How to resist the Fading Qualia Argument

Andreas Mogensen (Global Priorities Institute, University of Oxford)

Global Priorities Institute | March 2024

GPI Working Paper No . 5-2024

Please cite this working paper as: Mogensen, A. How to resist the Fading Qualia Argument. *Global Priorities Institute Working Paper Series, No. 5-2024.* Available at: https://globalprioritiesinstitute.org/how-to-resist-the-fading-qualia-argument-mogensen

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

How to Resist the Fading Qualia Argument

Andreas L. Mogensen

Abstract

The Fading Qualia Argument is perhaps the strongest argument supporting the view that in order for a system to be conscious, it does not need to be made of anything in particular, so long as its internal parts have the right causal relations to each other and to the system's inputs and outputs. I show how the argument can be resisted given two key assumptions: that consciousness is associated with vagueness at its boundaries and that conscious neural activity has a particular kind of holistic structure. I take this to show that what is arguably our strongest argument supporting the view that consciousness is substrate independent has important weaknesses, as a result of which we should decrease our confidence that consciousness can be realized in systems whose physical composition is very different from our own.

1 Introduction

Many believe that in order for a system to be conscious, it does not need to be made of anything in particular, so long as its internal parts have the right causal relations to each other and to the system's inputs and outputs. As a result, many also believe that the right software could in principle allow there to be something it is like to inhabit a digital computer, controlled by an integrated circuit etched in silicon. A recent expert report concludes that if consciousness requires only the right causal relations among a system's inputs, internal states, and outputs, then "conscious AI systems could realistically be built in the near term." (Butlin et al. 2023: 6) If that were to happen, it could be of enormous moral importance, since digital minds could have superhuman capacities for well-being and ill-being (Shulman and Bostrom 2021).

But is it really plausible that any system with the right functional organization will be conscious - even if it is made of beer-cans and string (Searle 1980) or consists of a large assembly of people with walky-talkies (Block 1978)? My goal in this paper is to raise doubts about what I take to be our strongest argument supporting the view that consciousness is *substrate independent* in something like this sense.[1] The argument I have in mind is Chalmers' *Fading Qualia Argument* (Chalmers 1996: 253–263). I show how it is possible to resist the argument by appeal to two key assumptions: that consciousness is associated with vagueness at its boundaries and that conscious neural activity has a particular kind of holistic structure. Since these assumptions are controversial, I claim only to have exposed important weaknesses in the Fading Qualia Argument.

I'll begin in section 2 by explaining what the Fading Qualia Argument is supposed to show and the broader dialectical context it inhabits. In section 3, I give a detailed presentation of the argument. In section 4, I show how the argument can be answered given the right assumptions about vagueness and the structure of conscious neural activity. At this point, I rely on the assumption that vagueness gives rise to truth-value gaps. In section 5, I explain how the argument can be answered even if we reject that assumption. In section 6, I say more about the particular assumption about the holistic structure of conscious neural activity needed to resist the Fading Qualia Argument in the way I outline. I take the need to rely on this assumption to be the greatest weakness of the proposed response.

2 Organizational Invariance and Two Arguments

Chalmers argues for the *principle of organizational invariance*, according to which, given any system that has conscious experiences in any nomologically possible world, any other system with the same fine-grained functional organization will have qualitatively identical experiences in any nomologically possible world, regardless of its material composition.

Two physical systems are said to realize the same *functional organization* if they can be divided into physical components with the same causal dependencies relating their in-

[1] See the third paragraph in section 2 for discussion of two ways in which the conclusion supported by this argument is weaker than some may expect a principle of substrate independence to be.

puts, outputs, and internal states. For sameness of *fine-grained* functional organization, we require that the sameness of causal dependencies relating the systems' inputs, outputs, and internal states holds at a sufficiently fine-grained level of description that the two systems determine exactly the same behavioural capacities. Moreover, we require that the systems realize corresponding states at the relevant times. Any two systems sharing the same fine-grained functional organization, so understood, are said to be *functional isomorphs*.

It's worth noting at the outset that the principle of organizational invariance is weaker in two respects than we might have wanted a principle of substrate independence to be. Firstly, because it's phrased as a conditional, it's consistent with the hypothesis that there are in fact no nomologically possible conscious systems that are not composed of neurons, glia, and the like, because the laws of nature sharply constrain the kinds of materials that can realize the functional organization required for a system to be conscious (see Chalmers 1996: 260).[2] Secondly, it is consistent with the hypothesis that there are nomologically impossible worlds within which functional isomorphs differ in the character of their experience or in whether they have any experience at all because of differences in their material composition (see Chalmers 1996: 257, 269).[3]

Chalmers (1996) presents two arguments in support of the principle of organizational invariance: the *Fading Qualia Argument* and the *Dancing Qualia Argument*. In the first instance, the Fading Qualia Argument is intended to rule out the nomological possibility of functional isomorphs that differ in respect of being conscious at all, while the Dancing Qualia Argument is intended to rule out the nomological possibility of functional isomorphs that are both conscious but nonetheless differ in the character of their experiences.

I focus my discussion on the Fading Qualia Argument. Some readers of *The Conscious Mind* may recall that Chalmers (1996: 270) also argues that the Dancing Qualia Argument

---

[2] See Godfrey-Smith (2016) and Cao (2022) for arguments supporting this hypothesis.

[3] It also allows for nomologically impossible worlds containing unconscious functional isomorphs of systems that are conscious in the actual world, i.e., *zombies* (Chalmers 1996: 94–99).

can be transposed through minor alterations to buttress the Fading Qualia Argument.[4] He also suggests that this modified Dancing Qualia Argument is actually a stronger argument. However, Chalmers has since changed his mind on this point and now views the Fading Qualia Argument as the stronger of the two (see Chalmers 2010: 34). Moreover, Chalmers' transposition of the Dancing Qualia Argument is intended to buttress the conclusion that the spectrum of cases involved in the Fading Qualia Argument does not involve fading qualia. The response to the Fading Qualia Argument that I'm going to outline is designed to allow us to answer the argument without committing ourselves to the hypothesis that fading qualia are involved.

3 The Fading Qualia Argument

Suppose that there exists a system, Robot, with the same fine-grained functional organization as Dave during a time when Dave undergoes a rich conscious experience, such as the experience of watching a loud, colourful basketball game. Suppose for reductio that Robot lacks any conscious experience altogether, because Robot is built from silicon chips, and silicon chips are the the wrong sort of components for realizing consciousness, in spite of their ability to duplicate the functional organization of Dave's brain.

We then imagine a spectrum of cases ranging from Dave to Robot, across which more and more of Dave's brain is replaced with functionally isomorphic components made of silicon chips. At the first step, we imagine replacing a single biological neuron with an artificial replacement component made of silicon that performs the same fine-grained functional role. As we move across the spectrum, we replace additional neurons, one by one, always maintaining the same fine-grained functional organization. Eventually, no biological material remains.[5]

---

[4] He denies that the Fading Qualia Argument can be transposed to yield an argument against functional isomorphs that are both conscious but that differ in the character of their experiences. See Chalmers (1996: 266).

[5] Even if we grant the possibility of a functional isomorph of the human brain realized in silicon, we might doubt that it is possible to replace individual neurons with silicon components while retaining the same fine-grained functional organization. That is because there would be serious difficulties involved in getting neurons and

By assumption, Dave is conscious and Robot isn't. The question is how this transition from consciousness to unconsciousness plays out as we move along the spectrum. One possibility is that consciousness degrades gradually as we imagine more and more of Dave's brain replaced. Another is that there is no gradual degradation and instead a point at which rich and detailed conscious experience abruptly switches off completely between two neighbouring cases. Each of these possibilities is argued to be extremely implausible, and, with it, the assumption that Robot isn't conscious.

Chalmers makes quick work of the idea that there might be *suddenly disappearing qualia*, i.e., a point at which consciousness abruptly switches from rich and detailed to totally absent. He rejects this possibility because, he says, an abrupt transition from consciousness to unconsciousness of this kind would imply "brute discontinuities in the laws of nature unlike those we find anywhere else" and because any point of discontinuity would be "entirely arbitrary" (Chalmers 1996: 255).

As for the possibility of *fading qualia*, Chalmers thinks we can reject this possibility for the following reason. Imagine a system halfway along the spectrum between Dave and Robot. Call this system Joe. By assumption, Joe's experience is degraded relative to Dave's. Thus, where Dave sees glaring, bright red and yellow uniforms worn by the players, Joe might experience only a tepid pink and murky brown; where Dave experiences the loud noises of the roaring crowd, Joe experiences only a sort of distant rumble. Nonetheless, Joe isn't able to report any of this. Since he is Dave's functional isomorph, he, like Dave, reports having vivid experiences of sound and colour. Nor is Joe in a position to notice that these reports systematically misdescribe what he's experiencing, assuming that noticing requires a particular kind of cognitive processing that supervenes on the fine-grained functional organization of the brain. After all, Dave is Joe's functional isomorph, and he fails to notice any such discrepancy, since none exists for him. On a functional construal of belief, Joe will even count as believing, like Dave, that his experiences really are strong

---

silicon components to interact appropriately with one another while retaining sameness of fine-grained functional organization: see Cao (2022: 5–10). Here, I set aside any worries about the feasibility of single-neuron replacement. See Chalmers (1996: 259–260) for suggestions about how the argument can be reformulated or amended to address such concerns.

and vivid, although they are really dim and murky. All this, Chalmers (1996: 257) claims, is extremely implausible: "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."[6]

4 Vagueness and the Fading Qualia Argument

The Fading Qualia Argument is obviously similar to the Sorites Paradox. Still, Chalmers (1996: 261) insists that his argument isn't soritical. It doesn't say: Dave is conscious; if X_i is conscious, then if X_{i+1} is obtained by replacing a single neuron in X_i with silicon, X_{i+1} is conscious also; therefore, by repeated application of *modus ponens*, Robot is conscious. In conceding, at least for the sake of argument,[7] that consciousness might in principle fade gradually to nothing as we make small changes, Chalmers seems to allow that some small change in the system, like replacement of a single neuron, could take us from a case of faint and badly degraded conscious experience to no conscious experience at all. Therefore, he does not assume the conditional premise stated above.

Nonetheless, it seems to me that we can resist the Fading Qualia Argument by interpreting the spectrum of cases to which it appeals as giving rise to vagueness. Here is how.

---

[6] Block (2023: 454-455) suggests that, on a natural reading, Joe *can* be said to suffer from functional pathology, since the cognitive processes that are recruited in phenomenal introspection are completely misfiring in his case. Arguing for the contrary conclusion, Chalmers (1996: 257) tells us that "Joe's processes are *functioning* as well as [Dave's] – by hypothesis, he is functionally isomorphic." (Emphasis in original.) Block (2023: 455) understandably objects that it is question-begging to define pathology as supervening on fine-grained functional organization in the present context. I think it's clear that the assumptions about the accessibility of phenomenal character to introspection on which the argument relies need to be made clearer and/or put on a firmer basis. However, since it's obvious that something intuitively bizarre is going on in Joe's case, I wouldn't be surprised to learn of a successful patch for this vulnerability in the argument.

[7] Elsewhere, he writes: "There is something odd about the idea that a system with n elements could not be conscious but a system with $n+1$ elements could be." (Chalmers 1996: 297)

At the outset, we say that it is determinate that the individual has a richly detailed experience of the basketball game unfolding. The colours are bright, the sounds loud. Once we've replaced a good portion of the original neural tissue, it's not the case that the individual is determinately conscious and experiencing the world as murky and muffled. Instead, it's indeterminate whether the individual is having an experience that is qualitatively identical to the original experience, and determinate that they aren't having any other experience. At this point, it is therefore indeterminate whether they are having any experiences at all. After enough of the brain has been replaced, it becomes determinate that they aren't having an experience that is qualitatively identical to the original experience, and it remains determinate that they aren't having any other experience. At that point, it's therefore determinate that they aren't having any experiences at all.

Suppose we enumerate the members of the sequence as $X_1, X_2, \ldots, X_n$, where X_1 is Dave and X_n is Robot. Let instances of the sentence schema $\ulcorner X_i$ has $E\urcorner$, obtained by replacing the variable i with the name of an integer between 1 and n, assert that the corresponding element in the sequence has exactly the same richly detailed experience of the basketball game unfolding that Dave has. Assume we reject an epistemicist theory of vagueness (Williamson 1994) and treat vagueness as giving rise to truth-value gaps. Assume, furthermore, that we accept the description of the sequence outlined in the previous paragraph. We can then say that the sentence schema $\ulcorner X_i$ has E and X_{i+1} does not have $E\urcorner$ has no true instances. In that sense, there are no suddenly disappearing qualia.

Assume, furthermore, that we accept a theory of vagueness that recognizes degrees of truth (e.g., Goguen 1969; Lewis 1970; Kamp 1975; Machina 1976; Edgington 1992, 1997; Smith 2008). For concreteness, suppose we accept a supervaluationist theory that lets us talk about smaller and larger subsets of the set of permissible sharpenings and treats sentences that are true according to larger subsets of sharpenings as having greater degrees of truth (see Williamson 1994: 154–156 and Keefe 2000: 171–172 for discussion). Then we can say that instances of the sentence schema $\ulcorner X_i$ has $E\urcorner$ decrease bit by bit in their degree of truth as we move across the spectrum. This provides some sense in which consciousness 'degrades' gradually, although $\ulcorner X_i$ does not have $E\urcorner$ has no true instances prior to the

point where it is also true that the system is completely unconscious.[8] In that sense there is no determinate change in phenomenal character as we move across the spectrum, setting aside the eventual loss of consciousness. But nor are there suddenly disappearing qualia.

Besides allowing us to go between the horns of Chalmers' dilemma, why might someone who rejects the principle of organizational invariance wish to interpret the $X_1, X_2, \ldots, X_n$ spectrum in this way? Suppose that we are attracted to a biological theory of consciousness on which consciousness essentially is a particular kind of functional organization of neural tissue, and so requires a neurophysiological substrate, much as water essentially is the liquid phase of H_2O, and not the liquid phase of just any polar molecule that is potable, transparent, etc. Evidence suggests that conscious experiences in human subjects involve what Dehaene (2014: 137) calls "a coherent brain-scale assembly": a state involving synchronized and integrated information processing that binds together distributed populations of neurons and facilitates their shared encoding of a coherent neural representation.[9] Therefore, one natural way to develop a biological theory of consciousness is to propose that E is some particular holistic pattern of brain activity, N, and any

---

[8] A somewhat similar idea is proposed by Bostrom (2006), who suggests that there is no difference in *quality* between Dave and Joe's experiences, but a difference in the *quantity* of experience that each has. However, in Bostrom's conception, the sense in which Dave has more experience than Joe is supposed to be the same sense in which there is more pain when more individuals are in pain. This, I take it, is distinct from the idea that ascriptions of phenomenal consciousness admit of degrees of truth and that such ascriptions are less true in respect of Joe than in respect of Dave. (Presumably the sentence 'Someone is in pain' does not take a higher degree of truth when two people are in pain than when only one is.) Moreover, the hypothesis that there is only a difference in quantity between Dave and Joe's experiences does not suffice to rebut the Fading Qualia Argument. We need to know what happens at other points in the spectrum. For example, is quality of experience preserved for all decreases in the quantity of experience until the quantity of experience is exactly zero? If so, we have suddenly disappearing qualia, whereas Bostrom makes no objection to Chalmers' dismissal of the possibility of suddenly disappearing qualia.

[9] Of course, it is hardly uncontroversial that consciousness neural activity involves global patterns of activity of this kind: see Lamme (2006, 2010) and Block (2007). Nonetheless, global theories of the neural correlates of consciousness are reasonable to believe given current evidence and it would be a significant liability if the Fading Qualia Argument required us to assume their falsity. Moreover, as I argue at the end of section 6, local theories appear to put independent pressure on the epistemological assumptions made by the argument.

experience, E', that is determinately different in character is a holistic pattern of brain activity, N', of a determinately different kind.[10] Arguably, what happens as more and more neurons are replaced isn't that we transition between different holistic patterns of brain activity. At each stage, if there is some holistic pattern of brain activity going on, it can only be N. However, it becomes less and less clear that the pattern of activity that's in place can be described as a pattern of *brain* activity. Exactly what counts as *brain* activity is surely vague. Just how much active matter needs to be neural matter? There's presumably no sharp boundary. Once we're far enough along the sequence - but not too far - it's simply going to be indeterminate whether the particular assembly of neural cells and silicon parts counts as a brain, as opposed to something else.[11]

Assume we are happy to go along with the assumptions I've outlined. Assuming we do not endorse an epistemicist theory of vagueness and instead treat vagueness as giving rise to truth-value gaps, we ought then to say that the schema $\ulcorner X_i$ has N and X_{i+1} does not have $N\urcorner$ has no true instances. Assuming that we accept a theory of vagueness that recognizes degrees of truth, such as the a supervaluationist theory that identifies higher degrees of truth with larger subsets of the set of permissible sharpenings, we can also say that instances of the sentence schema $\ulcorner X_i$ has $N\urcorner$ decrease bit by bit in their degree of truth. If we then also count all instances of the biconditional schema $\ulcorner X_i$ has E iff X_i has $N\urcorner$ as determinately true, then, on the standard supervaluationist logic, indeterminacy of any instance of $\ulcorner X_i$ has $N\urcorner$ entails that $\ulcorner X_i$ has $E\urcorner$ must also be indeterminate. In this way, we can end up committed to saying everything I said above.

If we can say all that, we appear to have defeated the Fading Qualia Argument. We don't have suddenly disappearing qualia, in the sense that $\ulcorner X_i$ has E and X_{i+1} does not have $E\urcorner$ has no true instances.[12] Nor do we have fading qualia, in the sense Chalmers imagines.

---

[10] Here, 'brain activity' may be understood loosely, as a pattern of activity occurring in tissue composed of neurons, glia, and the like, as opposed to a stricter sense of activity occurring within tissues organized to play a particular role within the nervous system of an animal. Thanks to Brad Saad for noting this distinction and pressing me to clarify this issue.

[11] The same can be said in respect of whether the system counts as *biological*.

[12] If we are supervaluationists, we do, however, remain committed to the truth of the sentence, "There is some i

Joe is not determinately having an experience as of muted colours and muffled sounds. Rather, it is indeterminate that Joe is having any experience at all, since it is indeterminate that he has the same richly detailed experience of the basketball game Dave has, and determinate that he has no other experience. Nothing in the set-up leads to bizarre failures of introspective access to the character of experience on behalf of any subject that can truthfully be described as conscious.

Of course, the argument assumes that there are borderline cases, where a system is neither determinately phenomenally conscious nor determinately unconscious. That is a controversial assumption. Many philosophers have the intuition that there are no indeterminate cases of phenomenal consciousness (Searle 1992; McGinn 1996; Tye 2021). That includes Chalmers (1996: 105, 297). However, explicit arguments for this claim are few (Antony 2008; Simon 2017) and appear to face significant challenges (Hall 2023). Given that the evolutionary trajectory from the earliest single-celled prokaryotes to human beings proceeds gradually as a result of the accumulation of selectively advantageous mutations over the course of billions of years, the hypothesis of a sharp boundary that separates conscious from unconscious life seems incredible (Godfrey-Smith 2020; Schwitzgebel 2021). Furthermore, the assumption that consciousness is sharp is close to question-begging in the current context. It is obviously vague what counts as brain activity. Therefore, anyone who finds plausible the view that consciousness is essentially a form of brain activity has good reason to demur from the assumption that consciousness cannot have borderline instances, given our current assumptions about vagueness (compare Antony 2006: 521–525).

5 An Apology for Suddenly Disappearing Qualia

In the previous section, I relied on the idea that a plausible way to develop a biological theory of consciousness is one on which E is some particular holistic pattern of brain activity, N, and what happens as more and more neurons are replaced isn't that we transition between different holistic patterns of brain activity, but simply that it becomes less and less

such that X_i has E and X_{i+1} does not have E." If that worries you, read on to section 5.

clear that the pattern of activity that's in place can be described as a pattern of *brain* activity, although it is determinate throughout that no other pattern of brain activity with the holistic character needed for consciousness is involved. In that case, epistemicism entails that ⌜X_i has N and X_{i+1} does not have N⌝ has a true instance. If we assume that all instances of the biconditional schema ⌜X_i has E iff X_i has N⌝ hold, it follows that ⌜X_i has E and X_{i+1} does not have E⌝ must also have a true instance. Assuming no other experience is had at any point in the spectrum, epistemicism thereby commits us to the possibility of suddenly disappearing qualia.[13]

If we find epistemicism plausible,[14] this might seem like an embarrassment.[15] After all, Chalmers tells us that the possibility of suddenly disappearing qualia is absurd and implies that for it to be plausible that replacement of a single neuron results in the loss of consciousness, the person's experience must already be of severely muted colours and badly muffled sounds. I disagree. If we think that brain activity is necessary for consciousness and take seriously the idea that replacement of *this* particular neuron can make the difference as to whether a system constitutes a brain – as epistemicists surely must – we need feel no embarassment in supposing that that cut-off is associated with a sudden loss of rich and detailed conscious experience, as opposed to needing to occur at a point where consciousness has already become thoroughly degraded. At the least, I claim, that holds true for any sense of 'thoroughly degraded' that we should be shocked to find a person

[13] As noted previously, it also follows from supervaluationism that the sentence "There is some i such that X_i has E and X_{i+1} does not have E" comes out true, although ⌜X_i has E and X_{i+1} does not have E⌝ has no true instances. The discussion in this section may therefore serve also to reassure readers who have residual doubts about the ability of supervaluationism to avoid an implausible commitment to suddenly disappearing qualia.

[14] This may not include Chalmers: see Chalmers (2012: 288–289).

[15] Epistemicism does potentially allow us to drop one of the philosophically controversial assumptions outlined at the end of the previous section: namely, the assumption that 'phenomenally conscious' is a vague predicate. Even if 'phenomenally conscious' is coextensive with some neurophysiological predicate that is vague, it need not follow that 'phenomenally conscious' is itself vague (compare Antony 2006: 522–523). Epistemicism arguably entails that any vague predicate is coextensive with some sharp predicate (if only in an idealized language): namely, a sharp predicate that picks out the location of the borderline between the vague predicate's extension and its complement.

failing to notice.

To explain why, I rely on an analogy. Imagine that I use a pencil to write out the sentence, "Kim is wearing brightly coloured yellow socks," on a sheet of paper. This sentence has a certain intentional content. It represents Kim as the wearer of bright yellow socks. Imagine now that I write out this sentence using a successively fainter and fainter hand, until I apply so little pressure to the page that no sentence at all is inscribed thereon. As you scan your eyes down the page, the sentence "Kim is wearing brightly coloured yellow socks" becomes gradually fainter and fainter, until it disappears altogether. But the same doesn't hold true of the colour of Kim's socks, as represented by the sentences I write. Those colours don't get any fainter. So long as there is a determinate sentence there to be read at all, the sentence represents Kim's socks as being exactly as brightly coloured as its predecessors, no matter how faint my handwriting. The content of the sentence does not change, insofar as there is any determinate sentence present to have a content at all. It is only the vehicle that gradually degrades.

Consciousness too has intentional content. Some, including Chalmers (2004), go so far as to hold that phenomenal properties are identical to certain representational properties (Harman 1990; Dretske 1995; Tye 1995; Lycan 1996; Byrne 2001). As with the case of written sentences, we need to distinguish the properties represented *in* experience from properties *of* the experience. Recall Dave, who is watching a loud and colourful basketball game in action. Suppose part of Dave's experience is of a yellow triangle of fabric, seen as part of a basketball player's jersey. That experience is presumably not itself either yellow or triangular.

A plausible hypothesis, then, is that there exists a spectrum of possible cases involving conscious experience that is like the sequence of ever-fainter sentences that I described above. In other words, we can imagine a spectrum of cases involving conscious experiences with the same intentional content, but which differ in that the vehicle of content gradually degrades from one case to the next, up to the point at which there is finally nothing there at all. The content remains the same throughout, until there's no content. Moreover, we can imagine that the spectrum appealed to in the Fading Qualia Argument works like that. If so, I think the argument can be defeated while conceding a commitment to

suddenly disappearing qualia.

A key motivation for identifying phenomenal and intentional properties is a line of argument going back at least to Moore (1903), which maintains, roughly, that our ordinary mode of access to the character of experience is indirect and goes via access to the intentional content of experience. If I try to bring my attention to the character of my experience of the blue of the sky, I find that I end up focusing on the blue of the sky, as represented in my experience. There doesn't seem to be any separate mental blue in my mind. We needn't here assume the *strong transparency hypothesis* - rejected by Moore - on which it is *impossible* to become directly aware via introspection of intrinsic properties of experience distinct from the experience's intentional content (Harman 1990; Tye 1995). In what follows, I assume only the more plausible *weak transparency hypothesis*, on which it's at the very least rare and/or difficult (Kind 2003).

Suppose Joe has had a large part of his brain replaced with silicon chips, but is still conscious. Assuming weak transparency, for Joe to be thoroughly and bafflingly out of touch with the character of his experiences, he arguably needs to be ignorant of their intentional content. Even if intentionalism is false and there is some non-intentional difference in the character of Joe's and Dave's experiences, given weak transparency, Joe's failure to notice would hardly be baffling or extraordinary. If Joe is to be ignorant of the intentional content of his experiences, that presumably requires their intentional content to differ from the content of Dave's experiences. But that need not be true, since Joe's experience might be degraded relative to Dave's only in the sense that a faint inscription of "Kim is wearing brightly coloured yellow socks" is degraded relative to an inscription made by pressing hard with a newly sharpened pencil. Just as those two sentences nonetheless have the same intentional content, so Dave and Joe would have experiences with the same intentional content, and those experiences would be very difficult, if not impossible, to distinguish introspectively.

More generally, as we traverse the spectrum, the world is not experienced as dimmer and fainter. In that sense, there are no fading qualia. It is the vehicle that gradually decays, not the content. The intentional content of experience remains fixed up to the very end, when the vehicle of content becomes so utterly degraded that it fails to support any

content at all. In that sense, we have a case of richly detailed consciousness disappearing suddenly as a result of a minor transition. But there needn't be anything arbitrary or mysterious going on here, since the analogy with fainter and fainter inscriptions of "Kim is wearing brightly coloured yellow socks" shows us that this sort of change is perfectly intelligible. It's just a fact about the way intentional contents relate to their vehicles that exactly the same intentional content can be supported by a badly degraded vehicle all the way up to the point at which additional degradation of the vehicle means there's actually no vehicle left to support any content at all.

6 Holism

We never said explicitly according to what pattern neurons are replaced with silicon chips as we make the transition from Dave to Robot. It might have been natural to imagine that individual neurons were being replaced at random. To assess the robustness of the proposed response to the Fading Qualia Argument, we should make sure to explicitly consider alternative scenarios that might be used to run the argument.

Imagine, then, that we do not randomly replace individual neurons, but first replace all and only those neurons encoding, say, visual information. Suppose also that Joe occurs at a point along the spectrum at which replacement of all and only those neurons encoding visual information is complete. Someone might imagine that those who treat conscious experience as being essentially a pattern of brain activity will think that in this scenario, Joe's experience will be like Dave's except that it has no visual qualia, since we have preserved those patterns of brain activity that encode the remaining information of which Dave is conscious (compare Searle 1992: 66-67). Maybe it was indeterminate at some point along the way whether the system still had visual qualia. Even so, we might think, if conscious experience is essentially a pattern of brain activity, then once we've replaced all the neurons encoding visual information, it has to be determinate that visual qualia are gone.

If Joe's experience were like Dave's except that it had no visual qualia, then Joe would exhibit a bizarre failure of introspective access to the character of his current experience if

he failed to notice that he lacks any conscious visual experiences.[16] The conclusion that he would fail to notice can easily be derived by repurposing the arguments set out in section 3 for thinking that Joe would not be able to notice if his visual qualia are faded. Thus, if we should expect that Joe's experience will be like Dave's but for the absence of visual qualia when the transition from Dave to Robot involves first replacing all and only those neurons encoding visual information, the Fading Qualia Argument is alive and kicking. To be able to respond to the argument in full generality, we have to say that replacement of all and only those neurons encoding visual information does not result in a case where Joe has an experience much like Dave's but without visual qualia.

As a matter of fact, I did say that. At least, I said something that directly entails it. In section 4, I said that as more and more neurons are replaced with silicon components as we make the transition from Dave to Robot, we do not transition between different holistic patterns of brain activity and so do not transition between qualitatively different conscious experiences. That entails that it is not true that Joe is having an experience qualitatively different from Dave's, as would be the case if Joe were having an experience much like Dave's but without visual qualia.

Still, it might not have been obvious that that entailment was supposed to hold even in the case when the transition from Dave to Robot involves first replacing all and only those neurons encoding visual information – a case that might not have been salient to you. Noticing that entailment might lead you to have (new) doubts about the assumption on which I relied previously about the holistic character of conscious neural processing.

I claim that how seriously we should take those doubts depends on how seriously we should take the view known as *phenomenal holism* in the philosophy of consciousness (Searle 2000; Bayne and Chalmers 2003; Bayne 2010): very roughly, the view that the conscious whole is prior to its parts. That's because the assumption on which I've relied so

---

[16] Note, however, that cases of *Anton syndrome* exhibit something like this profile, with patients who are cortically blind insisting that they can see. Thus, Chalmers needs to say why a commitment to the nomological possibility of Joe's unnoticed blindness, as described here, is significantly harder to accept than a commitment to the nomological possibility of Anton syndrome, which everyone must accept. See Chalmers (1996: 260–261) and Block (2023: 455–456) for discussion. Compare the discussion in footnote 6.

far follows from phenomenal holism in conjunction with identity statements it would be natural for adherents of a biological theory of consciousness to endorse.

I'll start by doing more to clarify the assumption on which I've relied about the holistic character of conscious neural processing. I'll call this assumption *neural holism* from now on. Consider some overarching pattern of brain activity that binds together activity in different populations of neurons, thereby realizing an integrated and coherent *total phenomenal state*, understood as the overarching experience that subsumes each of the more particular experiences had by a conscious subject at a given point in time.[17] Neural holism says that those of the network's proper parts that encode experiences subsumed by the total phenomenal state do not realize conscious experiences in and of themselves: they do not realize any conscious experience except by virtue of forming part of an overarching pattern of brain activity that similarly integrates information across distributed populations of neurons so as to realize a coherent neural representation.

It follows that replacement of all and only those neurons encoding visual information does not result in Joe having an experience much like Dave's but without visual qualia. Those patterns of brain activity that encode the non-visual information of which Dave is conscious don't suffice on their own for realizing proper parts of the original experience. If the overarching pattern of information processing of which they form part is correctly described as a pattern of brain activity, then they realize the relevant proper parts of the original experience, which is itself realized in its entirety. If it does not, they do not. If it is indeterminate, then it is indeterminate whether any proper part of the original experience occurs.[18]

---

[17] On how to understand the relation of subsumption, see Bayne and Chalmers (2003).

[18] If that's what neural holism says about Joe, does it also entail, absurdly, that people who are blind as a result of lesions to the visual cortex aren't having experiences that are like the experiences of sighted people but lacking in visual qualia? No. Here's the crucial difference. In cases of cortical blindness, neural populations encoding sensory information participate in an overarching pattern of brain activity encoding a coherent content from which visual information is missing. In Joe's case, by contrast, if those same patterns of activation occur, they do so as part of a neuron-silicon assembly encoding a coherent content to which visual information contributes. Only in the latter case does the substantial replacement of neural matter by silicon call into question the biological status of the overarching information-processing network responsible for integrating informa-

I claim that neural holism follows from phenomenal holism, given identities that it would be natural for adherents of a biological theory of consciousness to accept. Phenomenal holism is the view that the total phenomenal state is basic and its parts are derivative. The specific version of phenomenal holism I have in mind says that any component of any total phenomenal state is conscious only because there exists a total phenomenal state of which it is part. In this sense, consciousness is inherited by the parts from the whole.

Bayne (2010) formulates holism about consciousness slightly differently, namely as the view that "the components of the phenomenal field are conscious only as the components of *that* field." (Bayne 2010: 225) (My emphasis.) I take him as denying that one and the same token phenomenal state can occur within the context of different total phenomenal states.[19] We may, however, find it plausible that one and the same pain can persist throughout changes in the overall character of a person's experience and so participate in distinct total phenomenal states. In my formulation, phenomenal holism allows for this possibility, but insists that any component of any phenomenal field depends for its being conscious at a time on its inclusion at that time in the given conscious whole of which it then forms part. In a similar fashion, a sound wave may pass from a body of air to a body of water, but is in each case nothing more than a pattern of disturbance in the given medium. In a roughly similar sense, phenomenal holism understands the components of the phenomenal field "as bumps or forms or features in the unified field of consciousness." (Searle 2000: 574)

Holism in my formulation remains supported by the same arguments relied on by Bayne. For example, one key argument for holism (Bayne and Chalmers 2003; Bayne 2010: 236–238) is that it explains *phenomenal unity*, understood as the idea that for any set of experiences had by any conscious subject at a given point in time, there is an overarch-

---

tion across sensory channels.

[19] Bayne (2010: 243) writes that "tokens of a single fine-grained phenomenal state type can occur within the context of various total phenomenal state types." Other philosophers such as Sprigge (1983) may be understood as putting forward a form of holism on which tokens of a fine-grained phenomenal state type can only occur as parts of one and the same total phenomenal state type, because the phenomenal character of every part of any total phenomenal state reflects the character of every other. See Dainton (2000: 182–239) for discussion.

ing experience that subsumes them. Phenomenal unity is a claim about how things stand for the subject at a fixed point in time and so doesn't touch on the ability of one and the same token experience to participate in different total phenomenal states at other times. It stands to reason that phenomenal unity is explained at least as well by my formulation of phenomenal holism as by Bayne's.

Now to show that the desired entailment exists between neural and phenomenal holism given the right identities. Say that any total phenomenal state is identical to some overarching pattern of brain activity that binds together distributed populations of neurons to realize an integrated and coherent neural representation. Identify its components with those proper parts of the overarching pattern that encode different aspects of the coherent representational state realized across the pattern as a whole. These identities strike me as ones it would be natural for adherents of a biological theory of consciousness to endorse. And with these identities in place, phenomenal holism entails neural holism:[20] that the components of a given total phenomenal state are conscious only because there exists some total phenomenal state of which they are part entails that the proper parts of an overarching pattern of brain activity identified with a given total phenomenal state are conscious only as the parts of an overarching pattern of brain activity of that kind.

I noted at the start of this section that in order to be able to respond to the Fading Qualia Argument in full generality, we have to say that replacement of all and only those neurons encoding visual information does not result in a case where Joe has an experience much like Dave's but without visual qualia. As I've now argued, that conclusion follows from neural holism, which itself follows from phenomenal holism in conjunction with identity statements involving total phenomenal states and their components that it would be natural for adherents of a biological theory of consciousness to endorse. Obviously, phe-

[20]The converse entailment also holds.

nomenal holism is controversial.[21] Still, Chalmers accepts holism[22] (Bayne and Chalmers 2003), and there are good arguments in its favour. As noted, it explains the unity of consciousness (Bayne and Chalmers 2003, Bayne 2010: 236–238). It also gains support from scientific theories of consciousness like the *global neuronal workspace theory* (Dehaene 2014) (see Bayne 2010: 228–229).

It is also worth noting that some of the best reasons to doubt phenomenal holism may also cast doubt on the Fading Qualia Argument itself. Here is the particular example I have in mind. As noted by Bayne (2010: 228), phenomenal holism can be challenged by evidence supporting theories of consciousness like Lamme's *local recurrent processing theory* (Lamme 2006, 2010). According to Lamme, visual experience can be realized by local activity within circumscribed regions at the back of the brain and does not depend on global patterns of neural activity (see also Zeki and Bartels 1999; Block 2007). This suggests an atomistic picture of perceptual consciousness as built up from different streams of information processing, each conscious in itself in virtue of local recurrent processing in circumscribed cortical areas. But local theories like Lamme's also call into doubt the assumptions about the accessibility of phenomenal character to introspection on which the Fading Qualia Argument relies. Local processing of visual stimuli at the back of the brain need not recruit areas at the front of the head involved in introspective judgment. As a result, theories like Lamme's strongly suggest the possibility of dissociations between experience and introspective access to the character of experience even in the absence of any functional pathology (see Block 2023: 451-459).[23]

---

[21] Though it is worth emphasizing here that, as explained previously, the version of holism I outline is weaker than others, and so escapes objections that specifically target the hypothesis that the same experience cannot occur as part of distinct total phenomenal states: for example, the argument against holism put forward by Lee (2014: 296–297).

[22] Albeit with some caveats, see Chalmers (2014: 792–796).

[23] Note also that for present purposes we do not need holism to be true of all minds, but only of those minds for which we expect the character of experience to be introspectable. If there are counter-examples to phenomenal unity in, say, cases of epileptic patients treated with commissurotomy (Nagel 1971; Lockwood 1989; Bayne 2008; Schechter 2014), then holism may be false of such minds. However, such putative cases of dis-

7 Conclusion

I've shown how the Fading Qualia Argument can be resisted given assumptions to the effect that consciousness is associated with vague boundaries and that conscious neural activity has a particular kind of holistic structure. I take this to show that our strongest argument supporting the view that consciousness is substrate independent has important weaknesses, as a result of which we should decrease our confidence that conscious AI systems can be built.

Since the assumptions needed to resist the Fading Qualia Argument in the way I've outlined are controversial, I claim only to have exposed important weaknesses in the argument. I have certainly not refuted it. Thus, while I take my argument to show that we should decrease our confidence that conscious AI systems can be built, it is an open question by how much, and whether the decrease should be substantial. It depends on what attitude we take to the two assumptions on which I've relied. There is a lot of scope for reasonable disagreement here. For my own part, I feel confident that consciousness is in fact associated with vagueness at its boundaries. By contrast, I go back and forth on whether I find the kind of holism I outlined in the previous section to be believable.

References

Antony, M. V. 2006. Vagueness and the metaphysics of consciousness. *Philosophical Studies 128*(3): 515–538.

Antony, M. V. 2008. Are our concepts Conscious State and Conscious Creature vague? *Erkenntnis 68*(2): 239–263.

Bayne, T. 2008. The unity of consciousness and the split-brain syndrome. *Journal of Philosophy 105*(6): 277–300.

---

unified consciousness are also associated with significant disruptions of introspective access to aspects of the disunified conscious stream. Consider the well-known case of patient P.S. who was unable to report the visual presentation of a snow scene available to only his left eye when explaining his left hand's choice of a snow shovel as an appropriate item to pair with the associated picture (Gazzaniga and LeDoux 1978).

Bayne, T. 2010. *The Unity of Consciousness*. Oxford: Oxford University Press.

Bayne, T. J. and D. J. Chalmers. 2003. What is the unity of consciousness? In *The Unity of Consciousness*, eds. C. Frith and A. Cleeremans, 23–58. Oxford: Oxford University Press.

Block, N. 1978. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9: 261–325.

Block, N. 2007. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences 30*(5): 481–548.

Block, N. 2023. *The Border Between Seeing and Thinking*. Oxford: Oxford University Press.

Bostrom, N. 2006. Quantity of experience: Brain-duplication and degrees of consciousness. *Minds and Machines 16*(2): 185–200.

Butlin, P., R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon, and R. VanRullen. 2023. Consciousness in artificial intelligence: Insights from the science of consciousness. https://arxiv.org/abs/2308.08708.

Byrne, A. 2001. Intentionalism defended. *Philosophical Review 110*(2): 199–240.

Cao, R. 2022. Multiple realizability and the spirit of functionalism. *Synthese 200*(6): 1–31.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, D. J. 2004. The representational character of experience. In *The Future for Philosophy*, ed. B. Leiter, 153–181. Oxford University Press.

Chalmers, D. J. 2010. *The Character of Consciousness*. Oxford: Oxford University Press.

Chalmers, D. J. 2012. *Constructing the World*. Oxford: Oxford University Press.

Chalmers, D. J. 2014. Strong necessities and the mind-body problem: A reply. *Philosophical Studies 167*(3): 785–800.

Dainton, B. 2000. *Stream of Consciousness: Unity and Continuity in Conscious Experience*. London: Routledge.

Dehaene, S. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York, NY: Viking Press.

Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Edgington, D. 1992. Validity, uncertainty and vagueness. *Analysis 52*(4): 193–204.

Edgington, D. 1997. Vagueness by degrees. In *Vagueness: A Reader*, eds. R. Keefe and P. Smith, 617–630. Cambridge, MA: MIT Press.

Gazzaniga, M. S. and J. E. LeDoux. 1978. *The Integrated Mind*. New York, NY: Springer US.

Godfrey-Smith, P. 2016. Mind, matter, and metabolism. *Journal of Philosophy 113*(10): 481–506.

Godfrey-Smith, P. 2020. *Metazoa: Animal Minds and the Birth of Consciousness*. London: William Collins.

Goguen, J. A. 1969. The logic of inexact concepts. *Synthese 19*(3-4): 325–373.

Hall, G. 2023. Is consciousness vague? *Australasian Journal of Philosophy 101*(3): 670–684.

Harman, G. 1990. The intrinsic quality of experience. *Philosophical Perspectives* 4: 31–52.

Kamp, J. A. W. 1975. Two theories about adjectives. In *Formal Semantics of Natural Language*, ed. E. L. Keenan, 123–155. Cambridge University Press.

Keefe, R. 2000. *Theories of Vagueness*. Cambridge: Cambridge University Press.

Kind, A. 2003. What's so transparent about transparency? *Philosophical Studies 115*(3): 225–244.

Lamme, V. A. F. 2006. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences 10*(11): 494–501.

Lamme, V. A. F. 2010. How neuroscience will change our view on consciousness. *Cognitive Neuroscience 1*(3): 204 – 220.

Lee, G. 2014, 10). Experiences and Their Parts. In *Sensory Integration and the Unity of Consciousness*. The MIT Press.

Lewis, D. K. 1970. General semantics. *Synthese 22*(1-2): 18–67.

Lockwood, M. 1989. *Mind, Brain and the Quantum: The Compound "I"*. Oxford: Oxford University Press.

Lycan, W. G. 1996. *Consciousness and Experience*. Cambridge, Mass.: MIT Press.

Machina, K. F. 1976. Truth, belief, and vagueness. *Journal of Philosophical Logic 5*(1): 47–78.

McGinn, C. 1996. *The Character of Mind: An Introduction to the Philosophy of Mind*. Oxford: Oxford University Press.

Moore, G. E. 1903. The refutation of idealism. *Mind 12*(48): 433–453.

Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese 22*(3-4): 396–413.

Schechter, E. 2014. Partial unity of consciousness: A preliminary defense. In *Sensory Integration and the Unity of Consciousness*, eds. D. Bennett and C. Hill, 347–374. Cambridge, MA: MIT Press.

Schwitzgebel, E. 2021. Borderline consciousness: When it's neither determinately true nor determinately false that experience is present. http://faculty.ucr.edu/ eschwitz/SchwitzAbs/BorderlineConsciousness.htm.

Searle, J. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3): 417–57.

Searle, J. R. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Searle, J. R. 2000. Consciousness. *Annual Review of Neuroscience 23*(1): 557–578.

Shulman, C. and N. Bostrom. 2021. Sharing the world with digital minds. In *Rethinking Moral Status*, eds. S. Clarke, H. Zohny, and J. Savulescu, 306–326. Oxford: Oxford University Press.

Simon, J. A. 2017. Vagueness and zombies: Why 'phenomenally conscious' has no borderline cases. *Philosophical Studies 174*(8): 2105–2123.

Smith, N. J. J. 2008. *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.

Sprigge, T. 1983. *The Vindication of Absolute Idealism*. Edinburgh: Edinburgh University Press.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.

Tye, M. 2021. *Vagueness and the Evolution of Consciousness: Through the Looking Glass*. Oxford: Oxford University Press.

Williamson, T. 1994. *Vagueness*. New York: Routledge.

Zeki, S. and A. Bartels. 1999. Toward a theory of visual consciousness. *Consciousness and Cognition 8*(2): 225–259.