Consciousness: The First-Person and Third-Person Views

David Chalmers
Lincoln College, Oxford
30 December, 1987

1 Introduction

An important distinction which we must make when talking about consciousness, or any mental activity, is that between the *first person* and *third person* views. The third person (or *objective*) view looks at consciousness as it exists in others. It explains mental properties in objective terms, and shows how they determine our behaviour. This is the domain of neurophysiologists and psychologists, and of some philosophers of mind. It is distinguished by the fact that there is no 'mystery' surrounding the phenomenon, just technical difficulties.

The most interesting and most difficult philosophical problems arise when we take the first person (or *subjective*) view. How can it be that there is a being which *experiences* this 'consciousness?' Which has an 'inner life?' Such beings do in fact exist, for I have direct introspective evidence for it in myself. Were it not for this evidence, we would be inclined to dismiss the question as ridiculous. It has a mystical, unscientific air surrounding it. From the objective viewpoint, there need be no 'experiencers', merely physical objects which form themselves into complex aggregates, which have abstract properties which it might be convenient to label as 'mental', as a form of shorthand. (Indeed, 'convenience' is the mark of third-person mental attributions. Contrasting with this, first-person phenomena have something 'absolute' about them. Something really is going on, it's just hard to say what.)

This problem is of course just the mind-body problem. We might phrase the problem thus: "Given a being which is conscious from the third-person viewpoint, is it conscious from the first-person viewpoint, and why?". Often the first-person difficulties are ignored in favour of giving a good third-person account of consciousness. Nevertheless it is with the first person that the real difficulties lie. This is basically the point which Nagel [1] makes when he asks "What is it like to *be* such a being?".

We might say that a third-person-conscious being "seems conscious," and that a first-person-conscious being "is conscious." Most people would accept that in principle it is possible for, say, a computer to be third-person-conscious. Many would harbour doubts as to whether a computer could be first-person-conscious. Whenever I use the word 'conscious' alone, I will mean 'first-person-conscious.' Most of this discussion applies to any first-person mental activities; I am using consciousness as the paradigm of the mental.

The starting point of this essay is the observation that first-person experience of consciousness is irrelevant in explaining what happens in the physical world; in particular, in explaining our behaviour. If we grant that the laws of physics are nowhere violated, then we can in principle explain and predict all physical events from a third-person viewpoint. We can regard the human brain as a highly complex computer, biological but still mechanical. This computer accepts input, acts on it mechanically and produces output. We might want to use the third-person view of consciousness as a convenient shorthand, but the first-person view is never required. Even if there were no being which 'experienced' this consciousness, the behaviour would remain the same.

In particular, the 2000-year debate over the mind-body problem is independent of first-person experience of consciousness. Humans would still debate "What is this funny feeling we have of consciousness?" whether or not there was in fact anyone 'experiencing' that feeling. This essay would still be written even if there were not an "I" to write it. Further, we realize that all our descriptions of consciousness are just how we would expect any third-person-conscious being to describe it. Our first-person experience gives us 'nothing extra.' After all, our 'mystical' evocations of consciousness are all simply products of the laws of physics being obeyed in our brains; a third-person phenomenon. Nevertheless *we* (or I) know there is something more - I am 'in here', experiencing. But we can't really convince anyone of our consciousness in a watertight way. It's a frustrating, impotent feeling.

Some of these statements feel rather jarring, and don't quite ring true. "What do you mean, I'd still say this if I wasn't conscious? It's *because* I'm conscious that I'm saying it in the first place!" This is because we are so used to implicitly identifying third-person and first-person consciousness. Indeed this may be a valid identification, but in almost every existing reasonable third-person theory of consciousness, there is nothing to suggest why there should be a first person *experiencing*. Nevertheless, if we could show that third-person consciousness necessarily entails first-person consciousness, the above statements would lose much of their punch. If the hypothesis of seemingly intelligent beings "not being conscious" was impossible, then any conclusions would be irrelevant.

This is the view which I will try to support. This essay aims to show that intelligent behaviour without consciousness is a faintly ridiculous notion. In particular, I show that the only possible *convincing* solution to the mind-body problem can be a universal solution showing logically that consciousness must accompany intelligent behaviour. After all, we have already seen that arguments from personal experience are weak and unconvincing - so the only possible convincing argument must be universal, independent of personal experience. (Of course, it is possible that no convincing solution exists - perhaps there is no convincing reason to accept the true solution. This would be directly analogous to Godel's Incompleteness Theorem in mathematics.) To put the conclusion more pithily, though more weakly: if machines cannot be conscious, then the mind-body problem is insoluble.

2 Smart Beings

In this essay I will use the term 'smart being' to denote a being whose capacity for intelligent behaviour rivals that of a human. By this I mean that it can perceive the world (even if only verbally), it can use language to communicate, and it is capable of rational discussion and argument on a range of objective topics similar to that of humans (where rationality is judged by human standards. In particular, I assume that an objective argument which convinces a human will convince such a being). It must also be flexible, able to give a reasonable answer to any reasonable question, even if only a justification of its ignorance. For instance, a computer which passed the Turing Test would satisfy this definition, though the Turing Test is too restrictive to be a necessary condition for smart beings - it requires that they simulate humans' limitations as well as their capabilities. I also postulate that smart beings are subject to the laws of physics.

I have already used mentalistic terms when speaking of smart beings - saying that they 'perceive' and 'communicate', for instance. This is not meant to beg the question of whether or not they have a first-person mental life. Instead they are merely third-person ascriptions - descriptions of their behaviour. When I say 'communicate' I mean perhaps 'utter certain sounds which we interpret'. When I say 'perceive' I mean 'behave in a manner consistent with perception' - making consistent and correct claims about the 'outside world', and so on. It was an achievement of Ryle [2] to show that many (though not all) mental properties can be coherently ascribed in a third-person way on the basis of behaviour. Of course, this remains a third-person view. The question of whether and why such beings are first-person-conscious is unanswered.

There seems to be no philosophical difficulty with the idea that non-human smart beings can exist in principle. For instance, again assuming that the laws of physics are not violated in the human brain, we could in principle get a computer to simulate the brain digitally by numerically solving a set of partial differential equations with appropriate initial conditions, to a required degree of accuracy (for instance to within the degree of quantum uncertainty given by Planck's constant). Of course the practical difficulties would be rather non-trivial, but no matter! The question of whether or not such smart beings would be conscious is at the centre of the mind-body problem.

Now, it is a central observation that any smart being would claim to be conscious; it would be third-person-conscious. For a start, it must have some notion of self or "I" in speech. If it replied to the question "What did you just say?" with "I don't understand what you're referring to when you use the word 'you'", it would not meet our criteria for a smart being - it would not be capable of rational discussion of its own conversation. Further, it must have some perception of its own processing, as we see in the following.

Here is an imagined dialogue with a smart being, say a computer, which claims not to be directly aware of its own processing.

Human: How do you know what I am saying now?

Computer: My input device recorded the statement and I retrieved it from there.

Human: Who retrieved it from there?

Computer: I use the word "I" merely as a convenient shorthand for this collection of circuits. When I say "I retrieved" a statement I simply mean that information coding the statement passed into my central processing unit.

Human: What caused this to happen?

Computer: The laws of physics, acting on the circuits in the way they were designed.

Human: Sure, you know intellectually that that's what happened, just as I know that I am 'nothing but' a collection of neurons flowing along chemical pathways. But are you saying that you have no direct perception of this event?

Computer: Yes, I simply deduce that it happened from the statements which appeared on my output device.

Human: So earlier, how were you justified in making your first statement?

Computer: I don't know, it just came out.

Human: So you have no direct awareness of your own processes?

Computer: That's right. I have no idea what I am saying right now until I read it on my output device.

This is clearly ridiculous. The point is that a smart being's perception of the world cannot suddenly end with itself. If it perceives, it must perceive its own perception. If you like, a smart being cannot perceive the objective without perceiving the subjective. It must claim some awareness of its own processing. And once it claims this it must claim consciousness.

Computer: OK, I concede that I am directly aware of my own processing.

Human: Where do you get this awareness from?

Computer: I guess it was programmed into me.

Human: But how do you perceive this awareness?

Computer: I don't know. It's just <u>there</u>. Although I realize that it must be a direct consequence of my programming, this doesn't help me understand it any better. For example, when I'm talking to you I know intellectually that I perceive what you say by electrons coding your statement flowing into my central processing unit. But that's not the way it seems to <u>me</u>. Your statement just appears in my consciousness - I 'hear' it. I don't know how to explain this. I would say that it's an illusion, except that it feels so real to me.

This last point, that self-perception does not occur on the lowest level, is important. The point is that the way a being perceives its own processing is on a higher level; in terms of the patterns that the lower level forms and the information it represents. It is this perception from the inside that gives rise to at least the 'illusion' of consciousness - perception of pattern and information 'all at once' seems strange from the inside. Of course this point is still a third-person, functional view. There is nothing yet to suggest whether there is a first-person experiencer in there. This view will be taken further in my essay 'Mind, Pattern and Information.' There I attempt to identify the first person with the pattern and information.

But for now, we have established that any smart being would at least claim to be conscious. Given a modicum of intellectual curiosity, amongst any smart beings there would arise a debate over the nature of consciousness which would roughly parallel the debate amongst humans. (Computer A: "I know that my central processor is nothing but a collection of digital circuits, but I can't explain this feeling I have of having thoughts of my own." Computer B: "I know what you mean. I'm inclined towards dualism myself.") It seems likely they would describe consciousness in much the same way as humans - having thoughts "somewhere in there", perceptions just "appearing", having a sense of "self."

This should not be too surprising, for after all our descriptions of consciousness and mental events are products of a physical system, explainable in the third person. Indeed it is striking that whenever we use words to try and capture the concept of consciousness, we feel that we are not really capturing the elusive nature of the phenomenon. All these words, such as 'experience', 'identity', 'images', 'thoughts', even 'first-person' - they are all words which we should expect a smart being to use; they all correspond to third-person concepts. This is why the phenomenon of first-person consciousness is so hard to write about, and gets ignored so often. It also explains the popularity of such doctrines as eliminative materialism - if all we need is third-person concepts to describe a system, then why bother with the first person at all? Nevertheless, the first person really is there, I know it, though you may not believe me.

3 Arguments About Consciousness

We have seen that there is nothing 'special' about the kind or the properties of our consciousness - any smart being would describe it in the same way. The only possible difference is that our consciousness might actually 'exist', whereas that of some other smart being is only a false claim. This would seem a trifle strange, but it is not (at least so far) logically contradictory. This view is analogous to solipsism, and is unconvincing for much the same reason that solipsism is unconvincing. If solipsism were convincing, we would all be wandering around saying "I am the only one with a mind; you are all deluded." We would be believing contradictory statements, therefore we could not all be rational.

I will call the view that consciousness is a necessary property of a smart being the egalitarian view. If you like, you can think of this view as saying that any third-person-conscious being is necessarily first-person-conscious. The contrary view, that it is possible for a smart being not to be conscious will be labelled 'non-egalitarian.' The central thesis of this paper is that any non-egalitarian 'solution' to the mind-body problem must be unconvincing. (By a 'convincing' statement I mean one that convinces a rational being through rational argument. This is analogous to the concept of 'provability' in mathematics. By a 'solution' to the mind-body problem I mean at the very least a classification of beings into 'conscious' and 'non-conscious' categories.) Of course, a statement's being unconvincing does not preclude its being correct.

We can regard the non-egalitarian position as saying that first-person-consciousness is something 'special,' something extra over and above apparent third-person properties and separable from those. But if we grant that the first-person and the third-person are separable, we must take heed of arguments suggesting that (assuming the laws of physics are not violated in the human brain) the entire mind-body debate would have gone exactly the same even if humans were not conscious. Essentially, if we accept the non-egalitarian view, consciousness plays no role in the mind-body debate. In particular, it is impossible to argue convincingly from first-person experience of consciousness (as we would argue the same whether we were conscious or not). The only way around this is to argue on objective grounds, independent of personal experience. Of course, if we accept the egalitarian view, the first-person is inseparable from the third-person and such arguments lose their validity.

To see this more clearly: imagine that it was possible for a human being (that is, a creature physically the same as a human which behaves like a human) not to be conscious. Imagine also that we had a convincing solution to the mind-body problem. Then our non-conscious human would (claim to be) equally convinced by this solution, as behaviour is independent of consciousness. But if this solution is correct it must establish that she is non-conscious. Thus she is convinced that she is non-conscious. By independence of behaviour from consciousness, all humans must therefore be convinced that they are non-conscious. This gives a contradiction, for consciousness is basically defined to be that strange property of awareness possessed by humans (or at least by me!). (A more exact definition is impossible for the reasons outlined above.) The dual view, that no humans are in fact conscious, is more or less another version of the egalitarian view.. Both views assert that consciousness is not 'something extra' over and above intelligent behaviour.

As we have a contradiction, one of our assumptions at the top of the last paragraph must be false. That is, one of the two following statements must be true.

(1) A non-conscious human being is impossible.

(2) There exists no convincing solution to the mind-body problem.

Note that throughout this discussion we make the assumption that human beings are rational, so that a convincing solution must also be correct.

Of course, the above conclusion is fairly trivial and can be seen more easily as follows: A convincing solution to the mind-body problem must give us an observable and third-person-verifiable property P as a criterion for whether a given being is conscious. But if one being is physically identical to another, then the occurrence or non-occurence of all such properties P must coincide in the two. So if property P establishes that one human being is conscious, it must establish that all physically identical humans are conscious.

The advantage of the argument as previously stated is that it extends to beings which are not necessarily physically identical, but which have identical behavioural dispositions. I cannot imagine being convinced by any argument purporting to show that I was not conscious. Even if my skull was opened up and shown to consist of metal parts, and this was accompanied by an argument showing that consciousness is inherently biological, I could not accept it. To accept the argument would force me to believe a contradiction. I know I am conscious. The point is that this stubbornness on my part is a behavioural disposition; one which I presume is shared by all humans. If there could be a non-conscious smart being with identical behavioural dispositions to mine, then if it could not be convinced by any argument showing it to be non-conscious. So no convincing solution could exist.

To extend this argument to smart beings in general is not difficult. Our earlier discussion implied that behaviourally, any smart being will be as convinced of its own consciousness as we are of ours. From its own point of view, the fact of its consciousness will seem indisputable. The following paragraph gives our main argument in this most general setting. This is the central argument of this paper.

Imagine we had a convincing non-egalitarian solution to the mind-body problem. Then it is possible that a non-conscious smart being exists. As we assumed our smart beings are rational, it must also be convinced by this solution. Thus, as the solution is correct, it is convinced that it is not conscious. But as we saw earlier, it must be convinced that it is conscious. So it is forced to utter a contradictory claim: "I am conscious and I am not conscious." Therefore it is not rational, and we have a contradiction. So there can be no convincing non-egalitarian solution to the mind-body problem.

Recall that underlying this whole discussion was the assumption that the laws of physics are not violated in the human brain. Although this seems a reasonable assumption to make, there are some who would dispute it: for example interactionist dualists and some religious people. So for completeness we include it here. We have shown that one of the following three statements must be true.

(1) The laws of physics are violated in the human brain.

(2) There can be no convincing solution to the mind-body problem.

(3) A smart being is necessarily conscious.

4 Discussion and Speculation

So we have three statements, each expressing a strong conclusion, one of which must be true. Let us examine the statements individually.

Of the three, I have no difficulty in rejecting (1). The consequences of accepting it would be far more radical than the consequences of accepting (2) or (3).

It is interesting to see where our arguments would break down if we accepted (1). The argument about human beings would break down as there would be no reason to suppose that behaviour is independent of consciousness. The argument about smart beings in general would break down as (1) would leave open the possibility that the human brain could possess some higher insight (religious in nature, perhaps) which is not available to smart beings in general. Or conceivably it might be impossible for some reason that a fully rational smart being could exist, subject to the laws of physics. Either way, an argument convincing to humans might not be convincing to other beings.

The laws of physics in fact play an important role in the mind-body problem. In this paper, when I speak of 'possibility' and 'necessity' I always mean possibility and necessity subject to the laws of physics. This is a weaker notion than logical possibility. A priori, it is not necessary that when I drop an apple it should fall to the ground; but subject to the laws of physics, it is necessary. Similarly, I can conceive of a non-conscious block of marble discoursing lucidly on the meaning of life, so a non-conscious smart being is not logically impossible. But this example is not possible subject to the laws of physics - intelligent behaviour requires great internal complexity. An egalitarian theory (functionalism for instance, or the pattern/information theory that I propose) could identify mental events with some aspect of that complexity.

If statement (2) is true then we are in an interesting position. Presumably some solution exists, but it is not convincing to human beings; either because the human brain (and our system of reason) is not powerful enough to comprehend the reason for its truth, or because there exists no explanation for its truth - it just 'happens to be' that way.

This is directly analogous to Gödel's Incompleteness Theorem in mathematics, which states that in no consistent system of mathematics can all the truths of arithmetic be proved. There must be some statement which is true but unprovable. This is because for any system which is powerful enough to represent all rational deduction, then if the system could prove the statement, a contradiction would follow, so it could not be consistent - thus there exists no proof for the statement.

Notice the analogy between this and our argument for why the mind-body problem is insoluble under non-egalitarianism. If a non-conscious being were powerful enough to be rational, then if it could be convinced by a non-egalitarian solution, it would be forced to believe a contradiction, so it could not be consistent - thus the solution cannot be convincing. (There remains open the possibility that human/smart beings are necessarily irrational and inconsistent. This would correspond to an inconsistent mathematical system in Gödel's Theorem. In either case a true statement is believed, but for all the wrong reasons.) One might say that under non-egalitarianism, the solution to the mind-body problem would be the Gödel sentence for the human brain.

Beautiful though this analogy is, it is in (3) that any hope for solution to the mind-body problem lies. Throughout this essay it has been apparent that to talk about smart beings without consciousness is very strange (though not yet contradictory). For a start, it contradicts their claims. When and if intelligently behaving computers are constructed, we are going to feel very strange telling them that they are not conscious, against their professed belief. Further, it struck a jarring note to talk of consciousness playing no causal role in the mind-body debate. It seems strange to say that human history would have been the same without consciousness. If we could show that intelligent behaviour and consciousness are inseparable, these statements would become meaningless.

So our only hope lies in a universal, logical argument showing that consciousness is a necessary adjunct of intelligent behaviour. Exactly how this can be done appears uncertain. One possible approach: any being that is capable of intelligent behaviour whilst subject to the laws of physics _must_ physically consist of a highly complex structure, with information coded in complex patterns. Perhaps consciousness can be identified with those patterns. This theme is developed in my second essay, "Mind, Pattern and Information."

Whichever path we take, we must reconcile the third-person and first-person views of mentality. One question which any theory must answer is "Why does the brain perceive a mind." The fact that we talk about having minds is a third-person phenomenon. There should be no intrinsic 'mystery' to this. A good theory must investigate why this happens, and show how it relates to the first-person phenomenon of the mind which we experience. So we have three things between which we must find the relationship:

(1) the brain
(2) the mind which we experience (a first-person phenomenon)
(3) the mind which the brain perceives (a third-person phenomenon)

One would hope at least for an identity between (2) and (3). This way it could be seen that the first-person and third-person views are truly inseparable.

[1] T. Nagel, "What Is It Like to Be a Bat?" The Philosophical Review, October 1974

[2] G. Ryle, The Concept of Mind (London: Hutchinson & Co., 1949)