The partial brain thought experiment: partial consciousness and its implications

by Dr. Jacques Mallah (jackmallah@yahoo.com)

Abstract:

The 'Fading Qualia' thought experiment of Chalmers purports to show that computationalism is very probably true even if dualism is true by considering a series of brains, with biological parts increasingly substituted for by artificial but functionally analagous parts in small steps, and arguing that consciousness would not plausibly vanish in either a gradual or sudden way. This defense of computationalism inspired an attack on computationalism by Bishop, who argued that a similar series of substitutions by parts that have the correct physical activity but not the correct causal relationships must likewise preserve consciousness, purportedly showing that 'Counterfactuals Cannot Count' and if so ruining a necessary condition for computation to meaningfully distinguish between physical systems. In this paper, the case in which a series of parts are simply removed and substituted for only by imposing the correct boundary conditions to exactly preserve the functioning of the remaining partial brain is described. It is argued that consciousness *must* gradually vanish in this case, not by fading but by becoming more and more partial. This supports the non-centralized nature of consciousness, tends to support the plausibility of physicalism against dualism, and provides the proper counterargument to Bishop's contention. It also provides an avenue of attack against the "Fading Qualia" argument for those who remain dualists.

Fading Qualia:

Chalmers [1] describes the Fading Qualia thought experiment as follows:

"In this thought-experiment, we assume for the purposes of *reductio* that absent qualia are empirically possible. It follows that there can be a system with the same functional organization as a conscious system (such as me), but which lacks conscious experience entirely due to some difference in non-organizational properties. Without loss of generality, suppose that this is because the system is made of silicon chips rather than neurons. Call this functional isomorph Robot. …

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. …

The question arises: *What is it like to be the systems in between?* ...

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*.

On the second hypothesis, the replacement of a single neuron could be responsible for the vanishing of an entire field of conscious experience. ... This seems antecedently implausible, if not entirely bizarre. …

This leaves the first hypothesis, Fading Qualia. To get a fix on this hypothesis, 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. …

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 (otherwise we are left with the problem of Suddenly Disappearing Qualia). Similarly, there is presumably a point at which subtle distinctions in my experience are no longer present in an intermediate system's experience; if we are to suppose that all the distinctions in my experience are present right up until a moment when they simultaneously vanish, we are left with another version of Suddenly Disappearing Qualia.

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. …

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. … In short, Joe is utterly out of touch with his conscious experience, and is incapable of getting in touch.

There is a significant implausibility here. This is a being whose rational processes are functioning and who is in fact *conscious*, but who is completely wrong about his own conscious experiences. Perhaps in the extreme case, when all is dark inside, it is reasonable to suppose that a system could be so misguided in its claims and judgments - after all, in a sense there is nobody in there to be wrong. But in the intermediate case, this is much less plausible. In every case with which we are familiar, conscious beings are generally capable of forming accurate judgments about their experience, in the absence of distraction and irrationality. For a sentient, rational being that is suffering from no functional pathology to be so systematically out of touch with its experiences would imply a strong dissociation between consciousness and cognition. We have little reason to believe that consciousness is such an ill-behaved phenomenon, and good reason to believe otherwise.”

The Fading Qualia argument in favor of functionalism is aimed at dualists, like Chalmers himself. (Physicalists would tend to believe in functionalism already.) Chalmers believes in Qualia, and believes that there are simple laws of nature connecting them to the physical world. He believes it would be merely unlikely (but, in his view, concievable) that the natural laws could be such that for example a person's qualia are based on what is happening in his right foot, rather than what his brain is doing. In the same way, he thinks it would be unlikely for a person to have wrong beliefs about his own qualia. Thus, since the brain replacement with silicon preserves beliefs he thinks it will also be likely to preserve qualia.

For a reductive functionalist such as myself, who finds the logical possibility of a not-conscious-in-the same-sense-we-are zombie Robot – a system that would have the same information that we do about the color red, since epiphenomal qualia could not by definition influence our mathematical properties such as information and patterns of thought - not plausible in the first place, the Fading Qualia argument has little direct relevance. However, as a practical matter, it is convenient for the reductive functionalist to have this argument presented to dualists, since they may then become allies on the narrow issue of functionalism while keeping their dualism, as Chalmers has.

Removing Counterfactuals:

It is a requirement of any viable criterion for implementation of a computation by a physical system (Chalmers [2], [Mallah]) that the proper behavior as specified by that computation would occur if any component of a computer were to receive a different (counterfactual) input.

[Bishop] presented an interesting variation on the Fading Qualia argument. For simplicity, one can assume that the brain has been replaced already by an artificial computer, as the Fading Qualia argument would have us believe preserves consciousness. It is then easy enough to then gradually replace the components, which base their next state on their current state and inputs, with components that merely pass through a predetermined sequence of states; they don't have the normal sensitivity to counterfactual inputs.

Would the intermediate systems be conscious, and if so, what would that be like? If the logic of the Fading Qualia argument applies here – if there is no plausible way to pass from normal qualia to no qualia for such a series of partially replaced systems – then one would have to conclude that they would have normal consciousness. But then counterfactuals could not be part of the requirements for computation, and (as Bishop argues) this would ruin the ability to distinguish between computers of interest and systems such as rocks that should not be seen as performing those computations.

Chalmers gives the following comments on his web site:

http://consc.net/responses.html#bishop

"… He runs a version of the fading qualia argument, suggesting that we can remove unused state-transitions one-by-one, thus removing counterfactual sensitivity, while (he argues) preserving consciousness.

… this process will gradually transform a counterfactually-sensitive system into a "wind-up" system that implements just one run. This plausibly will affect the system's cognitive states (such as beliefs), gradually destroying them, so the fading qualia argument (which relies on preserving cognitive states) doesn't apply. …

Maybe it is initially hard to see just how mere counterfactual sensitivity can affect an intrinsic property such as consciousness. But it's hard to see how any physical property can affect consciousness. …"

In short: In the case of "wind-up" substitution, beliefs will gradually dissappear along the series of substitutions, and if they do then so can qualia.

This reply is reasonable given that the basis of the Fading Qualia argument is that it is implausible for a conscious system to be mistaken about its own qualia. If counterfactuals are required for cognitive states, then removing them removes the beliefs. However, for certain intermediate cases such as those in which the higher-order belief centers are preserved while other areas of the brain have been substituted, it does not seem an adequate answer if one maintains that the system would not be mistaken about its own qualia.

In addition, the commited anti-functionalist dualist then can reply to the original Fading Qualia argument in the same way. As the last paragraph quoted above might suggest, any property can be substituted for counterfactual sensitivity in the argument. Chalmers admits that *beliefs plausibly can* gradually dissappear along a series of substitutions – and that if they do, qualia plausibly can vanish too. A dualist might believe that substituting biological components with artificial functional analogues could affect the system's cognitive states (such as beliefs), gradually destroying them, while preserving its behavior. After all, dualists believe that a functionally identical system without qualia (a zombie) would also have no beliefs. Therefore, the Fading Qualia argument has no force.

The Partial Brain Argument (PBA):

The possibility of substituition with components lacking counterfactuals shows that the computationalist cannot take the logic of the Fading Qualia argument too far. I will now argue that the basis for the argument is in fact mistaken: in fact, it is *not* always implausible for a conscious system to be mistaken about its own "qualia" (even if such things as qualia exist). This will come as no surprise to the eliminativists, who maintain that the very belief that there are any qualia is such a mistake.

Substitution with the dummy components is in fact a red herring, leading one to focus on those added components. A more revealing thought experiment is simply to remove components, leaving behind a series of smaller and smaller Partial Brains, with (highly specific and normally improbable) boundary conditions imposed on them such that the functioning within the Partial Brains is identical to what it would have been if the brain had been left intact.

What would it be like to be such a partial brain? Some important features seem obvious: it is not plausible that as we let the partial brain decrease in size, consciousness would vanish suddenly. Nor is it possible that consciousness will remain unchanged.

Therefore, progressively less and less of its consciousness will remain. In a sense it can't notice this - it beliefs will dissappear as certain parts of the brain vanish, but they won't otherwise change - but that just means its beliefs will become more wrong until they vanish. For example, if the higher order belief center remains intact but the visual system is gone, the partial brain will believe it is experiencing vision but will in fact not be.

While the partial brain's consciouness would be *indistinguishable for it* from the normal brain's consciousness to whatever extent it could still *evaluate the question*, it would not be the same. There would be less of it. This contradicts the homunculus fallacy; there is no unified mind's eye.

Consciousness can be tricky to think about, but knowledge will serve to illustrate what may happen. Consider an example of doing addition. Compare the knowledge of an intact brain and a partial brain, where the partial brain has input to mimic part of the intact brain as usual. This example is not based on a specific model of neural architecture, but seems a reasonable example that could be produced by removing a specific set of components.

I will assume that the person's brain can add 2-digit numbers in his head.

Intact Brain: "23 plus 45 equals 68. I didn't need to carry the one."

Partial Brain "2_ plus 4_ equals 6_. I didn't need to carry the one."

Does Partial Brain know something is missing? No, it's just doing its job, 'assured' by the inputs that all is well in what would normally be the rest of the brain. It thinks that a 2-digit addition has been performed. Is its consciousness the same as that of Intact Brain? Of course not; it knows nothing of the "3 plus 5 equals 8" business.

It is possible that the remaining consciousness would become more "fuzzy" as parts are removed rather than simply being present or not present. Nonetheless, the brain would not detect this; its functioning would be such that it could not; it would not have the resources to do so. The two important features of the argument are the gradual loss of consciousness along the series of decreasing partial brains, and the partials brains' inability to know that they are not just part of a full normal brain.

Conclusions:

The PBA shows that consciousness can indeed vanish along a near-continuous progression of brain types *while the changes nontheless remain " undetectable" as far as the remaining brain is able to evaluate*. For the computationalist, Bishops' series of substitutions with increasing amounts of dummy components is equivalent to a series of partial brains of decreasing size, and therefore his argument (that consciousness must be preserved) has no force.

The PBA also has implications for the way we should think about the brain's knowledge of qualia. Even a partial brain with no visual system would still think it has visual qualia, contradicting the unified mind's-eye (homunculous) fallacy.

Why then should a normal brain's belief that it has visual qualia be taken to imply that it does in fact have them? It is true that this does not mean that it can't; after all, the partial brain in the addition example was mistaken about performing a 2-digit addition, but the normal brain was not mistaken about it. However, if some aspects of qualia must be mysterious and beyond physical possibility, then given the fact that a brain *can* be wrong about its own qualia, surely the burden of proof rests on those making such assertions.

There is no doubt that the brain tends to think it has qualia, and clearly, unless we are partial brains subject to unlikely boundary conditions, there is something that plays the functional role that qualia intuitively seem to play. I will call this the Functional Aspects of Qualia (FAQ), as distinct from the dualists' hypothesized Epiphenomenal Aspects of Qualia (EAQ). Like the partial brain was mistaken even about FAQ, the normal brain could be mistaken about EAQ.

Functional explanation of the FAQ would account for all of the known facts without any need for dualism. One of the most fundament aspects of qualia is that they don't seem epiphenomenal; we very much seem to be able to be influenced by them and to report on our observations of them, which by definition only FAQ could account for and not EAQ. Anything that could prompt a brain to investigate that thing is by definition not epiphenomenal and must therefore admit of a functionalist explanation.

Finally, diehard anti-functionalist dualists could counter the Fading Qualia argument by supposing that only the non-artificial part of the brain gives rise to any consciousness, so that the consciousness could vanish gradually not by fading but by becoming more and more partial as in the PBA.

Acknowledgement:

The author would like to thank Kory Heath for an email discussion that helped inspire this paper.

References:

Mark Bishop, Counterfactuals cannot count: A rejoinder to David Chalmers. ***Consciousness and Cognition*****, 11:642-52, 2002.**
http://www.doc.gold.ac.uk/~mas02mb/Selected%20Papers/2002%20Counterfactuals%20cant%20count.pdf

Chalmers' response to Bishop: http://consc.net/responses.html#bishop

David J. Chalmers [1], Absent Qualia, Fading Qualia, Dancing Qualia. Published in ***Conscious Experience*****, 1995.**
http://consc.net/papers/qualia.html

David J. Chalmers [2], Does a Rock Implement Every Finite-State Automaton? ***Synthese*** **108:309-33, 1996.**
http://consc.net/papers/rock.html

Jacques Mallah, The Many Computations Interpretation (MCI) of Quantum Mechanics. ArXiv manuscript.* http://arxiv.org/abs/0709.0544

* Note: I will post a manuscript containing a slightly revised account of implementation and the MCI in 2009.