Merricks on the existence of human organisms

Cian Dorr

August 24, 2002

Merricks's Overdetermination Argument against the existence of baseballs depends essentially on the following premise:

BB Whenever a baseball causes an event, the baseball's constituent atoms also cause that event, and the baseball is causally irrelevant to whether those atoms cause that event.

This argument can be transformed into an argument against the existence of human organisms by replacing BB with

HO Whenever a human organism causes an event, the human organism's constituent atoms also cause that event, and the human organism is causally irrelevant to whether those atoms cause that event.

Since Merricks believes in human organisms, he needs to explain why the considerations that support BB do not support HO equally well. What is the relevant disanalogy between hypothetical baseballs and hypothetical human organisms?

Merricks's answer to this question is this: If baseballs exist, all the facts about them are metaphysically determined by the microphysical facts. By contrast, human organisms, if

they exist, have some properties—including consciousness and various more specific conscious mental properties—whose instantiation is not metaphysically determined by the microphysical facts. In the first part of this paper, I will make some critical remarks about Merricks's argument for the thesis that consciousness is special in this way. In the second, I will express some doubts about Merricks's claim that if this thesis is true, we have no reason for believing HO. Finally, I will present a dilemma for Merricks's view about the sort of causal efficacy enjoyed by human organisms.

1

Let us understand the facts about the *arrangement* of some things to include all the facts about their intrinsic properties and the spatiotemporal and causal relations among them. Using this abbreviation, Merricks's argument for the claim that the facts about consciousness do not supervene on the microphysical facts can be reconstructed as follows:

P1 Whenever some atoms compose something with a conscious mental property, it is possible that some atoms arranged in exactly the same way do not compose something with that property.[1]

P2 All "conscious and subjective" mental properties are intrinsic.[2]

P3 If the facts about the arrangement of some atoms do not determine whether they

[1] As Merricks puts it: 'the existence of an object with an intrinsic conscious mental property is not entailed by the intrinsic properties and spatiotemporal and causal interrelations of that object's constituent atoms' (p. 94).

[2] We need not worry about the meaning of "subjective": all that matters is that a 'healthy, awake, adult human person' (p. 93) is supposed to have many such properties, and that possession of sufficiently many such properties is supposed to be necessary and sufficient for consciousness.

> compose something that has a certain intrinsic property, then the microphysical facts do not determine whether those atoms compose something that has that property.[3]

Therefore,

> Whenever some atoms compose something that has a conscious and subjective mental property, it is possible that those atoms do not compose anything that has that property even though the microphysical facts are just the same.

This argument is valid, and it makes a welcome change from sitting around contemplating our intuitions about the possibility of zombies; but I don't think it should convince anyone. Serious doubts can be raised about each of its premises, and there is good reason to think that at least one of them is false.

Merricks argues for P1 as follows: For any normal person, the atoms that are parts of that person but not, as we would ordinarily put it, part of that person's left index finger—for short, the "finger-complement atoms"—do not compose any conscious being. But if the "finger" atoms were annihilated, the finger-complement atoms would compose a conscious being, without (at least initially) differing in their arrangement.

This argument seems convincing, but it does not establish P1: it only establishes the weaker claim there are *some* possible cases in which the fact that some atoms compose a conscious being is not determined by their arrangement.[4] Someone might accept this while holding that there are many other cases where the arrangement of some atoms *does* determine

[3]As Merricks puts it: 'I assume that if an *intrinsic* property is not fixed, of necessity, by its atoms, it is not so fixed by its atoms plus the atoms that fill its environment' (p. 91, footnote 5); 'Intrinsic properties, just by their very nature, *supervene* either locally or not at all' (p. 100, footnote 10).

[4]This weaker claim is the negation of Merricks's claim (C) (p. 94).

that they compose a conscious being. How could we rule out this view? One way to do so would be to appeal to the following somewhat plausible premise: if a single new atom were embedded somewhere deep inside a conscious being, without disturbing the arrangement of any other atoms, the new atom would become part of that conscious being, so that the old atoms would no longer compose any conscious being. If this is true, all conscious beings are in the same situation as people whose fingers have just been annihilated. However, someone might reject this premise on the grounds that a new atom embedded in a person would not be part of that person until it had been "taken up into the person's life", and thereby affected the arrangement of the person's other atoms.[5] Thus, it remains unclear whether P1 can be established by any argument along the lines of Merricks's.

Merricks's argument for P2 consists, essentially, in a challenge to opponents of P2 to give a credible theory of consciousness which explains how relational facts about an object can be partly constitutive of its conscious and subjective mental properties. Merricks considers one possible response to this challenge, due to Theodore Sider (MS). Sider suggests the following schematic account of what it is to be conscious: there is some intrinsic property (Merricks calls it 'pseudo-consciousness'), such that for an object to be conscious is for it to have that property and not be a proper part of anything else that has that property. Similarly for more specific conscious mental properties. Merricks's objections to this view all depend on imputing to Sider the additional thesis that 'there is *no phenomenological difference* between the conscious and the merely pseudo-conscious' (p. 101). This really is a very odd thing for anyone to think. Surely we should all agree that without consciousness there is no such thing as "phenomenology"; so of course there is a "phenomenological difference" between any conscious being and any being that is not conscious. I see no reason for a proponent of

[5]Cf. van Inwagen's claim that a strontium atom embedded in a person's tissues would not be part of that person (van Inwagen 1990, p. 95).

Sider's view to hold this problematic thesis; without it, the view looks quite promising, and is immune to Merricks's objections.

What about P3? Although Merricks treats this premise as if it were obvious, it is in fact quite a strong principle. Since trivial properties like *self-identity* are intrinsic, P3 entails the following controversial principle about composition:

CP If the facts about the arrangement of some atoms do not determine whether they compose something, then the microphysical facts do not determine whether those atoms compose something.

Many of Merricks's opponents reject CP. For example, anyone who believes in statues but denies the doctrine of arbitrary undetached parts (van Inwagen 1981) should, I think, deny CP. Take some atoms which in fact compose a statue: if these atoms were arranged in the same way while seamlessly embedded in a block of marble, they wouldn't compose anything. Given CP, it would follow that the fact that these atoms compose something is not determined by the microphysical facts. If you wanted to accept this conclusion, you would have the makings of a reply to the overdetermination argument against the existence of statues analogous to Merricks's reply to the overdetermination argument against the existence of human organisms.[6] But the conclusion is deeply implausible. It would be much more natural to deny CP, and with it P3. Insofar as Merricks's formulations of P3—e.g. 'intrinsic properties, just by their very nature, *supervene* either locally or not at all' (p. 100, footnote 10)—seem obvious, this is because they can be read as expressing the following alternative thesis:

[6]Merricks considers such a reply: his response (p. 106) is that we have no reason to believe the premise that whenever some atoms compose a statue, the fact that they compose something is not fixed by their intrinsic properties and spatiotemporal and causal relations. But surely van Inwagen's arguments against the doctrine of arbitrary undetached parts do at least give us *some* reason to accept that premise.

P3* If the facts about the arrangement of some atoms, together with the fact that those atoms compose a certain object o, do not determine whether o has a certain intrinsic property, then the microphysical facts, together with the fact that those atoms compose o, do not determine whether o has that property.

P3* does not entail CP, and seems like a much better candidate than P3 to be an analytic truth about intrinsicness. However, the substitution of P3* for P3 would render Merricks's argument invalid.

Different philosophers will wish to reject different ones of Merricks's premises; but everyone should reject at least one of them, for in conjunction they have some very implausible consequences. If Merricks's argument for P1 is sound, the facts about the arrangement of my finger-complement atoms do not suffice to determine that these atoms do not compose a conscious being. So by P2 and P3, the microphysical facts do not suffice to determine that my finger-complement atoms do not compose a conscious being. In other words, there is a possible world microphysically just like this one, in which my finger-complement atoms compose a conscious being. Could this really be true? I can certainly make sense of the possibility that, despite appearances, I don't have any atoms arranged left-index-fingerwise as parts. Perhaps, unbeknownst to me, my left index finger is some sort of prosthesis; perhaps the region I think of as occupied by my left index finger is entirely empty, and I am prevented from noticing this fact by the machinations of an evil demon. But when I try to entertain the supposition that those atoms are not part of me even though they are related to my other atoms in more or less the same ways as the atoms "in my other fingers", I find myself at a loss. The hypothesis I am being asked to entertain seems impossible.[7]

[7]Those who deny this modal intuition face some difficult epistemological questions. Surely we can at least agree that I know that this hypothesis is false. But how do I know this, if I don't know it in the way I know metaphysically necessary truths? I don't have conclusive

2

If Merricks is right that baseballs, if they existed, would have no properties that do not supervene on the microphysical facts, whereas human organisms, if they existed, would have many such properties, he at least has a good response to the charge that it is arbitrary to endorse BB while rejecting HO. But Merricks aspires to do more than to avoid the charge of arbitrariness: he wants to use his conclusions about consciousness to show that there is no good reason to believe HO.

I think this argument is unsuccessful: the argument which, to my mind, provides the most important reason to believe HO is largely unaffected by Merricks's conclusions about consciousness. The argument I have in mind is an argument from the empirical successes of microphysics, and goes something like this: Microphysics has been successful in finding explanations of a wide variety of physical events. As Merricks points out, it hasn't yet explained *every* physical event—indeed, it never will; there are far too many of them. Nevertheless, many of us think that the empirical success that microphysics has met with so far is good reason to believe—or "bet", as Merricks puts it (p. 111)—that some theory of the same general sort as our current best microphysical theories is true. But our current theories are all either deterministic, or if they are indeterministic, they predict that the chance of any microphysical event occurring is fixed by other microphysical facts. If any theory of this sort is true, then every microphysical event either has a wholly microphysical cause to which non-microphysical entities are causally irrelevant, or has no cause at all. But if human organisms are ever non-overdetermining causes, they must be non-overdetermining causes of

empirical evidence against the hypothesis, since my experiences would be just as they actually are if the hypothesis were true. True, the hypothesis is rather arbitrary and inelegant; but could I really *know* it to be false if I have no better reason for rejecting it than this?

some microphysical events. Hence, every event caused by a human organism has a wholly microphysical cause to which non-microphysical entities are causally irrelevant.

What effect should the discovery that conscious mental properties don't supervene on the microphysical have on this argument? Not much, I think. Indeed, there is some risk involved in drawing conclusions about the behaviour of microphysical entities that are parts of beings with non-supervenient properties from experimental evidence which mostly concerns entities that are not parts of entities with non-supervenient properties. But this sort of risk occurs all the time in science. One discovers that particles in a certain highly artificial experimental situation behave in a certain way; one infers that all particles behave in that way, not just in that experimental situation, but in outer space, and on the surface of the Sun, and inside our coffee cups. I don't see why the inference from the experimental evidence to conclusions about atoms that are parts of conscious beings should look any worse than those other inductions, even after we have discovered that consciousness does not supervene on the microphysical. We have no empirical evidence that it matters to the behaviour of an atom whether it is part of a conscious being; we do have empirical evidence that it doesn't matter *very much* to the behaviour of an atom whether it is part of a conscious being; under such circumstances, the assumption that it doesn't matter at all to the behaviour of an atom whether it is part of a conscious being seems to be justified by standard scientific methodology.

Thus, the question whether facts about consciousness supervene on microphysical facts seems to be more or less irrelevant to the question whether the argument from the empirical success of microphysics gives us good reason to believe HO. Of course, the argument may fail for other reasons. But if it does, Merricks has not told us what they are.[8]

[8]In chapter 6, Merricks gives an independent argument against HO, based on considerations about free will. But even if this worked, it would not show that the argument from the empirical success of microphysics gives us *no reason* to believe HO.

3

Merricks holds that people, as well as their mental properties, have many physical properties (like mass and shape) that might be shared by inanimate composite objects, if there were any such things. Call these *straightforwardly physical* properties. I am not sure whether Merricks thinks that people ever cause things by having straightforwardly physical properties. But I see trouble either way.

If Merricks says that people do cause things in virtue of their straightforwardly physical properties, he will be forced to grant that there is systematic overdetermination, thereby undermining the Overdetermination Argument against the existence of baseballs. For it seems that whenever a person causes some event by having a straightforwardly physical property, the person's atoms also cause that event by virtue of their microphysical features, and the person's having that property is causally irrelevant to whether the atoms' having those microphysical features causes that event. Suppose for example that a person caused some atoms arranged scales-wise to read '75' at a certain time by weighing 75kg at that time. Then, surely, it is also true that the person's atoms caused the atoms arranged scales-wise to read '75' by collectively weighing 75kg at that time. And while the *person* may well be causally relevant to the atoms' causing the reading, in Merricks's sense—perhaps the atoms collectively weigh 75kg partly as a result of the person's making various conscious decisions in the past—*the person's weighing 75kg on this occasion* is not causally relevant to whether the atoms cause the reading by collectively weighting 75kg. (The person's weighing 75kg on this occasion doesn't cause the atoms to cause the reading; nor do the atoms, by collectively weighing 75kg, cause the person to cause the reading; and this isn't a case of joint causation.) So the scale-atoms' reading '75' is overdetermined by the person's weighing

75kg and the atoms' collectively weighing 75kg.[9] And the same is true of every event caused by a composite object in virtue of its straightforwardly physical properties.

Suppose, on the other hand, that Merricks denies that people ever cause things by having straightforwardly physical properties. This is a hard view to accept: we normally think, for example, that when a heavy material object is thrown at a window, the object will cause the window to break *by crashing into it*, not just by making conscious decisions and the like. Worse, this view conflicts with a piece of common sense which Merricks is very concerned to respect: the claim that people can be seen. As Merricks recognises, in order for one to see something, it must be a partial cause of one's visual experience. But obviously not just any causal relation to visual experience is enough—when one sees a person, one does not see all

[9]Would Merricks agree? On p. 58 he defines overdetermination as follows:

> An effect is overdetermined if... that effect is caused by an object; that object is causally irrelevant to whether some other—i.e. numerically distinct—object or objects cause that effect; and the other object or objects do indeed cause that effect.

Provided that the person is a cause (in virtue of past decisions) of the atoms' collectively weighing 75kg, the reading will not count as overdetermined in this sense. But Merricks's definition seems inadequate: an effect can be overdetermined by several *events* even if these events involve *objects* each of which plays some non-redundant role in causing the event. Suppose a prisoner is shot by a firing squad of two, one of whom also happens to be the judge who sentenced the prisoner to death, and the other of whom also happens to be the executioner who issued the order to fire. The death of the prisoner does not satisfy Merricks's condition for overdetermination: since the judge caused the executioner to cause the death, and the executioner caused the judge to cause the death, neither is causally irrelevant to whether the other causes the death. But intuitively, the death *is* overdetermined: the judge causes the death *by firing his gun*, and the executioner causes the death *by firing his gun*, and neither of the gun-firing events is causally relevant to whether the other one causes the death. The roles played further up the causal chain by the members of the firing squad are simply irrelevant. Moreover, systematic overdetermination of this sort—overdetermination involving events rather than objects—seems no less objectionable than systematic overdetermination of the sort Merricks focuses on.

that person's ancestors; one does not see a neurosurgeon who causes visual experiences by inserting electrodes into one's visual cortex. What is required, it seems, is that the object seen should cause one's visual experiences *by having certain visually detectable properties*: colours, shapes, textures, etc. If this is right, the claim that people don't cause anything to happen by having straightforwardly physical properties will force us to conclude that people are invisible.[10, 11]

References

Merricks, Trenton (2001). *Objects and Persons*. Oxford: Clarendon.

Sider, Theodore (MS). 'Merricks on Microphysical Supervenience.' Forthcoming in *Philosophy and Phenomenological Research*.

van Inwagen, Peter (1981). 'The Doctrine of Arbitrary Undetached Parts.' *Pacific Philosophical Quarterly* 62: 123–137.

— (1990). *Material Beings*. Ithaca: Cornell University Press.

[10]In fact, all we need to reach this conclusion is the relatively weak premise that whether one sees a certain object depends on the nature of the causal relations between that object and one's visual experience. If the only properties in virtue of which I cause anything to happen are mental properties, then as far as my causal relations to anyone's visual experiences are concerned, I might as well be an immaterial spirit. But surely immaterial spirits are invisible. Hence I am invisible, even if I have lots of causally irrelevant straightforwardly physical properties.

[11]Thanks to Jessica Moss and Ted Sider.