Fading qualia and dancing qualia

Jeff Speaks
PHIL 30304

December 4, 2018

1 Conscious machines

Could there be conscious machines? Could we design robots which not only performed many of the functions of human beings, but also were such that there is something it is like to be such a robot?

One natural way to argue that there could be conscious robots is to begin with a robot which (very) closely resembles a human being. One might try to first argue that such a being would be conscious. One might then try to extend the argument by arguing that if that sort of conscious robot is possible, then, presumably, so are others which don't so closely resemble human beings.

2 Resisting robot consciousness

Suppose that one wished to deny that robots would enjoy the same conscious experiences as the corresponding person. Then there are two options: (i) say that the robots would lack consciousness altogether, or (ii) say that the robot would have conscious experiences, but ones very different from ours.

Chalmers argues separately against these two lines of resistance. Once these are ruled out, that will lead to the principle that he calls 'organizational invariance', which says that

> 'given any system that has conscious experiences, then any system that has the same functional organization at a fine enough grain will have qualitatively identical conscious experiences.'

2.1 Fading qualia

Suppose one holds that a robot functionally identical to me would have no conscious experiences. Call this scenario 'absent qualia.' Then Chalmers argues as follows:

1. If absent qualia are empirically possible, then fading qualia are empirically possible.
2. Fading qualia are not empirically possible.

C. Absent qualia are not empirically possible.

What are fading qualia? This is best explained by working through Chalmers' example. Here, 'Robot' is a name for a robot which is functionally identical to me:

> 'Given this scenario, we can construct a series of cases intermediate between me and Robot such that there is only a very small change at each step and such that functional organization is preserved throughout. We can imagine, for instance, replacing a certain number of my neurons by silicon chips. In the first such case, only a single neuron is replaced. Its replacement is a silicon chip that performs precisely the same local function as the neuron. We can imagine that it is equipped with tiny transducers that take in electrical signals and chemical ions and transforms these into a digital signal upon which the chip computes, with the result converted into the appropriate electrical and chemical outputs. As long as the chip has the right input/output function, the replacement will make no difference to the functional organization of the system.
>
> In the second case, we replace two neighboring neurons with silicon chips. This is just as in the previous case, but once both neurons are replaced we can eliminate the intermediary, dispensing with the awkward transducers and effectors that mediate the connection between the chips and replacing it with a standard digital connection. Later cases proceed in a similar fashion...
>
> In the final case, every neuron in the system has been replaced by a chip, and there are no biochemical mechanisms playing an essential role. ...We can imagine that throughout, the internal system is connected to a body, is sensitive to bodily inputs, and produces motor movements in an appropriate way, via transducers and effectors. Each system in the sequence will be functionally isomorphic to me at a fine enough grain to share my behavioral dispositions. But while the system at one end of the spectrum is me, the system at the other end is essentially a copy of Robot.'

The key question: what is it like to be the thing undergoing the gradual replacement?

> 'Given that Robot, at the far end of the spectrum, is not conscious, it seems that one of two things must happen along the way. Either consciousness gradually fades over the series of cases, before eventually disappearing, or somewhere along the way consciousness suddenly blinks out,

> although the preceding case had rich conscious experiences. Call the first possibility Fading Qualia and the second Suddenly Disappearing Qualia.'

Problem for suddenly disappearing qualia: replacement of one neuron would lead to the complete vanishing of conscious experience. But this would make the psychophysical laws oddly discontinuous.

So suppose we go for fading qualia. This, Chalmers argues, leads to absurdity:

> '... consider a system halfway along the spectrum between me and Robot, after consciousness has degraded considerably but before it has gone altogether. Call this system Joe. What is it like to be Joe? Joe, of course, is functionally isomorphic to me. He says all the same things about his experiences as I do about mine. At the basketball game, he exclaims about the vivid bright red and yellow uniforms of the basketball players.
>
> By hypothesis, though, Joe is not having bright red and yellow experiences at all. Instead, perhaps he is experiencing tepid pink and murky brown. Perhaps he is having the faintest of red and yellow experiences. Perhaps his experiences have darkened almost to black. There are various conceivable ways in which red experiences might gradually transmute to no experience, and probably more ways that we cannot conceive. But presumably in each of these transmutation scenarios, experiences stop being bright before they vanish ...
>
> For specificity, then, let us imagine that Joe experiences faded pink where I see bright red, with many distinctions between shades of my experience no longer present in shades of his experience. Where I am having loud noise experiences, perhaps Joe is experiencing only a distant rumble. Not everything is so bad for Joe: where I have a throbbing headache, he only has the mildest twinge.
>
> The crucial point here is that Joe is systematically wrong about everything that he is experiencing. He certainly says that he is having bright red and yellow experiences, but he is merely experiencing tepid pink. If you ask him, he will claim to be experiencing all sorts of subtly different shades of red, but in fact many of these are quite homogeneous in his experience. He may even complain about the noise, when his auditory experience is really very mild. ... Joe will even judge that he has all these complex experiences that he in fact lacks. In short, Joe is utterly out of touch with his conscious experience, and is incapable of getting in touch.'

Joe is (in one sense at least) a fully rational subject, and yet is one whose judgements about the world and his own conscious experience are entirely out of step with what that conscious experience is like. Chalmers thinks that subjects of this kind are, even if not metaphysically impossible, still empirically (nomologically) impossible.

2.2 Dancing qualia

The preceding argument rules out the idea that Robot would have no conscious experiences. But it does not immediately rule out the possibility that Robot would have conscious experiences quite different than my own. To rule this out Chalmers gives an argument against the (empirical) possibility of what he calls 'dancing qualia':

> 'assume that inverted qualia are empirically possible. Then there can be two functionally isomorphic systems that are having different experiences. Suppose for the sake of illustration that these systems are me, having a red experience, and my silicon isomorph, having a blue experience (there is a small caveat about generality, which I discuss below). As before, we construct a series of cases intermediate between me and my isomorph. Here, the argument takes a different turn. ... All that matters is that there must be two points A and B in this series, such that no more than one tenth of the system is replaced between A and B, and such that A and B have significantly different experiences. To see that this must be the case, we need only consider the points at which 10 percent, 20 percent, and so on up to 90 percent of the brain has been replaced. Red and blue are sufficiently different experiences that some neighboring pairs here must be significantly different (that is, different enough that the difference would be noticeable if they were experienced by the same person); there is no way to get from red to blue by ten non-noticeable jumps.
>
> There must therefore be two systems that differ in at most one-tenth of their internal makeup, but that have significantly different experiences. ... let these systems be me and Bill. Where I have a red experience, Bill has a slightly different experience. We may as well suppose that Bill sees blue; perhaps his experience will be more similar to mine than that, but that makes no difference to the argument. The two systems also differ in that where there are neurons in some small region of my brain, there are silicon chips in Bill's brain. This substitution of a silicon circuit for a neural circuit is the only physical difference between me and Bill.
>
> The crucial step in the thought-experiment is to take a silicon circuit just like Bill's and install it in my head as a backup circuit. This circuit will be functionally isomorphic to a circuit already present in my head. We equip the circuit with transducers and effectors so that it can interact with the rest of my brain, but we do not hook it up directly. Instead, we install a switch that can switch directly between the neural and silicon circuits. Upon flipping the switch, the neural circuit becomes irrelevant and the silicon circuit takes over. We can imagine that the switch controls the points of interface where the relevant circuits affects the rest of the brain. When it is switched, the connections from the neural circuit are pushed out of the way, and the silicon circuit's effectors are attached. ...
>
> What will happen ... is that my experience will change 'before my eyes.' Where I was once experiencing red, I will now experience blue. All of a sudden, I will have a blue experience of the apple on my desk. We can

> even imagine flipping the switch back and forth a number of times, so that the red and blue experiences "dance" before my eyes. This might seem reasonable at first — it is a strangely appealing image — but something very odd is going on here. My experiences are switching from red to blue, but I do not notice any change.'

Again, Chalmers' verdict is that it is much more plausible to think that if there were a change in my experience of this sort, I would be able to notice it. And this is a good reason to think that cases like Bill are empirically impossible. (Again, there is no presumption that they are impossible simpliciter.)

3 Robot ethics

Suppose Chalmers is right that there could be conscious robots. It would not be unreasonable to think that there will be some, perhaps within our lifetimes.

Would it follow that we had moral obligations with respect to these robots?

If the robots were mentally equivalent to human beings, would we have the same moral obligations toward robots as toward human beings?

Suppose that robots had more intense sensations than human beings. Would it be morally required in some cases for one to harm a human being in order to spare a robot?

If conscious robots are possible, is it possible that we have already created one, and do not know it? How might we tell whether a robot is conscious?