Absent Qualia, Fading Qualia, Dancing Qualia

David J. Chalmers
Department of Philosophy
Washington University
St. Louis, Missouri 63130
dave@twinearth.wustl.edu

June 18, 1994

1 The principle of organizational invariance

It is widely accepted that conscious experience has a physical basis. That is, the properties of experience (phenomenal properties, or qualia) systematically depend on physical properties according to some lawful relation. There are two key questions about this relation. The first concerns the strength of the laws: are they logically or metaphysically necessary, or are they merely nomologically necessary? This question about the strength of the psychophysical link is the basis for debates over physicalism and property dualism. The second question concerns the nature of the laws: precisely how do phenomenal properties depend on physical properties? What sort of physical properties enter into the laws' antecedents, for instance; consequently, what sort of physical systems can give rise to conscious experience? It is this second question that I will address in this paper.

To put the issue differently, even once it is accepted that experience arises from physical systems, the question remains open: in virtue of what sort of physical properties does conscious experience arise? Some property that brains can possess, presumably, but it is far from obvious what sort of property this is. Some have suggested biochemical properties; some have suggested quantum-mechanical properties; many have professed uncertainty. A natural suggestion is that when experience arises from a physical system, it does so in virtue of the system's *functional organization*. On this view, the chemical and indeed the quantum substrate of the brain is not directly relevant to the existence of consciousness. What counts is rather the brain's abstract causal organization, an organization that might be realized in many different physical substrates.

In this paper I defend this view. Specifically, I defend a principle of *organizational invariance*, holding that experience is invariant across systems with the same fine-grained functional organization. More precisely, the principle states 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. A full specification of a system's fine-grained functional organization will fully determine any conscious experiences that arise.

To clarify this, we must first clarify the notion of functional organization. This is best understood as the *abstract pattern of causal interaction* between the components of a system, and perhaps between these components and external inputs and outputs. A functional organization is determined by specifying (1) a number of abstract components, (2) for each component, a number of different possible states, and (3) a system of dependency relations, specifying how the states of each component depends on the previous states of all components and on inputs to the system, and how outputs from the system depend on previous component states. Beyond specifying their number and their dependency relations, the nature of the components and the states is left unspecified.

A physical system *realizes* a given functional organization when the system can be divided into an appropriate number of physical components each with the appropriate number of possible states, such that the causal dependency relations between the components of the system, inputs, and outputs precisely reflect the dependency relations given in the specification of the functional organization. A given functional organization can be realized by diverse physical systems. For example, the organization realized by the brain at the neural level might in principle be realized by a silicon system.

A physical system has functional organization at many different levels, depending on how finely we individuate its parts and on how finely we divide the states of those parts. At a coarse level, for instance, it is likely that the two hemispheres of the brain can be seen as realizing a simple two-component organization, if we choose appropriate interdependent states of the hemispheres. It is generally more useful to view cognitive systems at a finer level, however. For our purposes I will always focus on a level of organization fine enough to determine the behavioral capacities and dispositions of a cognitive system. This is the role of the "fine enough grain" clause in the statement of the organizational invariance principle; the level of organization relevant to the application of the principle is one fine enough to determine a system's behavioral dispositions. In the brain, it is likely that the neural level suffices, although a coarser level might also work. For the purposes of illustration I will generally focus on the neural level of organization of the brain, but the arguments generalize.

Strictly speaking, for the purposes of the invariance principle we must require that for two systems to share their functional organization, they must be in corresponding states at the time in question; if not for this requirement, my sleeping twin might count as sharing my organization, but he certainly does not share my experiences. When two systems share their organization at a fine enough grain (including the requirement that they be in corresponding states), I will say that they are *functionally isomorphic* systems, or that they are *functional isomorphs*. The invariance principle holds that any functional isomorph of a conscious system has qualitatively identical experiences.

2 Absent qualia and inverted qualia

The principle of organizational invariance is far from universally accepted. Some have thought it likely that for a system to be conscious it must have the right sort of biochemical makeup; if so, a metallic robot or a silicon-based computer could never have experiences, no matter what its causal organization. Others have conceded that a robot or a computer might be conscious if it were organized appropriately, but have held that it might nevertheless have experiences quite different from the kind that we have. These two sorts of objections are often known as the *absent qualia* and *inverted qualia* objections to broadly functionalist theories of consciousness.

Arguments for the absent qualia objection usually consist in the description of a system that realizes whatever functional organization might be specified, but that is so outlandish that it is natural to suppose that it must lack conscious experience. For example, Ned Block[1] points out that the functional organization of the brain might be instantiated by the population of China, if they were organized appropriately, and argues that it is bizarre to suppose that this would somehow give rise to a group mind. In a similar way, John Searle[2] notes that a given organization might be realized by "a sequence of water-pipes, or a set of wind-machines", but argues that these systems would not be conscious.

Arguments for the inverted qualia objection often proceed from considerations about experiences of color. According to this line of argument, it is possible that a system might make precisely the same color discriminations that I do, but that when confronted by red objects it has the kind of experience that I have when confronted by blue objects. Further, it is argued that this might happen even when the systems are functionally isomorphic.[3] If this argument succeeds, then even if the appropriate functional organization suffices for the existence of conscious experience, it does not determine their specific nature. Instead, the specific nature of experiences must be dependent on non-organizational properties, such as specific neurophysiological properties.

Sometimes these arguments are intended as arguments for "possibility" only in some weak sense, such as logical or metaphysical possibility. These less ambitious forms of the arguments are the most likely to be successful. It seems difficult to deny that the absent qualia and inverted qualia scenarios are at least intelligible. With the aid of certain assumptions about possibility, this intelligibility can be extended into an argument for the logical and perhaps the metaphysical possibility of the scenarios. If successful, even these less ambitious arguments would suffice to refute some strong versions of functionalism, such as analytic functionalism and the view that phenomenal properties are identical to functional properties.

In the present paper I am not concerned with the logical or metaphysical possibility of these scenarios, however, but rather with their *empirical* (or nomological) possibility. The mere logical or metaphysical possibility of absent qualia is compatible with the claim that in the actual world, whenever the appropriate functional organization is realized, conscious experience is present. By analogy: many have judged it logically possible that a *physical* replica of a conscious system

[1] N. Block, "Troubles with functionalism", in (Block, ed) *Readings in the Philosophy of Psychology* (Cambridge, MA: Harvard University Press, 1980).

[2] J.R. Searle, "Minds, brains, and programs", *Behavioral and Brain Sciences* 3(1980):417-57.

[3] S. Shoemaker, "The inverted spectrum", *Journal of Philosophy* 79 (1982):357-81; T. Horgan, "Functionalism, qualia, and the inverted spectrum", *Philosophy and Phenomenological Research* 44 (1984):453-69.

might lack conscious experience, while not wishing to deny that in the actual world, any such replica will be conscious. It is the claim about empirical possibility that is relevant to settling the issue at hand, which concerns a possible lawful relation between organization and experience. Mere intelligibility does not bear on this, any more than the intelligibility of a world without relativity can falsify Einstein's theory.

On the question of empirical possibility, the success of the absent qualia and inverted qualia arguments is unclear. To be sure, many have found it *counterintuitive* that the population of China might give rise to conscious experience if organized intuitively. The natural reply, however, is that it seems equally counterintuitive that a mass of 10^{11} appropriately organized neurons should give rise to consciousness, and yet it happens. Intuition is unreliable as a guide to empirical possibility, especially where a phenomenon as perplexing as conscious experience is concerned. If a brain can do the job of enabling conscious experience, it is far from obvious why an appropriately organized population, or indeed an appropriate organized set of water-pipes, could not.

The debate over absent and inverted qualia tends to produce a stand-off, then. Both proponents and opponents claim intuitions in support of their positions, but there are few grounds on which to settle the debate between them. Both positions seem to be epistemic possibilities, and due to the notorious difficulties in collecting experimental evidence about conscious experience, things might seem likely to stay that way.

I believe that the stand-off can be broken, and in this paper I will offer considerations that offer strong support to the principle of organizational invariance, suggesting that absent qualia and inverted qualia are empirically impossible. These arguments involve thought-experiments about gradual neural replacement, and take the form of a *reductio*. The first thought-experiment demonstrates that if absent qualia are possible, then then a phenomenon involving what I will call *Fading Qualia* is possible; but I will argue that we have good reason to believe that Fading Qualia are impossible. The second argument has broader scope and is more powerful, demonstrating that if absent qualia *or* inverted qualia are possible, then a phenomenon involving what I will call *Dancing Qualia* is possible; but I will argue that we have even better reason to believe that Dancing Qualia are impossible. If the arguments succeed, we have good reason to believe that absent and inverted qualia are impossible, and that the principle of organizational invariance is true.

These arguments do not constitute conclusive *proof* of the principle of organizational invariance. Such proof is generally not available in the domain of conscious experience, where for familiar reasons one cannot even disprove the hypothesis that there is only one conscious being. Nevertheless, in the absence of proof we can bring to bear arguments for the plausibility and implausible of different possibilities, and not all possibilities end up equal. These thought-experiments constitute strong plausibility arguments for the principle of organizational invariance. If an opponent wishes to hold onto the possibility of absent or inverted qualia, she can do so only at significant cost.

3 Fading Qualia

The scenario that I will present here is a relatively familiar one,[4] but a correct analysis of it is important, and it is a necessary preliminary to the more powerful second argument. 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. The causal patterns in Robot's processing system are the same as mine, but there is nothing it is like to be 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. 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, with larger and larger groups of neighboring neurons replaced by silicon chips. Within these groups, biochemical mechanisms have been dispensed with entirely, except at the periphery. 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. (I abstract away here from detailed issues concerning whether, for instance, glial cells play a non-trivial role; if they do, they will be components of the appropriate functional organization, and will be replaced also.)

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.

[4]Neural replacement scenarios along the lines discussed in this section are discussed by Z. Pylyshyn, "The 'causal power' of machines", *Behavioral and Brain Sciences* 3 (1980):442-4; S. Savitt, "Searle's demon and the brain simulator reply", *Behavioral and Brain Sciences* 5 (1982):342-3; T. Cuda, "Against neural chauvinism", *Philosophical Studies* 48 (1985):111-27; and J.R. Searle, *The Rediscovery of the Mind*, Chapter 3 (Cambridge, MA: MIT Press).

To fix imagery, imagine that as the first system I am having rich conscious experiences. Perhaps I am at a basketball game, surrounded by shouting fans, with all sorts of brightly-colored clothes in my environment, smelling the delicious aroma of junk food and perhaps suffering from a throbbing headache. Let us focus in particular on the bright red and yellow experiences I have when I watch the players' uniforms. ("Red experience" should be taken as shorthand for "color experience of the kind I usually have when presented with red objects", and so on throughout.) The final system, Robot, is in the same situation, processing the same inputs and producing similar behavior, but by hypothesis is experiencing nothing at all.

The question arises: *What is it like to be the systems in between?* For those systems intermediate between me and Robot, what, if anything, are they experiencing? As we move along the spectrum of cases, how does conscious experience vary? Presumably the very early cases have experiences much like mine, and the very late cases have little or no experience, but what of the cases in the middle?

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. If so, we could switch back and forth between a neuron and its silicon replacement, with a field of experience blinking in and out of experience on demand. This seems antecedently implausible, if not entirely bizarre. If Suddenly Disappearing Qualia were possible, there would be brute discontinuities in the laws of nature unlike those we find anywhere else. Any specific point for qualia to suddenly disappear (50 percent neural? 25 percent?) would be entirely arbitrary. We might even run the experiment at a finer grain within the neuron, so that ultimately the replacement of a few molecules produces a sudden disappearance of experience. As always in these matters, the hypothesis cannot be disproved, but there is little reason to take it seriously.

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. 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 even 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 the 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. 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 he is only experiencing a distant rumble. Worse, on a functional construal of judgment, 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.

This seems to be vastly implausible. 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.

To be sure, Fading Qualia may be *logically* possible. Arguably, there is no contradiction in the notion of a system that is so wrong about its experiences. But logical possibility and empirical possibility are different things. One of the most salient empirical facts about conscious experience is that when a conscious being has experiences, it is at least capable of forming reasonable judgments about those experiences. Perhaps there are some cases where judgment is impaired due to a malfunction in rational processes, but this is not such a case. Joe's processes are *functioning* as well as mine—by hypothesis, he is functionally isomorphic. It is just that he happens to be completely misguided about his experience.

There are various cases of fading qualia in everyday life, of course. Think of what happens when one is dropping off to sleep; or think of moving back along the evolutionary chain from people to trilobites. In each case, as we move along a spectrum of cases, conscious experience gradually fades away. But in each of these cases, the fading is accompanied by a corresponding change in *functioning*. When I become drowsy, I do not believe that I am wide awake and having intense experiences (unless perhaps I start to dream, in which case I very likely *am* having intense experiences). The lack of richness in a dog's experience of color accompanies a corresponding lack of discriminatory power in a dog's visual mechanisms. These cases are quite unlike the case under consideration, in which experience fades while functioning stays constant. Joe's mechanisms can still discriminate subtly different wavelengths of light, and he certainly judges that such discriminations are reflected in his experience, but we are to believe that his experience does not reflect these discriminations at all.

Searle[5] discusses a thought-experiment like this one, and suggests the following possibility:

> ...as the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when the doctors test your vision, you hear them say, "We are holding up a red object in front of you; please tell us what you see." You want to cry out, "I can't see anything. I'm going totally blind." But you hear your voice saying in a way that is completely out of your control, "I see a red object in front of me." If we carry the thought-experiment out to the limit, we get a much more depressing result than last time. We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same.

Here, Searle embraces the possibility of Fading Qualia, but suggests that such a system need not be systematically mistaken in its beliefs about its experience. The system might have true beliefs about its experience, but beliefs that are impotent to affect its behavior.[6]

It seems that this possibility can be ruled out, however. There is simply no room in the system for any new beliefs to be formed. Unless one is a dualist of a very strong variety, beliefs must be reflected in the functioning of a system—*perhaps* not in behavior, but at least in some process. But this system is identical to the original system (me) at a fine grain. There is no room for new beliefs like "I can't see anything", new desires like the desire to cry out, and other new cognitive states such as amazement. Nothing in the physical system can correspond to that amazement. There is no room for it in the neurons, which after all are identical to a subset of the neurons supporting the usual beliefs; and Searle is surely not suggesting that the silicon replacement is itself supporting the new beliefs! Failing a remarkable, magical interaction effect

[5] Searle, *The Rediscovery of the Mind*, pp. 66-67.

[6] Searle also raises the possibility that upon silicon replacement, the system might be slowly reduced to paralysis, or have its functioning otherwise impaired. Such a scenario is irrelevant to the truth of the invariance principle, however, which applies only to systems with the appropriate functional organization. If a silicon system does not duplicate the organization of the original system, the principle does not even come into play.

between neurons and silicon—and one that does not manifest itself anywhere in processing, as organization is preserved throughout—such new beliefs will not arise.

While it might just seem plausible that an organization-preserving change from neurons to silicon might twist a few experiences from red to blue, a change in beliefs from "Nice basketball game" to "I seem to be stuck in a bad horror movie!" is of a different order of magnitude. If such a major change in cognitive contents were not mirrored in a change in functional organization, cognition would float free of internal functioning like a disembodied mind. If the contents of cognitive states supervened on physical states at all, they could do so only by the most arbitrary and capricious of rules (if this organization in neurons, then "pretty colors!"; if this organization in silicon, then "Alas!").

It follows that the possibility of Fading Qualia requires either a bizarre relationship between belief contents and physical states, or the possibility of beings that are massively mistaken about their own conscious experiences despite being fully rational. Both of these hypotheses are significantly less plausible than the hypothesis that rational conscious beings are generally correct in their judgments about their experiences. A much more reasonable hypothesis is therefore that when neurons are replaced, qualia do not fade at all. A system like Joe, in practice, will have conscious experiences just as rich as mine. If so, then our original assumption was wrong, and the original isomorph, Robot, has conscious experiences.

This thought-experiment can be straightforwardly extended to other sorts of functional isomorphs, including those that differ in shape, size, and physical makeup. All we need do is construct a sequence of intermediate cases, each with the same functional organization. In each case the conclusion is the same. If such a system is not conscious, then there exists an intermediate system that is conscious, has faded experiences, and is completely wrong about its experiences. Unless we are prepared to accept this massive dissociation between consciousness and cognition, the original system must have been conscious after all.

We can even extend the reasoning straightforwardly to the case of an appropriately-organized population: we simply need to imagine neurons replaced one-by-one with tiny homunculi, ending up with a network of homunculi that is essentially equivalent to the population controlling a robot. (If one objects to tiny homunculi, they can be external and of normal size, as long as they are equipped with appropriate radio connections to internal function when necessary.) Precisely the same considerations about intermediate cases arise. One can also imagine going from a multiple-homunculi case to a single-homunculus case, yielding something like Searle's "Chinese room" example. We need only suppose that the homunculi gradually "double up" on their tasks, leaving written records of the state of each component, until only a single homunculus does all the work. If the causal organization of the original system is preserved, even if it is only among a system of marks on paper, then the same arguments suggest that the system will have experiences. (Of course, we should not expect the homunculus itself to have the experiences; it is merely acting as a sort of causal facilitator.)

If Absent Qualia are possible, then Fading Qualia are possible. But I have argued above that it is very unlikely that Fading Qualia are possible. It follows that it is very unlikely that absent qualia are possible.

Some might object that these thought-experiment are the stuff of science fiction rather than the stuff of reality, and point out that this sort of neural replacement would be quite impossible in practice. But although it might be technologically impossible, there is no reason to believe that the neural replacement scenario should be nomologically impossible. We already have prosthetic arms and legs. Prosthetic eyes lie within the foreseeable future, and there seems to be no reason why a prosthetic neuron is impossible in principle. Even if it were impossible for some technical reason (perhaps there would not be enough room for a silicon replacement to do its work?), it is entirely unclear what bearing this technical fact would have on the principled force of the thought-experiment. There will surely be *some* systems between which gradual replacement is possible; will the objector hold that the invariance principle holds for those systems, but no other? If so, the situation seems quite arbitrary; if not, then there must be some deeper objection available.

Some might be temped to object that no silicon replacement could perform even the local function of a neuron, perhaps because neural function is uncomputable. There is little evidence for this, but it should be noted that even if it is true, it does not affect the argument for the invariance principle. If silicon really could not even duplicate the *function* of a neural system, then a functional isomorph made of silicon would be impossible, and the assessment of silicon systems would simply be irrelevant to the invariance principle. To evaluate the truth of the principle, it is only functionally isomorphic systems that are relevant.

Another objection notes that there are actual cases in which subjects are seriously mistaken about their experiences. For example, in cases of blindness denial, subjects believe that they are having visual experiences when they likely have none. In these cases, however, we are no longer dealing with fully rational systems. In systems whose belief-formation mechanisms are impaired, anything goes. Such systems might believe that they are Napoleon, or that the moon is pink. My "faded" isomorph Joe, by contrast, is a fully rational system, whose cognitive mechanisms are functioning just as well as mine. In conversation, he seems perfectly sensible. We cannot point to any unusually poor inferential connections between his beliefs, or any systematic psychiatric disorder that is leading his thought processes to be biased toward faulty reasoning. Joe is an eminently thoughtful, reasonable person, who exhibits none of the confabulatory symptoms of those with blindness denial. The cases are therefore disanalogous. The plausible claim is not that no system can be massively mistaken about its experiences, but that no rational system whose cognitive mechanisms are unimpaired can be so mistaken. Joe is certainly a rational system whose mechanisms are working as well as mine, so the argument is unaffected.

Some object that this argument has the form of a Sorites or "slippery-slope" argument, and observe that these arguments are notoriously suspect. Using a Sorites argument, we can "show" that even a grain of a sand is a heap; after all, a million grains of sand form a heap, and if we take a single grain away from a heap we still have a heap. This objection is based on a superficial reading of the thought-experiment, however. Sorites arguments gain their force by ignoring the fact that some apparent dichotomy is in fact a continuum; there are all sorts of vague cases between heaps and non-heaps, for instance. The Fading Qualia argument, by contrast, explicitly accepts the possibility of a continuum, but argues that intermediate cases are impossible for independent reasons. The argument is therefore not a Sorites argument.

Ultimately, the only tenable way for an opponent of organizational invariance to respond to this argument is to bite the bullet and accept the possibility of Fading Qualia, and the consequent possibility that a rational conscious system might be massively mistaken about its experience. This position is unattractive in its implication of a dissociation between consciousness and cognition, and seems much less attractive than the alternative, other things being equal; but it is a more tenable route than any of the objections above. The argument to follow provides an even more powerful case against the possibility of absent qualia, however, so opponents of organizational invariance cannot rest easily.

4 Dancing Qualia

If the Fading Qualia argument succeeds, it establishes that functional isomorphs of a conscious system will have conscious experience, but it does not establish that isomorphs have the *same* sort of conscious experience. The preceding argument has no bearing on the possibility of inverted qualia. For all that has gone before, where I am having a red experience, my silicon functional isomorph might be having a blue experience, or some other kind of experience that is quite foreign to me.

One might think that the Fading Qualia argument could be directly adapted to provide an argument against the possibility of inverted qualia, but that strategy fails. If I have a red experience and my functional isomorph has a blue experience, there is no immediate problem with the idea of intermediate cases with intermediate experiences. These systems might be simply suffering from milder cases of qualia inversion, and are no more problematic than the extreme case. These systems will not be systematically wrong about their experiences. Where they claim to experience distinctions, they may really be experiencing distinctions; where they claim to be having intense experiences, they may still be having intense experiences. To be sure, the experiences they call "red" differ from those I call "red", but this is already an accepted feature of the usual inversion case. The difference between these cases and the Fading Qualia cases is that these cases preserve the *structure* of experience throughout, so that their existence implies no implausible dissociation between experience and cognition.

Nevertheless, a good argument against the possibility of inverted qualia can be found in the vicinity. Once again, for the purposes of *reductio*, 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 here, which I will discuss below).

As before, we construct a series of cases intermediate between me and my isomorph. Here, the argument takes a different turn. We need not worry about the *way* in which experiences change as we move along the series. 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%, 20%, and so on up to 90% 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. For the purposes of illustration, 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. (We might imagine that the transducers for both circuits are attached the entire time, so that the state of both circuits evolves appropriately, but so that only one circuit at a time plays a role in processing. We could also run a similar experiment where both transducers and effectors are disconnected, to ensure that the backup circuit is entirely isolated from the rest of the system. This would change a few details, but the moral would be the same.)

Immediately after flipping the switch, processing that was once performed by the neural circuit is now performed by the silicon circuit. The flow of control with the system has been redirected. However, my functional organization is exactly the same as it would have been if we had not flipped the switch. The only relevant difference between the two cases is the physical makeup of one circuit within the system. There is also a difference in the physical makeup of another "dangling" circuit, but this is irrelevant to functional organization, as it plays no role in affecting other components of the system and directing behavior.

What happens to my experience when we flip the switch? Before installing the circuit, I was experiencing red. After we install it but before we flip the switch, I will presumably still be experiencing red, as the only difference is the addition of a circuit that is not involved in processing in any way; for all the relevance it has to my processing, I might as well have eaten it. *After* flipping the switch, however, I am more or less the same system as Bill. The only difference between Bill and me now is that I have a causally irrelevant neural circuit dangling from the system (we might even imagine that the circuit is destroyed when the switch is flipped). Bill, by hypothesis, was enjoying a blue experience. After the switch, then, I will have a blue experience too.

What will happen, then, 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*. Even as we flip the switch a number of times and my qualia dance back and forth, I will simply go about my business, not noticing anything unusual. My functional organization remains normal throughout. In particular, my functional organization after flipping the switch evolves just as it would have if the switch had not been flipped. There is no special difference in my behavioral dispositions. I am not suddenly disposed to say "Hmm! Something strange is going on!". There is no room for a sudden start, for an exclamation, or even for a distraction of attention. My cognitive organization is just as it usually is, and in particular is precisely as it would have been had the switch not been flipped.

Certainly, on any functional construal of judgment, it is clear that I do not make any novel judgments due to the flip. Even if one were to dispute a functional account of judgment, it is is extremely implausible that a simple organization-preserving replacement of a neural circuit by a silicon circuit could be responsible for the addition of significant new judgments such as "My qualia just flipped". As in the case of Fading Qualia, there is simply no room for such a change to take place, unless it is in an accompanying Cartesian disembodied mind.

We are therefore led once more into a *reductio ad absurdum*. It seems entirely implausible to suppose that my experiences could change in such a significant way, even with me paying full attention, without my being able to notice the change. It would suggest once again an extreme dissociation between consciousness and cognition. If this kind of thing could happen, then psychology and phenomenology would be radically out of step, much further out of step than even the Fading Qualia scenario would imply.

This "Dancing Qualia" scenario may be logically possible (although the case is so extreme that it seems *only just* logically possible), but that does not mean we should take it seriously as an empirical possibility, any more than we should take seriously the possibility that the world was created five minutes ago. As an empirical hypothesis, it is far more plausible that when one's experiences change significantly, then as long as one is rational and paying attention, one should be able to notice the change. If not, then consciousness and cognition are tied together by only the most slender of threads.

Indeed, if we are to suppose that Dancing Qualia are empirically possible, we are led to a worrying thought: they might be *actual*, and happening to us all the time. The physiological properties of our functional mechanisms are constantly changing. The functional properties of the mechanisms are reasonably robust; one would expect that this robustness would be ensured by evolution. But there is no adaptive reason for the non-functional properties to stay constant. From moment to moment there will certainly be changes in low-level molecular properties. Properties such as position, atomic makeup, and so on can change while functional role is preserved, and such change is almost certainly going on constantly.

If we allow that qualia are dependent not just on functional organization but on implementational details, it may well be that *our* qualia are in fact dancing before our eyes all the time. There seems to be no principled reason why a change from neurons to silicon should make a difference while a change in neural realization should not; the only place to draw a *principled* line is at the functional level. The reason why we doubt that such dancing is taking place in our own cases is that we accept the following principle: when one's experiences change significantly, one can notice the change. If we were to accept the possibility of Dancing Qualia in the original case, we would be discarding this principle, and it would no longer be available as a defense against skepticism even in the more usual cases.

It is not out of the question that we could actually perform such an experiment. Of course the practical difficulties would be immense, but at least in principle, one could install such a circuit in me and *I* could see what happened, and report it to the world. But of course there is no point performing the experiment: we know what the result will be. I will report that my experience stayed the same throughout, a constant shade of red, and that I noticed nothing untoward. I will become even more convinced than I was before that qualia are determined by functional organization. Of course this will not be a *proof*, but the evidence will be hard to seriously dispute.

I conclude that by far the most plausible hypothesis is that replacement of neurons while preserving functional organization will preserve qualia, and that experience is wholly determined by functional organization.

Once again, one can extend the thought-experiment to other functional isomorphs. For systems much larger than a brain, we may need a complex system of radio transmitters to act as a connection between neurons and an external circuit, but that is no problem in principle. A problem arises with isomorphs that are much faster or slower than the original system. In this case, we cannot simply substitute a circuit from one system into the other and expect everything to function normally. However, we can still perform the experiment on a slowed-down or speeded-up version of the system in question. At worst, we have left open the possibility that a change in speed might invert qualia; but this hypothesis was never very plausible in the first place, and it would seem quite arbitrary that this would be the *only* way to invert qualia.

There is another small caveat to the generality of the argument. The argument does not refute the possibility of very mild spectrum inversions, Between dark red and a slightly darker red, for instance, there may be nine intermediate shades such that no two neighboring shades are distinguishable. In such a case the Dancing Qualia scenario is not a problem; if the system notices no difference on flipping the switch, that is just what we would expect.

Of course, there is nothing special about the figure of one-tenth as the amount of difference between two neighboring systems. But we cannot make the figure too high. If we made it as high as one half, we would run into problems with personal identity: it might reasonably be suggested that upon flipping the switch, we are creating a new person, and it would not be a problem that the new person noticed no change. Perhaps we might go as high as 20possibility of very mild inversions, the kind that could be composed of four or five unnoticeable changes. We can reduce the impact of this worry, however, by noting that it is very unlikely that experience depends equally on all areas of the brain. If color experience depends largely on a small area of the visual cortex, say, then we could perform any qualia inversion in one fell swoop while only replacing a small portion of the system, and the argument would succeed against even the mildest noticeable qualia inversion.

In any case, the possibility of a mild underdetermination of experience by organization is a very unthreatening one. If we wished, we could accept it, noting that any differences between isomorphs would be so slight as to be uninteresting. More likely, we can note that this would seem an odd and unlikely way for the world to be. It would seem reasonable that experiences should be invertible across the board, or not invertible at all, but why should the world be such that a small inversion is possible but nothing more? This would seem quite arbitrary. We cannot rule it out, but it is not a hypothesis with much antecedent plausibility.

It should be noted that the Dancing Qualia argument works just as well against the possibility of absent qualia as against that of inverted qualia. If absent qualia are possible, then on the path to absent qualia we can find two slightly different systems whose experience differs significantly, and we can install a backup circuit in the same way. As before, the hypothesis implies that switching will cause my qualia to dance before my eyes, from vivid to tepid and back, without my ever noticing any change. This is implausible for the same reasons as before, so we have good reason to believe that absent qualia are impossible.

Overall, the Dancing Qualia argument seems to make an even more convincing case against absent qualia than the Fading Qualia argument does, although both have a role to play. Where an opponent might bite the bullet and accept the possibility of Fading Qualia, Dancing Qualia are an order of magnitude more difficult to accept. The very immediacy of the switch makes a significant difference, as does the fact that the subject cannot notice something so striking and dynamic. The possibility of Fading Qualia would imply that some systems are out of touch with their conscious experience, but Dancing Qualia would establish a much stranger gap.

5 Nonreductive functionalism

To summarize: we have established that if absent qualia are possible, then Fading Qualia are possible; if inverted qualia are possible, then Dancing Qualia are possible; and if absent qualia are possible, then Dancing Qualia are possible. But it is implausible that Fading Qualia are possible, and it is extremely implausible that Dancing Qualia are possible. It is therefore extremely implausible that absent qualia and inverted qualia are possible. It follows that we have good reason to believe that the principle of organizational invariance is true, and that functional organization fully determines conscious experience.

It should be noted that these arguments do not establish functionalism in the strongest sense, as they establish at best that absent and inverted qualia are empirically (or nomologically) impossible. There are two reasons why the arguments cannot be extended into an argument for logical or metaphysical impossibility. First, both Fading Qualia and Dancing Qualia seem to be intelligible hypotheses, even if they are very implausible. Some might dispute their logical possibility, perhaps holding that it is constitutive of qualia that subjects can notice differences between them. This conceptual intuition would be controversial, but in any case, even if we were to accept the logical impossibility of Fading and Dancing Qualia, there is a second reason why these arguments do not do not establish the logical or metaphysical determination of conscious experience by functional organization.

To see this second reason, note that the arguments take as an *empirical* premise certain facts about the distribution of functional organization in physical systems: that I have conscious experiences of a certain kind, or that some biological systems do. If we established the logical impossibility of Fading and Dancing Qualia, this might establish the logical necessity of the *conditional*: if one system with fine-grained functional organization F has a certain sort of conscious experiences, then any system with organization F has those experiences. But we cannot establish the logical necessity of the conclusion without establishing the logical necessity of the premise, and the premise is itself empirical. On the face of it, it is difficult to see why it should be logically necessary that *brains* with certain physical properties give rise to conscious experience. Perhaps the most tenable way to argue for this necessity is via a form of analytic functionalism; but in the context of using the Fading and Dancing Qualia arguments to *establish* this sort of functionalism, this strategy would be circular. It follows that the Fading and Dancing Qualia arguments are of little use in arguing for the logical and metaphysical impossibility of absent and inverted qualia.

The arguments therefore fail to establish a strong form of functionalism upon which functional organization is *constitutive* of conscious experience; but they succeed in establishing a weaker form, on which functional organization *suffices* for conscious experience with nomological necessity. We can call this view *nonreductive functionalism*, as it holds that conscious experience is determined by functional organization without necessarily being reducible to functional organization. As things stand, the view is just as compatible with certain forms of property dualism about experience as with certain forms of physicalism. Whether the view should be strengthened into a reductive version of functionalism is a matter that the Fading and Dancing Qualia arguments leave open.

In any case, the conclusion is a strong one. It tells us that systems that duplicate our functional organization will be conscious even if they are made of silicon, constructed out of water-pipes, or instantiated in an entire population. The arguments in this paper can thus be seen as offering support to some of the ambitions of artificial intelligence. The arguments also make progress in constraining the principles in virtue of which consciousness depends on the physical. If successful, they show that biochemical and other non-organizational properties are at best indirectly relevant to the instantiation of experience, relevant only insofar as they play a role in determining functional organization.

Of course, the principle of organizational invariance is not the last word in constructing a theory of conscious experience. There are many unanswered questions: we would like to know just what sort of organization gives rise to experience, and what sort of experience we should expect a given organization to give rise to. Further, the principle is not cast at the right level to be a truly *fundamental* theory of consciousness; eventually, we would like to construct a fundamental theory that has the principle as a consequence. In the meantime, the principle acts as a strong constraint on an ultimate theory.