Free Will as a Problem in Neurobiology[1]

JOHN R. SEARLE

I. The Problem of Free Will

The persistence of the traditional free will problem in philosophy seems to me something of a scandal. After all these centuries of writing about free will, it does not seem to me that we have made very much progress. Is there some conceptual problem that we are unable to overcome? Is there some fact that we have simply ignored? Why is it that we have made so little advance over our philosophical ancestors?

Typically, when we encounter one of these problems that seems insoluble it has a certain logical form. On the one hand we have a belief or a set of beliefs that we feel we really cannot give up, but on the other hand we have another belief or set of beliefs that is inconsistent with the first set, and seems just as compelling as the first set. So, for example, in the old mind-body problem we have the belief that the world consists entirely of material particles in fields of force, but at the same time the world seems to contain consciousness, an immaterial phenomenon; and we cannot see how to put the immaterial together with the material into a coherent picture of the universe. In the old problem of sceptical epistemology, it seems, on the one hand, according to common sense, that we do have certain knowledge of many things in the world, and yet, on the other hand, if we really have such knowledge, we ought to be able to give a decisive answer to the sceptical arguments, such as 'How do we know we are not dreaming, are not a brain in a vat, are not being deceived by evil demons, etc.?' But we do not know how to give a conclusive answer to these sceptical challenges. In the case of free will the problem is that we think explanations of natural

[1] This article is an extension of some of the ideas presented in my lecture to the Royal Institute of Philosophy, in February 2001. That lecture was based on an earlier article in *The Journal of Consciousness Studies*, 'Consciousness, Free Action and the Brain', volume 10, number 10, October 2000. Some of the arguments in the early part of this article are developed in more detail in my forthcoming book *Rationality in Action*, MIT Press.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

phenomena should be completely deterministic. The explanation of the Loma Prieta earthquake, for example, does not explain why it just happened to occur, it explains why it *had* to occur. Given the forces operating on the tectonic plates, there was no other possibility. But at the same time, when it comes to explaining a certain class of human behaviour, it seems that we typically have the experience of acting 'freely' or 'voluntarily' in a sense of these words that makes it impossible to have deterministic explanations. For example, it seems that when I voted for a particular candidate, and did so for a certain reason; well, all the same, I could have voted for the other candidate, all other conditions remaining the same. Given the causes operating on me, I did not *have* to vote for that candidate. So when I cite the reason as an explanation of my action I am not citing causally sufficient conditions. So we seem to have a contradiction. On the one hand we have the experience of freedom, and on the other hand we find it very hard to give up the view that because every event has a cause, and human actions are events, they must have sufficient causal explanations as much as earthquakes or rain storms.

When we at last overcome one of these intractable problems it often happens that we do so by showing that we had made a false presupposition. In the case of the mind-body problem, we had, I believe, a false presupposition in the very terminology in which we stated the problem. The terminology of mental and physical, of materialism and dualism, of spirit and flesh, contains a false presupposition that these must name mutually exclusive categories of reality—that our conscious states qua subjective, private, qualitative, etc, cannot be ordinary physical, biological features of our brain. Once we overcome that presupposition, the presupposition that the mental and the physical naively construed are mutually exclusive, then it seems to me we have a solution to the traditional mind-body problem. And here it is: All of our mental states are caused by neurobiological processes in the brain, and they are themselves realized in the brain as its higher level or system features. So, for example, if you have a pain, your pain is caused by sequences of neuron firings, and the actual realization of the pain experience is in the brain.[2]

[2] I am assuming for the sake of this article that the right functional level for explaining mental phenomena is the level of neurons. It might turn out to be some other level—micro-tubules, synapses, neuronal maps, whole clouds of neurons, etc.—but for the purposes of this article it does not matter what the right neurobiological explanatory level is, only that there is a neurobiological explanatory level.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

The solution to the philosophical mind-body problem seems to me not very difficult. However, the philosophical solution kicks the problem upstairs to neurobiology, where it leaves us with a very difficult neurobiological problem. How exactly does the brain do it, and how exactly are conscious states realized in the brain? What exactly are the neuronal processes that cause our conscious experiences, and how exactly are these conscious experiences realized in brain structures?

Perhaps we can make a similar transformation of the problem of free will. Perhaps if we analyse the problem sufficiently, and remove various philosophical confusions, we can see that the remaining problem is essentially a problem about how the brain works. In order to work toward that objective I need first to clarify a number of philosophical issues.

Let us begin by asking why we find the conviction of our own free will so difficult to abandon. I believe that this conviction arises from some pervasive features of conscious experience. If you consider ordinary conscious activities such as ordering a beer in a pub, or watching a movie, or trying to do your income tax, you discover that there is a striking difference between the passive character of perceptual consciousness, and the active character of what we might call 'volitional consciousness'. For example, if I am standing in a park looking at a tree, there is a sense in which it is not up to me what I experience. It is up to how the world is and how my perceptual apparatus is. But if I decide to walk away or raise my arm or scratch my head, then I find a feature of my experiences of free, voluntary actions that was not present in my perceptions. The feature is that I do not sense the antecedent causes of my action in the form of reasons, such as beliefs and desires, as setting causally sufficient conditions for the action; and, which is another way of saying the same thing, I sense alternative courses of action open to me.

You see this strikingly if you consider cases of rational decision making. I recently had to decide which candidate to vote for in a presidential election. Suppose for the sake of argument, that I voted for George W. Bush. I had certain reasons for voting for Bush, and certain other reasons for not voting for Bush. But, interestingly, when I chose to vote for Bush on the basis of some of those reasons and not others, and later when I actually cast a vote for Bush in a voting booth, I did not sense the antecedent causes of my action as setting causally sufficient conditions. I did not sense the reasons for making the decision as causally sufficient to force the decision, and I did not sense the decision itself as causally sufficient to force the action. In typical cases of deliberating and acting, there is, in short,

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

a gap, or a series of gaps between the causes of each stage in the processes of deliberating, deciding and acting, and the subsequent stages. If we probe more deeply we can see that the gap can be divided into different sorts of segments. There is a gap between the reasons for the decision and the making of the decision. There is a gap between the decision and the onset of the action, and for any extended action, such as when I am trying to learn German or to swim the English Channel, there is a gap between the onset of the action and its continuation to completion. In this respect, voluntary actions are quite different from perceptions. There is indeed a voluntaristic element in perception. I can, for example, choose to see the ambiguous figure either as a duck or a rabbit; but for the most part my perceptual experiences are causally fixed. That is why we have a problem of the freedom of the will, but we do not have a problem of the freedom of perception. The gap, as I have described it, is a feature of our conscious, voluntary activities. At each stage, the conscious states are not experienced as sufficient to compel the next conscious state. There is thus only one continuous experience of the gap but we can divide it into three different sorts of manifestations, as I did above. The gap is between one conscious state and the next, not between conscious states and bodily movements or between physical stimuli and conscious states.

This experience of free will is very compelling, and even those of us who think it is an illusion find that we cannot in practice act on the presupposition that it is an illusion. On the contrary, we have to act on the presupposition of freedom. Imagine that you are in a restaurant and you are given a choice between veal and pork, and you have to make up your mind. You cannot refuse to exercise free will in such a case, because the refusal itself is only intelligible to you as a refusal, if you take it as an exercise of free will. So if you say to the waiter, 'Look, I am a determinist—que sēra sēra, I'll just wait and see what I order', that refusal to exercise free will is only intelligible to you as one of your actions if you take it to be an exercise of your free will. Kant pointed this out a long time ago. We cannot think away our free will. The conscious experiences of the gap give us the conviction of human freedom.

If we now turn to the opposing view and ask why we are so convinced of determinism, the arguments for determinism seem just as compelling as the arguments for free will. A basic feature of our relation to the world is that we find the world causally ordered. Natural phenomena in the world have causal explanations, and those causal explanations state causally sufficient conditions. Customarily, in philosophy, we put this point by saying that that

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

every event has a cause. That formulation is, of course, much too crude to capture the complexity of the idea of causation that we are working with. But the basic idea is clear enough. In our dealings with nature we assume that everything that happens, occurs as a result of antecedently sufficient causal conditions. And when we give an explanation by citing a cause, we assume that the cause we cite, *together with the rest of the context*, was sufficient to bring about the event we are explaining. In my earlier example of the earthquake, we assume that the event did not just happen to occur, in that situation it had to occur. In that context the causes were sufficient to determine the event.

An interesting change occurred in the early decades of the 20th century. At the most fundamental level of physics, nature turns out not to be in that way deterministic. We have come to accept at a quantum mechanical level explanations that are not deterministic. However, so far quantum indeterminism gives us no help with the free will problem because that indeterminism introduces randomness into the basic structure of the universe, and the hypothesis that some of our acts occur freely is not at all the same as the hypothesis that some of our acts occur at random. I will have more to say about this issue later.

There are a number of accounts that seem to explain consciousness and even free will in terms of quantum mechanics. I have never seen anything that was remotely convincing, but it is important for this discussion that we remember that as far as our actual theories of the universe are concerned, at the most fundamental level we have come to think that it is possible to have explanations of natural phenomena that are not deterministic. And that possibility will be important when we later discuss the problem of free will as a neurobiological problem.

It is important to emphasize that the problem of free will, as I have stated it, is a problem about a certain kind of human consciousness. Without the conscious experience of the gap, that is, without the conscious experience of the distinctive features of free, voluntary, rational actions, there would be no problem of free will. We have the conviction of our own free will because of certain features of our consciousness. The question is: Granted that we have the experience of freedom, is that experience valid or is it illusory? Does that experience correspond to something in reality beyond the experience itself? We have to assume that there are causal antecedents to our actions. The question is: Are those causal antecedents in every case sufficient to determine the action, or are there some cases where they are not sufficient, and if so how do we account for those cases?

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

Let us take stock of where we are. On the one hand we have the experience of freedom, which, as I have described it, is the experience of the gap. The gap between the antecedent causes of our free, voluntary decisions and actions, and the actual making of those decisions and the performance of those actions. On the other hand we have the presupposition, or the assumption, that nature is a matter of events occurring according to causally sufficient conditions, and we find it difficult to suppose that we could explain any phenomena without appealing to causally sufficient conditions.

For the purposes of the discussion that follows, I am going to assume that the experiences of the gap are psychologically valid. That is, I am going to assume that for many voluntary, free, rational human actions, the purely *psychological* antecedents of the action are not causally sufficient to determine the action. This occurred, for example, when I selected a candidate to vote for in the last American presidential election. I realize that a lot of people think that psychological determinism is true, and I have certainly not given a decisive refutation of it. Nonetheless, it seems to me we find the psychological experience of freedom so compelling that it would be absolutely astounding if it turned out that at the psychological level it was a massive illusion, that all of our behaviour was psychologically compulsive. There are arguments against psychological determinism, but I am not going to present them in this article. I am going to assume that psychological determinism is false, and that the real problem of determinism is not at the psychological level, but at a more fundamental neurobiological level.

Furthermore, there are several famous issues about free will that I will not discuss, and I mention them here only to set them on one side. I will have nothing to say about compatibilism, the view that free will and determinism are really consistent with each other. On the definitions of these terms that I am using, determinism and free will are not compatible. The thesis of determinism asserts that all actions are preceded by sufficient causal conditions that determine them. The thesis of free will asserts that some actions are not preceded by sufficient causal conditions. Free will so defined is the negation of determinism. No doubt there is a sense of these words where free will is compatible with determinism (When for example people march in the streets carrying signs that say, 'Freedom Now' they are presumably not interested in physical or neurobiological laws), but that is not the sense of these terms that concerns me. I will also have nothing to say about moral responsibility. Perhaps there is some interesting connection between the problem of free

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

will and the problem of moral responsibility, but if so I will have nothing to say about it in this article.

II. How Consciousness Can Move Bodies

Because the problem of free will is a problem about the causal facts concerning certain sorts of consciousness, we need to explain how consciousness in general can function causally to move our bodies. How can a state of human consciousness cause a bodily movement? One of the most common experiences in our lives is that of moving our bodies by our conscious efforts. For example, I now intentionally raise my arm, a conscious effort on my part, and lo and behold, the arm goes up. What could be more common? The fact that we find such a banal occurrence philosophically puzzling suggests that we are making a mistake. The mistake derives from our inherited commitment to the the old Cartesian categories of the mental and the physical. Consciousness seems too weightless, ethereal and immaterial ever to move even one of our limbs. But as I tried to explain earlier, consciousness is a higher-level biological feature of the brain. To see how the higher level feature of consciousness has physical effects, consider how higher level features work in the case of metaphysically less puzzling phenomena.

To illustrate the relationships between higher-level or system features, on the one hand, and micro-level phenomena, on the other, I want to borrow an example from Roger Sperry. Consider a wheel rolling down hill. The wheel is entirely made of molecules. The behaviour of the molecules causes the higher level, or system feature, of solidity. Notice that the solidity affects the behaviour of the individual molecules. The trajectory of each molecule is affected by the behaviour of the entire solid wheel. But of course there is nothing there but molecules. The wheel consists entirely of molecules. So when we say the solidity functions causally in the behaviour of the wheel and in the behaviour of the individual molecules that compose the wheel, we are not saying that the solidity is something *in addition* to the molecules; rather it is just the *condition* that the molecules are in. But the feature of solidity is nonetheless a real feature, and it has real causal effects.

Of course there are many disanalogies between the relation of solidity to molecular behaviour, on one hand, and the relation of consciousness to neuronal behaviour, on the other. I will explain some of them later, but now I want to focus on the feature that we have just explored, and suggest that it applies to the relation of con-

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

sciousness and the brain. The consciousness of the brain can have effects at the neuronal level even though there is nothing in the brain except neurons (with glial cells, neuro-transmitters, blood flow, and all the rest). And just as the behaviour of the molecules is causally constitutive of solidity, so the behaviour of the neurons is causally constitutive of consciousness. When we say that consciousness can move my body, what we are saying is that the neuronal structures move my body, but they move my body in the way they do because of the conscious state they are in. Consciousness is a feature of the brain in a way that solidity is a feature of the wheel.

We are reluctant to think of consciousness as just a biological feature of the brain, in part because of our dualistic tradition, but also because we tend to suppose that if consciousness is *irreducible* to neuronal behaviour then it must be something extra, something 'over and above' neuronal behaviour. And of course consciousness, unlike solidity, is not ontologically reducible to physical micro-structures. This is not because it is some extra thing; rather it is because consciousness has a first-person, or subjective, ontology, and is thus not reducible to anything that has a third-person, or objective, ontology.[3]

In this brief discussion I have tried to explain how consciousness can have 'physical' causal consequences, and why there is nothing mysterious about that fact. My conscious intention-in-action causes my arm to go up. But of course, my conscious intention-in-action is a feature of my brain system, and as such at the level of the neurons it is constituted entirely by neuronal behaviour. There is no ontological reductionism in this account, because at no point are we denying that consciousness has an irreducible first-person ontology. But there is a causal reduction. Consciousness has no causal powers beyond the powers of the neuronal (and other neurobiological) structures.

III. The Structure of Rational Explanation

I said that the problem of free will is a problem about certain sorts of consciousness. If we look at the sorts of explanations that we give for actions which are manifestations of the gap, that is, actions which are expressions of our experience of free, rational decision-making, we find that the experience of free will is reflected in the

[3] For further discussion, see John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), especially Chapter 5.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

logical structure of action explanations. In a word, because of the gap, explanations that appeal to our rational decision-making processes are not deterministic in form in a way that typical explanations of natural phenomena are deterministic in form. To see how this is so, contrast the following three explanations:

1. I punched a hole in the ballot paper because I wanted to vote for Bush.
2. I got a bad headache because I wanted to vote for Bush.
3. The glass fell to the floor and broke because I accidentally knocked it off the table.

Of these examples, 1 and 2 look very similar in their syntactical structure, and they appear to be different from 3. I will argue, however, that 2 and 3 are the same in their underlying logical structure, and they both differ in this respect from 1. 3 is a standard causal explanation which states that one event or state caused another event or state. The logical form of 3 is simply: A caused B. But the form of 1 is quite different. We do not take statements of form 1 as implying that the event described by the clause before 'because' had to occur, given the occurrence of the event described after the 'because' and the rest of the context. We do not take 1 as implying that my desire to vote for Bush was such as to force me to punch a hole in the ballot paper, that given my psychological state at the time, I could not have done otherwise. Explanations of this form may on occasion cite causally sufficient conditions, but the form of the explanation does not require such conditions. If we compare 1 and 3, with 2 it seems to me that 2, like 3, is a matter of causally sufficient conditions. The form of 2, like 3, is simply: A caused B. In that context, the state of my desiring to vote for Bush was causally sufficient for the event of my getting a headache.

But this feature of rational explanation leaves us with a puzzle, almost a contradiction. It seems that if the explanation does not give causally sufficient conditions, it cannot really explain anything, because it does not answer the question why one event occurred as opposed to another event, which was also causally possible given exactly the same antecedent conditions. I think answering that question is an important part of the discussion of free will, so I want to spend a little bit of time on it.

As a matter of their logical structure, explanations of voluntary human actions in terms of reasons are different from ordinary causal explanations. The logical form of ordinary causal explanations is simply that event A caused event B. Relative to specific contexts, we typically take such explanations as adequate because

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

we assume that in that context, event A was causally sufficient for event B. Given the rest of the context, if A occurred then B had to occur. But the form of the explanation of human behaviour, where we say that a certain person performed act A by acting on reason R, has a different logical structure. It is not of the form 'A caused B'. I think you only understand that structure if you realize that it requires the postulation of a self or an ego. The logical form of the statement 'Agent S performed Act A because of reason R' is not of the form 'A caused B', it is of the form 'A self S performed action A, and in the performance of A, S acted on reason R'. The logical form, in short, of rational explanation is quite different from standard causal explanations. The form of the explanation is not to give causally sufficient conditions, but to cite the reason that the agent acted on.

But if that is right, then we have a peculiar result. It seems that rational action explanations require us to postulate the existence of an irreducible self, a rational agent, in addition to the sequence of events. Indeed, if we make explicit two further assumptions to those we have already been making, I think we can derive the existence of the self.

> Assumption 1: Explanations in terms of reasons do not typically cite causally sufficient conditions

and

> Assumption 2: Such explanations can be adequate explanations of actions.

How do I know that Assumption 2 is true? How do I know such explanations can be and often are adequate? Because in my own case I often know exactly what reasons I had for performing an action and I know that an explanation that cites those reasons is adequate, because I know that in acting I *acted on* those reasons and on those reasons alone. Of course we have to allow that there are all kinds of problems about the unconscious, self-deception, and all the rest of the unknown and unacknowledged reasons for action. But in the ideal case where I consciously act on a reason and am consciously aware of acting on a reason, the specification of the reason as the explanation of my action is perfectly adequate.

We have already been making a third assumption,

> Assumption 3: Adequate causal explanations cite conditions that, relative to the context, are causally sufficient.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

And this assumption just makes explicit the principle that if a causal statement is to explain an event, then the statement of the cause must cite a condition that in that particular context was sufficient to bring about the event to be explained. But from Assumptions 1 and 3 we can derive:

> Conclusion 1: Construed as ordinary causal explanations, reason explanations are inadequate.

If we were to assume that reason explanations are ordinary causal explanations we would have a straight contradiction. To avoid the contradiction we have to conclude:

> Conclusion 2: Reason explanations are not ordinary causal explanations. Though they have a causal component, their form is not, A caused B.

That leaves us with a problem. How are we to explain the adequacy of these explanations if they have a causal component, and, nonetheless, are not standard causal explanations? I think the answer is not hard to find. The explanation does not give a sufficient cause of an event, rather it gives a specification of how a conscious rational self acted on a reason, how an agent made a reason effective by freely acting on it. But when spelled out, the logical form of such explanations requires that we postulate an irreducible, non-Humean self. Thus:

> Conclusion 3: Reason explanations are adequate because they explain why a self acted in a certain way. They explain why a rational self acting in the gap, acted one way rather than another, by specifying the reason that the self acted on.

There are thus two avenues to the gap, an experiential and a linguistic. We experience ourselves acting freely in the gap, and this experience is reflected in the logical structure of explanations that we give for our actions. We experience ourselves acting as rational agents, and our linguistic practice of giving explanations reflects the gap (because the explanations do not cite causally sufficient conditions); and for their intelligibility these explanations require that we recognize that there must be an entity—a rational agent, a self, or an ego—that acts in the gap (because a Humean bundle of perceptions would not be enough to account for the adequacy of the explanations). The necessity of assuming the operation of an irreducible, non-Humean, self is a feature both of our actual experience of voluntary action and the practice that we have of explaining our voluntary actions by giving reasons.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

Of course such explanations, like all explanations, allow for further questions about why those reasons were effective and not other reasons. That is, if I say that I voted for Bush because I wanted an improvement in the educational system, there is a further question, why did I want that improvement? And why was that reason more compelling to me than other reasons? I agree that such a demand for explanations can always be continued, but that is true of any explanation. Explanations, as Wittgenstein reminded us, have to stop somewhere, and there is nothing inadequate about saying that I voted for Bush because I wanted an improvement in the educational system. It does not show that my answer is inadequate to show that it admits of further questions.

I am here summarizing briefly a complex argument that I have spelled out in more detail in Chapter 3 of *Rationality in Action* (MIT Press, forthcoming). But the bare bones of the argument can be conveyed even in this brief summary: We have the first-person conscious experience of acting on reasons. We state these reasons for action in the form of explanations. The explanations are obviously quite adequate because we know in our own case that, in their ideal form, nothing further is required. But they cannot be adequate if they are treated as ordinary causal explanations because they do not pass the causal sufficiency test. They are not deterministic in their logical form as stated, and they are not deterministic in their interpretation. How can we account for these facts? To account for these explanations we must see that they are not of the form A caused B. They are of the form, a rational self S performed act A, and in performing A, S acted on reason R. But that formulation requires the postulation of a self.

Conclusion 3 does not follow deductively from the assumptions. The argument as presented is a 'transcendental' argument, in one of Kant's senses of that term. Assume such and such facts and ask what are the conditions of possibility of these facts. I am claiming that the condition of possibility of the adequacy of rational explanations is the existence of an irreducible self, a rational agent, capable of acting on reasons.

Let us take stock again of where we are. We saw, first, that the problem of free will arises because of a special feature of a certain type of human consciousness, and we saw, second, that in order to explain our apparently free behaviour, we have to postulate an irreducible notion of the self. This, by the way, is typical of philosophy —in order to solve one problem you have to solve a bunch of others, but so far, I seem to have given you three problems for one. We started with the problem of free will, and we now have the problems

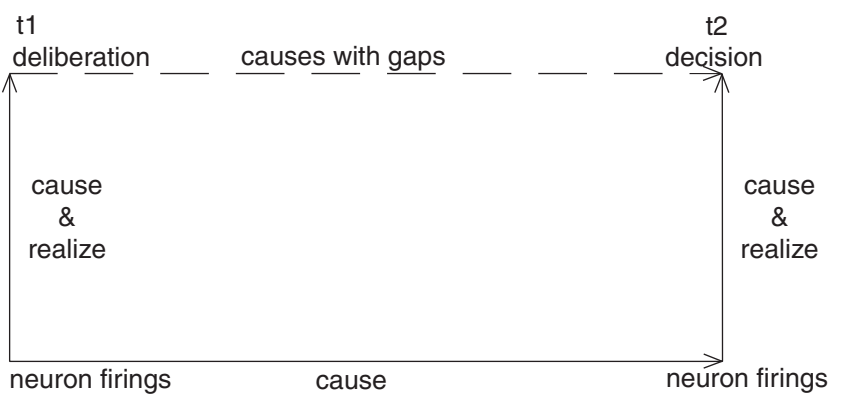
https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

of free will, of consciousness, and of the self, and they all seem to hang together.

IV. Free Will and the Brain

I now turn to the main question of this article: How could we treat the problem of free will as a neurobiological problem? And the assumption that I am making is that if free will is a genuine feature of the world and not merely an illusion, then it must have a neurobiological reality; there must be some feature of the brain that realizes free will. I said earlier that consciousness is a higher level, or system, feature of the brain caused by the behaviour of lower-level elements, such as neurons and synapses. But if that is so, what would the behaviour of the neurons and the synapses have to be like if the conscious experience of free will were to be neurobiologically real?

I have said that the philosophical solution to the traditional mind-body problem is to point out that all of our conscious states are higher-level or systemic features of the brain, while being at the same time caused by lower-level micro-processes in the brain. At the system level we have consciousness, intentionality, decisions, and intentions. At the micro level we have neurons, synapses, and neurotransmitters. The features of the system level are caused by the behaviour of the micro-level elements, and are realized in the system composed of the micro-level elements. In the past I have described the set of causal relations between decision making and acting in terms of a parallelogram where at the top level we have decisions leading to intentions-in-action, and at the bottom level we have neuron firings causing more neuron firings. Such a picture gives us a parallelogram that looks like this:



https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

The question is, if we suppose there is a gap at the top level in the case of rational decision-making, how might that gap be reflected at the neurobiological level? There are, after all, no gaps in the brain. In order to explore alternative hypotheses we need to consider an example.

A famous, if mythological, example is the judgment of Paris. Confronted with three beautiful Goddesses, Hera, Aphrodite and Pallas Athena, Paris was required to deliberate and reach a decision as to which should receive the golden apple, inscribed 'For the fairest'. He was not to decide this by appraising their beauty but by choosing among the bribes each offered. Aphrodite promised that he would possess the most beautiful woman in the world, Athena that he would lead the Trojans to victory over the Greeks, and Hera offered to make him ruler of Europe and Asia. It is important that he has to make a decision as a result of deliberation. He does not just spontaneously react. We also assume that he was operating in the gap: He consciously felt a range of choices open to him; and his decision was not forced by lust, rage or obsession. He made a free decision after deliberation.

We can suppose there was an instant when the period of reflection began, call it t1, and that it lasted until he finally handed the apple to Aphrodite at t2. In this example we will stipulate that there was no further stimulus input between t1 and t2. In that period he simply reflected on the merits and the demerits of the various offers. All the information on the basis of which he makes his decision is present in his brain at t1, and the processes between t1 and t2 are simply a matter of deliberation leading to the choice of Aphrodite.

Using this example we can now state the problem of the freedom of the will with somewhat more precision than we have been able to do so far. If the total state of Paris's brain at t1 is causally sufficient to determine the total state of his brain at t2, in this and in other relevantly similar cases, then he has no free will. And what goes for Paris goes for all of us. If the state of his brain at t1 is not causally sufficient to determine the subsequent states of his brain up to t2, then, given certain assumptions about consciousness that I need to make clear, he does have free will. And again, what goes for Paris goes for all of us.

Why does it all come down to this? The answer is that the state of his brain immediately prior to t2 is sufficient to determine the beginning of the muscle contractions that caused and realized his action of handing the apple to Aphrodite. Paris was a mortal man with neurons like the rest of us and as soon as the acetylcholene

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

reached the axon end plates of his motor neurons, then, assuming the rest of his physiology was in order, his arm, with apple in hand, started to move toward Aphrodite by causal necessity. The problem of free will is whether the conscious thought processes in the brain, the processes that constitute the *experiences* of free will, are realized in a neurobiological system that is totally deterministic.

So we have two hypotheses, first that the state of the brain is causally sufficient, and second that it is not. Let us explore each in turn. On Hypothesis 1 let us suppose that the antecedently *insufficient* psychological conditions leading up to the choice of Aphrodite at t2, the conditions that led us to the postulation of the gap, are matched at the lower neurobiological level by a sequence of neurobiological events each stage of which is causally *sufficient* for the next. On this hypothesis we would have a kind of neurobiological determinism corresponding to a psychological libertarianism. Paris has the experience of free will, but there is no genuine free will at the neurobiological level. I think most neurobiologists would feel that this is probably how the brain actually works, that we have the experience of free will but it is illusory; because the neuronal processes are causally sufficient to determine subsequent states of the brain, assuming there are no outside stimulus inputs or effects from the rest of the body. But this result is intellectually very unsatisfying because it gives us a form of epiphenomenalism. It says that our experience of freedom plays no causal or explanatory role in our behaviour. It is a complete illusion, because our behaviour is entirely fixed by the neurobiology that determines the muscle contractions. On this view evolution played a massive trick on us. Evolution gave us the illusion of freedom, but it is nothing more than that—an illusion.

I will say more about Hypothesis 1 later, but first let us turn to Hypothesis 2. On Hypothesis 2 we suppose that the absence of causally sufficient conditions at the psychological level is matched by an absence of causally sufficient conditions at the neurobiological level. Our problem is, what could that possibly mean? There are no gaps in the brain. In order to take seriously the hypothesis that the free will that is manifested in consciousness has a neurobiological reality, we have to explore the relation of consciousness to neurobiology a little more closely. Earlier I described consciousness as a higher level feature of the brain system. The metaphor of higher and lower, though it is common in the literature (my own writings included), I think is misleading. It suggests that consciousness is, so to speak, like the varnish on the surface of the table; and that is wrong. The idea we are trying to express is that consciousness is a

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

feature of the whole system. Consciousness is literally present throughout those portions of the brain where consciousness is created by and realized in neuronal activity. It is important to emphasize this point, because it runs contrary to our Cartesian heritage that says consciousness cannot have a spatial location: consciousness is located in certain portions of the brain and functions causally, relative to those locations.

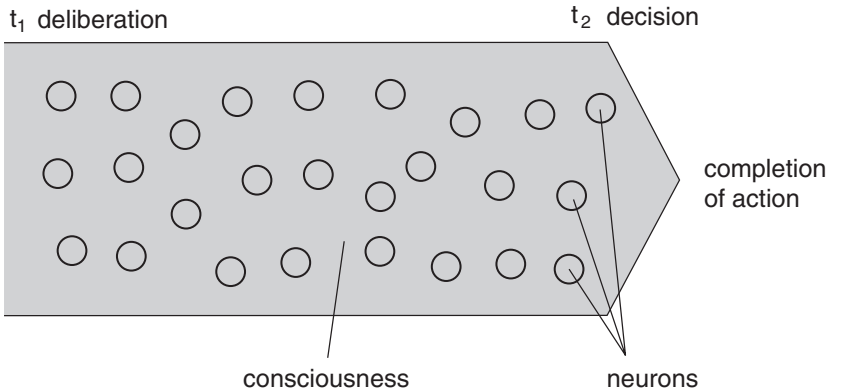
I explained earlier how consciousness could function causally, by giving an analogy between the consciousness of the brain and the solidity of the wheel, but if we carry that analysis a step further, we see that on Hypothesis 2 we have to suppose that the logical features of volitional consciousness of the entire system have effects on the elements on the system, even though the system is composed entirely of the elements, in the same way that the solidity of the wheel has effects on the molecules, even though the wheel is composed of molecules.

The point of the analogy was to remove the sense of mystery about how consciousness could affect neuronal behaviour (and thus move human bodies) by showing how, in unmysterious cases, a system feature can affect micro-level elements in a system composed entirely of the micro-level level elements, in which all causal powers are reducible to the causal powers of the micro-level elements. But of course any analogy goes only so far. The analogy: solidity is to molecular behaviour as consciousness is to neuronal behaviour, is inadequate at, at least, two points. First, we take the wheel to be entirely deterministic, and the hypothesis we are examining now is that the conscious voluntary decision-making aspects of the brain are not deterministic. Second, the solidity of the wheel is ontologically reducible to the behaviour of the molecules, and not just causally reducible. In the case of consciousness, though we suppose that consciousness is causally reducible to the behaviour of the micro elements, we cannot make a similar ontological reduction for consciousness. This is because the first person ontology of consciousness is not reducible to a third person ontology.

So far then, in our preliminary formulation of Hypothesis 2 we have three claims. First, the state of the brain at t1 is not causally sufficient to determine the state of the brain at t2. Second, the movement from the state at t1 to the state at t2 can only be explained by features of the whole system, specifically by the operation of the conscious self. And third, all of the features of the conscious self at any given instant are entirely determined by the state of the micro elements, the neurons, etc. at that instant. The systemic features are entirely fixed at any given instant by the micro

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

elements, because, causally speaking, there is nothing there but the micro elements. The state of the neurons determines the state of consciousness. But any given state of neurons/consciousness is not causally sufficient for the next state. The passage from one state to the next is explained by the rational thought processes of the initial state of neurons/consciousness. At any instant the total state of consciousness is fixed by the behaviour of the neurons, but from one instant to the next the total state of the system is not causally sufficient to determine the next state. Free will, if it exists at all, is a phenomenon in time. Diagrammatically the best I can do is this:



I have stated both Hypothesis 1 and Hypothesis 2 very swiftly, and it is now time to go over them a bit more slowly to see what is involved.

V. Hypothesis 1 and Epiphenomenalism

The best way to think of Hypothesis 1 is to think of it as an engineering problem. Imagine you are building a conscious robot. You build it in such a way that when confronted with choices it has the conscious experience of the gap. But you construct its hardware in such a way that each stage is determined by the preceding stages and by the impact of outside stimuli. Each movement of the robot's body is entirely fixed by its internal states. Indeed, we already have a model for this part of the technology in traditional artificial intelligence. We simply put in computer programs that will give the robot an algorithmic solution to the problems posed by the input stimuli and the states of the system. On Hypothesis 1, Paris's judgment was preprogrammed in advance.

I have said that an objection to Hypothesis 1 is that it leads to

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

epiphenomenalism. The distinctive features of conscious rational decision-making would have no real influence in the universe. Paris's judgment, my behaviour and the robot's behaviour are all entirely causally determined by the activity going on at the micro-level. But, someone might challenge me, why is the supposition involved in Hypothesis 1 any more epiphenomenal than any other account of the relationship of consciousness to the physiological functioning of the human body?

I have claimed that once we abandon the traditional dualistic categories there is no mystery at all about how consciousness can function causally. It is simply a matter of a higher-level, or system, feature functioning causally. And, furthermore, the account that I gave does not postulate any causal over-determination. There are not two sets of causes, the consciousness and the neurons; there is just one set, described at different levels. Consciousness, to repeat, is just the state that the system of neurons is in, in the same way that solidity is just a state that the system of molecules is in. But now, on my own account, why should Hypothesis 1 imply epiphenomenalism any more than Hypothesis 2? The answer is this. Whether a feature is epiphenomenal depends on whether the *feature* itself functions causally. Thus there are many features of any event that are causally irrelevant. For example, it is a feature of the event where I accidentally knocked the glass off the table that I was wearing a blue shirt at the time. But the blue shirt was not a causally relevant aspect of the event. It is true to say, 'The man in the blue shirt knocked the glass off the table', but the blue shirt is epiphenomenal—it does not matter. So when we say of some feature of an event that it is epiphenomenal, what we are saying is that that feature played no causal role. The suggestion that I am making is that on Hypothesis 1 the essential feature of rational decision making, namely the experience of the gap—the experience of alternative possibilities open to us, the experience that the psychological antecedents of the action are not causally sufficient to compel the action, and the experience of the conscious thought processes where we make up our minds and then act—all of those features of the experience do not matter. They are irrelevant. The specific determinate forms of those features whereby we anguish over a decision and consider various reasons are as irrelevant as the blueness of my shirt when I knocked the glass over. The judgment of Paris was already determined by the antecedent state of Paris's neurons, regardless of all of his cogitations.

The mere fact that a system feature is fixed by the micro elements does not show that the system feature is epiphenomenal. On the

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

contrary, we saw how consciousness could be fixed by neuronal behaviour and still not be epiphenomenal. To show that something is epiphenomenal, we have to show that the feature in question is not a causally relevant aspect in determining what happens. The epiphenomenalism in this case arises because the causal insufficiency of the experiences of the gap and the effort to resolve the insufficiency by making up our minds is simply not a causally relevant aspect in determining what actually happens. Our decision was already fixed by the state of our neurons even though we thought we were going through a conscious process of making up our minds among genuine alternatives, alternatives that were genuinely open to us, even given all of the causes.

Epiphenomenalism is sometimes said to be explained by counterfactuals. Multiple causes apart, the truth of 'Even if A had not occurred then B would still have occurred' is supposed to be the test for whether A is epiphenomenal. But this test is at best misleading. Assuming that both the experiences of the gap and the final decisions are fixed at the neuronal level, then if the experiences had not occurred the decision would not have occurred, or at least its occurrence would not have been guaranteed, because they are both caused by the same neuronal processes. So if one is absent the cause of the other must have been removed as well. But this does not show that the experiences were not epiphenomenal. The test for epiphenomenalism is not the truth of the counterfactual, but the reasons for its truth. The test for epiphenomenalism is whether the feature in question is a causally relevant aspect. On Hypothesis 1 the distinctive features of the gap and of rational decision making are causally irrelevant.

Well, what's wrong with epiphenomenalism? As we come to understand better how the brain works, it may turn out to be true. In the present state of our knowledge, the main objection to accepting epiphenomenalism is that it goes against everything we know about evolution. The processes of conscious rationality are such an important part of our lives, and above all such a biologically expensive part of our lives, that it would be unlike anything we know in evolution if a phenotype of this magnitude played no functional role at all in the life and survival of the organism. In humans and higher animals an enormous biological price is paid for conscious decision making, including everything from how the young are raised to the amount of blood flowing to the brain. To suppose that this plays no role in inclusive fitness is not like supposing the human appendix plays no role. It would be more like supposing that vision or digestion played no evolutionary role.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

VI. Hypothesis 2. The Self, Consciousness and Indeterminism

Hypothesis 1 is unattractive, but at least it is coherent and fits in with a lot of what we know about biology. The brain is an organ like any other and is as deterministic in its functioning as the heart or the liver. If we can imagine building a conscious machine then we can imagine building a conscious robot according to Hypothesis 1. But how would one treat Hypothesis 2 as an engineering problem? How would we build a conscious robot, where every feature of consciousness is entirely determined by the state of the micro elements, and at the same time the consciousness of the system functions causally in determining the next state of the system by processes that are not deterministic but are a matter of free decision making by a rational self, acting on reasons. So described, it does not sound like a promising project for Federal funding. The only reason for taking it seriously is that as far as we can tell from our own experiences of the gap, together with what we know about how the brain works, that is precisely the condition we are in. We are conscious robots whose states of consciousness are fixed by neuronal processes, and at the same time we sometimes proceed by nondeterministic conscious processes (hence neuronal processes) that are matters of our rational selves making decisions on reasons.

How could the brain work so as to satisfy all those conditions? Notice that I do not ask, 'How *does* the brain work so as to satisfy all those conditions?' because we don't know for a fact that it does satisfy the conditions, and if it does, we have no idea how it does so. At this point all we can do is describe various conditions that the brain would have to meet if Hypothesis 2 is true.

It seems to me there are three conditions, in ascending order of difficulty and an account of brain functioning in accord with Hypothesis 2 would have to explain how the brain meets these conditions.

1. Consciousness, as caused by neuronal processes and realized in neuronal systems, functions causally in moving the body.

I have already explained in some detail how this is possible.

2. The brain causes and sustains the existence of a conscious self that is able to make rational decisions and carry them out in actions.

It is not enough that consciousness should have physical effects on the body. There are many such cases that have nothing to do with rational free actions, as when a man gets a stomach ache from worry, or throws up at a disgusting sight, or gets an erection from erotic thoughts. In addition to a neurobiological account of mental

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

causation one needs a neurobiological account of the rational, volitional self. How does the brain create a self, how is the self realized in the brain, how does it function in deliberation, how does it arrive at decisions, and how does it initiate and sustain actions?

In the sense in which I introduced the notion of the self by the transcendental argument of section III, the self is not some extra entity, rather, in a very crude and oversimplified fashion, one can say that conscious agency plus conscious rationality = selfhood. So if you had an account of brain processes that explained how the brain produced the unified field of consciousness,[4] together with the experience of acting, and in addition how the brain produced conscious thought processes, in which the constraints of rationality are already built in as constitutive elements, you would, so to speak, get the self for free. To spell this out in a little more detail, the elements necessary for an organism to have a self in my sense are first, it must have a unified field of consciousness; second, it must have the capacity for deliberating on reasons, and this involves not only cognitive capacities of perception and memory but the capacity for coordinating intentional states so as to arrive at rational decisions; and third, the organism must be capable of initiating and carrying out actions (in the old time jargon, it must have 'volition' or 'agency').[5]

There is no additional metaphysical problem of the self. If you can show how the brain does all that—how it creates a unified field of consciousness capable of rational agency in the sense just explained, then you have solved the neurobiological problem of the self. Notice that, as far as the experiences are concerned, both Hypothesis 1 and Hypothesis 2 need to meet this condition. Indeed, any theory of brain function has to meet this condition, because we know that the brain gives us all these sorts of experiences. The difference between Hypothesis 1 and Hypothesis 2 is that on 1 rational agency is an illusion. We have the experience of rational agency but it makes no difference to the world.

3. The brain is such that the conscious self is able to make and carry out decisions in the gap, where neither decision nor action is

[4] For the importance of the unified field, see John R. Searle, 'Consciousness,' *Annual Review of Neuroscience*, 2000, Vol. 23, pp. 557–78.

[5] On my view rationality is not a separate faculty, rather the constraints of rationality are already built into intentional phenomena such as beliefs and desires and into thought processes. So a neurobiological account of mental phenomena would already be an account of the rational constraints on such phenomena. For more detailed presentation of this view and the reasons for it, see my *Rationality in Action*, MIT press, forthcoming, 2001.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

determined in advance, by causally sufficient conditions, yet both are rationally explained by the reasons the agent is acting on.

This is the trickiest condition: How could the gap be neurobiologically real, given all that I have just said? Assume we had an account of how the brain produces mental causation, and an account of how it produces the experiences of rational agency, how do you get rational indeterminism into your account of brain function?

The only way I know to approach such a problem is to begin by reminding ourselves of what we already know. We know, or at least we think we know, two things that bear on the case. First we know that our experiences of free action contain both indeterminism and rationality and that consciousness is essential to the forms that these take. Second we know that quantum indeterminism is the only form of indeterminism that is indisputably established as a fact of nature.[6]

It is tempting, indeed irresistible, to think that the explanation of the conscious experience of free will must be a manifestation of quantum indeterminism at the level of conscious rational decision making. Previously I never could see the point of introducing quantum mechanics into discussions of consciousness. But here at least is a strict argument requiring the introduction of quantum indeterminism.

Premise 1. All indeterminism in nature is quantum indeterminism.
Premise 2. Consciousness is a feature of nature that manifests indeterminism.
Conclusion: Consciousness manifests quantum indeterminism.

Our aim now is to keep following relentlessly the implications of our assumptions. If Hypothesis 2 is true and if quantum indeterminism is the only real form of indeterminism in nature, then it follows that quantum mechanics must enter into the explanation of consciousness. This conclusion does not follow on Hypothesis 1. As long as the gap is epiphenomenal, then no indeterminism in the causal apparatus is essential to explain how consciousness is caused by and realized in brain processes. This is important for contemporary research. The standard lines of research, both on the building block model and the unified field model, make no appeal to

[6] Chaos theory, as I understand it, implies unpredictability but not indeterminism.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

quantum mechanics in explaining consciousness. If Hypothesis 2 is true these cannot succeed, at least not for volitional consciousness.[7]

But even assuming we had a quantum mechanical explanation of consciousness, how do we get from indeterminism to rationality? If quantum indeterminacy amounts to randomness then quantum indeterminacy by itself seems useless in explaining the problem of free will because free actions are not random. I think we should take the question, 'What is the relation between quantum indeterminacy and rationality?' in the same spirit in which we take the question, 'What is the relation between brain micro processes and consciousness?' or the question, 'What is the relation between visual stimuli, brain processes and visual intentionality?' In the latter two cases we know in advance that the system features are caused by and realized in the microprocesses, so we know that the causal features of the system level phenomena are entirely explainable by the behaviour of the micro phenomena. As I have repeated to the point of tedium, the causal relations have the same *formal* structure as the causal relations between molecular movements and solidity. We also know that it is a fallacy of composition to suppose that the properties of the individual elements must be properties of the whole. Thus for example, the electrical properties of the individual atoms are not properties of the whole table, and the fact that a particular action potential is at 50 Hz does not imply that the whole brain is oscillating at 50Hz. Now exactly analogously, the fact that individual micro phenomena are random does not imply randomness at the system level. The indeterminacy at the micro level, may (if Hypothesis 2 is true) explain the indeterminacy of the system, but *the randomness at the micro level does not thereby imply randomness at the system level*.

Conclusion

I said at the beginning that obdurate philosophical problems arise when we have a conflict between deeply held inconsistent theses. In the case of the mind body problem we resolved the inconsistency by a kind of compatibilism. Once we abandon the assumptions behind the traditional Cartesian categories then naive materialism is consistent with naive mentalism. We could not make such a compatibilism work for the free will problem, because the thesis that every human act is preceded by causally sufficient conditions remains

[7] For an explanation of the distinction between the building block model and the unified field model, see John R. Searle, 'Consciousness' *Annual Review of Neuroscience. 2000*, Vol. 23, pp. 557–78.

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press

incompatible with the thesis that some are not. Once we sorted out the issues we found two possibilities, Hypothesis 1 and Hypothesis 2. Neither is very appealing. If we had to bet, the odds would surely favour Hypothesis 1, because it is simpler and fits in with our overall view of biology. But it gives a result that is literally incredible. When I gave this lecture in London someone in the audience asked, 'If Hypothesis 1 were shown to be true would you accept it?' The form of the question is: 'If free rational decision making were shown not to exist, would you freely and rationally make the decision to accept that it does not exist?' Notice that he did not ask, 'If hypothesis 1 were true would the neuronal processes in your brain produce the result that your mouth made affirmative noises about it?' That question at least is in the spirit of Hypothesis 1, though even that goes too far, because it asks me freely and rationally to make a prediction, something that is impossible on the Hypothesis.

Hypothesis 2 is a mess, because it gives us three mysteries for one. We thought free will was a mystery, but consciousness and quantum mechanics were two separate and distinct mysteries. Now we have the result that in order to solve the first we have to solve the second and invoke one of the most mysterious aspects of the third to solve the first two. My aim in this article is to continue the line of attack begun in my earlier writings and to follow out the competing lines of reasoning as far as they will go. There is, I am sure, much more to be said.

University of California at Berkeley

https://doi.org/10.1017/S0031819101000535 Published online by Cambridge University Press