscious abilities that he loses (e.g., the ability to have red qualia, the ability to either believe or disbelieve what he sometimes says, etc.). Hence, intuition simply does not offer any evidence that (2) is true.

CONCLUSION

This paper has, I hope, supported the conclusion that functional equivalence to a human at a very fine level, is a sufficient condition for an organism to have conscious states. It has done this by arguing that the contrary position entails a proposition (i.e., (2)) that we have good reason to believe to be false. The fine level of functional organization alluded to, involves reproducing the

functional role of each neuron in a normal human brain. Call this circuit functional equivalence.

However functional theories are more attractive, if they do not require as a necessary condition for conscious states, anything as fine grained as circuit functional equivalence. So one thing that would be worth doing would be to show that functional equivalence at some coarser level is sufficient for having conscious states. And I think that this paper can help do this by weakening one's beliefs to the contrary. (By a coarser level, I mean any level of description X, such that circuit functional equivalence entails equivalence at the X level, but equivalence at the X level does not entail circuit functional equivalence.)

To be more specific, consider some of the arguments of Block, Searle and others to the contrary ([1] and [2]). In these arguments, creatues are described which are, at some level coarser than the circuit functional, functionally equivalent to a human, but which are, according to these authors, such that they lack conscious states.

However, there seem to be at least two reasons why one might believe that these creatures are not conscious. One reason might be based on the belief that the functional equivalence that the creatures share with a human, is not at the relevant level of organization. The other reason, and I believe the dominant reason, is that one feels at first glance, that they are just not made of the right kind of stuff (e.g., they are made of homunculi).

This paper then, should help to weaken intuitions that are based on what the organisms are made of. I say this because I think it has been shown that what is important is not what an organism is made of, but rather functional organization *at some level*. Hence, if one wishes to maintain that such organisms do not have conscious states, then one is going to have to do this on the grounds that the functional equivalence that they share with a human is not at the relevant level, and not on the grounds that they are not made of the proper material.

BIBLIOGRAPHY

[1] Block, N.: 1978, 'Troubles with functionalism', in Savage, C. W., ed., Perception and Cognition. Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science, Vol. 9 (University of Minnesota Press, Minneapolis), pp. 261–325.

[2] Searle, J. R.: 1980, 'Minds, brains, and programs', The Behavioral and Brain Sciences 3, pp. 417–457.

NOTES

* I am especially indebted to Hartry Field and Brian Loar for many discussions on earlier drafts of this paper, and again to Loar for many discussions on the philosophy of mind in general. I am also indebted, for their valuable comments, to: Janet Fleetwood, Janet Levin, John L. Pollock, Stephen R. Schiffer, Thomas Uebel and Richard Warner.

[1] Searle (*op. cit.* p. 424) tells us that "...*only* a machine could think, and indeed only very special kinds of machines, namely brains *and machines that had the same causal power as brains*" (italics mine). And a few sentences later he says: "Whatever else intentionality is, it is a biological phenomenon, and it is likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomenon." I am not sure, but I would imagine that Searle is claiming here that organisms made of things other than neurons might be able to have consciousness, so long as their components have the right causal powers. Unfortunately, he does not tell us what the components must have in common with neurons in order to have the right causal powers, except that it has something to do with biochemistry. At any rate, all that is important for this paper, is that Searle does not consider functional organization alone to be sufficient for having conscious states, and that he would not consider things made of homunculi as candidates for consciousness, regardless of their functional organization.

[2] If one is worried that the operations themselves might disturb the states of the neurons in some $Fred_i$ so that they are not in the same state as their counterparts, simply imagine that the $Fred_i$s have the states of their brains 'frozen' during the operations, then 'unfrozen' afterward, and that the $Fred_i$*s have their brains 'frozen' and 'unfrozen' also, but don't have any operations in between. And if one is worried that quantum indeterminancies might cause some $Fred_i$ to not have his neurons in the same state as their counterparts, even though each $Fred_i$* goes through the same circumstances, other than neuron removal, that $Fred_i$ goes through, just remember that in spite of quantum indeterminancies, there is still a non zero probability that the $Fred_i$ and $Fred_i$* sequences remain synchronized. Hence just have the $Fred_i$* sequence be by stipulation, such that the two sequences bear the proper relationship. Also I would like to note that my introduction of the $Fred_i$* sequence is only a heuristic device, that the point of the paper could be made without assuming such a sequence.

[3] One might argue that in such cases the homunculi do not play a role in tokening the mental state of noticing that some part of consciousness has been lost, but rather that it is tokened entirely within the remaining neurons, even though they are in the same state as their counterparts. And that the neurons of some $Fred_i$ can token this mental state entirely within themselves, even though it is not tokened in $Fred_i$*, and even though they are in the same state as their counterparts, because the mental state is tokened simply by having less neurons than $Fred_i$*. However, not only is such a claim wildly ad hoc, it is almost certainly false as well. For if I were to *notice* that I was blind (as opposed to caused to become blind) simply because some neurons were removed, while the remaining neurons did not change state, I would not even be able to utter the sentence, 'I am blind', for there would be no change in any neurons that could trigger its utterance.

[4] Sydney Shoemaker has made a similar point, in that he has argued that an organism which has all of the intentional states of a normal human being (i.e., a human being which has qualia) cannot be lacking qualia. He leaves it open as to whether or not the qualia would have to be the same for two organisms that had the same type of intentional states however (e.g., the spectrum of one may be the inversion of the other's). See for example his paper, 'Functionalism and qualia', from, Philosophical Studies, Vol. 27, 1975, pp. 291–315.

[5] One might argue that I am begging the question here on the grounds that, for all we

know, such patients are completely unconscious, and only similar to normal humans in functional respects. The only thing that I have to say to such an objection, is to point out that anyone who raised it, would be embracing a very radical scepticism in order to maintain that Fred_n is unconscious and would simply be reducing potential problems with functionalism, to the problem of other minds.

Department of Philosophy,
University of Southern California,
Los Angeles, CA 90007,
U.S.A.