BORDERLINE CASES OF CONSCIOUSNESS

BRAINS, ROCKS, AND THE MACHINES IN BETWEEN

BY

DANIEL KOKOTAJLO

SUBMITTED TO THE

DEPARTMENT OF PHILOSOPHY

UNIVERSITY OF NOTRE DAME

THESIS ADVISOR: JEFF SPEAKS

2014

0. INTRODUCTION

If a computer were to simulate a human brain, complete with an accompanying virtual environment, on a low enough level that the simulated brain behaves as if it were conscious, would that computer contain a conscious person? This is a specific instance of the general question this paper seeks to answer, which is: Which kinds of physical systems are conscious?

This question is venerable and popular in the literature.[1] Answering it correctly is essential to understanding and solving many important problems in philosophy, cosmology, artificial intelligence, and religion.[2] I will not argue for a particular answer, though I will present considerations for and against some proposed answers. Instead, I will present a new methodology for making progress on this issue, and use this methodology to make some progress.

Section 1, the Setup, will define the target question of this paper, explain the literature which inspired it, and describe the methodology that this paper will use to make progress.

Section 2, the General Argument, will present a general argument that a certain kind of transformation, dubbed a Continuum, preserves consciousness. This argument is an expanded version of Chalmers' "Fading Qualia" argument.

Section 3, the Transformations, will present a series of possible transformations of one system into another. The first transformation starts with a conscious human brain and the last transformation ends with an ordinary rock.[3]

Section 4, the Properties, will discuss each of the transformations listed in section 3; specifically, it will discuss whether or not they preserve consciousness. If they all do, then absurd conclusions seem to follow; hence, at least one of them must fail to preserve consciousness. This section will argue that, unfortunately, they all seem to preserve consciousness.[4]

Section 5, the Conclusion, will discuss the implications of section 4. Since section 4 is not the final word on the matter, the problem it presents may be solvable through further exploration.

---

[1] For an example, see (Chalmers' "Absent Qualia…")
[2] Respectively: mind-body problem and many more, measure problem for multiverse theories, AI safety, and the problem of evil. Generally speaking, almost everything hinges on this question.
[3] Technically, an ordinary rock with a clock grafted on, or else a fusion of a clock and an ordinary rock.
[4] Although this paper will not argue very strongly against the early transformations; this is because this paper is more concerned with addressing functionalism/computationalism. See subsection 1.4.

That being said, the problem can be avoided by Externalist theories, eliminativist theories, and further-fact theories, so this paper renders those theories more plausible.

TABLE OF CONTENTS

1. SETUP

In this section, I will define the target question of this paper (1.1), explain the literature which motivated this project (1.2 & 1.3) and describe the general strategy that this paper will use to make progress (1.4 & 1.5)

1.1 Which question is this paper trying to answer?

Loosely following Chalmers, I distinguish between two important questions about consciousness: ("Absent Qualia…" 1)

Essence: What does it mean for something to be conscious?

Extension: Which kinds of physical systems are conscious?

The two questions are certainly related. A natural approach to answering them would be to first answer *Essence* and then use that answer to help answer *Extension*. This paper does not do that; instead, it presents a series of considerations and arguments about *Extension*, which will be indirectly relevant to *Essence*.

For this reason, I will not give a definition of consciousness here; I will merely refer the reader to the literature (see footnote).[5] The arguments in this paper are designed to be valid when applied to a wide variety of definitions and theories of consciousness. If an argument with an undesirable conclusion is invalid when applied to a particular theory, then so much the better for the theory. Indeed this is the general form that my conclusions will take.

Note 1: The question I am trying to answer is technically more complicated than stated above; more precisely, it is this: I am trying to find a way to tell, given a physical description[6] of a system S, whether or not S is conscious.

Note 2: For this paper, I will settle for being able to tell from a physical description of a system whether or not it *or some part of it* is conscious. For example, I might be unsure about whether (a) the brain is conscious, or (b) the human organism (with a brain) is conscious, or (c) some subsystem within the brain is conscious, and yet be very confident that the human organism (with a brain) *contains consciousness.*[7] Thus, when I speak of systems being conscious in this paper, what I am really interested in is whether or not they *contain consciousness*. For brevity, I will simply talk about systems being conscious, and trust the reader to understand that I really mean systems *containing consciousness.*

Note 3: I will often use phrases like "consciousness disappears during this transformation" or "This transformation fails to preserve consciousness." By this I mean: The transformation transforms a conscious system into a system which is not conscious. (Remember, this really means '…from a system which *contains consciousness* to a system which does not.')

1.2 Historical Context: Chalmers' Fading Qualia Argument

There are two important sources of inspiration for this paper. Understanding them is helpful for understanding this paper, but not necessary. This subsection briefly describes the first source of inspiration; it will be explored and critically analyzed in Section 2.

David Chalmers' "Fading Qualia" thought experiment imagines your brain slowly transformed into a computer by a process that replaces each neuron, one by one, with a silicon surrogate. ("Absent Qualia…" section 3) Each surrogate is a tiny machine that behaves exactly as

[5] See e.g. Van Gulick
[6] A description in the language of physics.
[7] In case it isn't clear: A system contains consciousness iff either it or some part of it is conscious.

the neuron it replaced would have behaved, so that the rest of the brain is (in a sense[8]) unaffected by the replacement: The remaining neurons continue to interact with each other, and with the surrogate, exactly as they would have done if it were a real neuron, and so the brain as a whole does not behave any differently than it would have without the replacement.

Note 1: Though I speak of replacing neurons, everything could just as well be done on a level lower than neurons. For example, tiny sections of cell wall could be replaced by mechanical surrogates, instead of entire neurons.

Note 2: Though I speak of replacing the brain, everything could just as well be done on a wider or smaller range of parts. For example, we could replace the entire nervous system, or the entire body, cell by cell; alternatively we could focus on some small section of the brain.

Chalmers' then argues that this transformation would preserve consciousness. His argument is that (1) consciousness could not disappear after a single neuron-replacement, because that would be arbitrary and unparsimonious, and (2) consciousness could not disappear gradually over the course of several neuron-replacements, because (since your overall behavior is the same) this would require you to be radically mistaken about your own consciousness, which is absurd. (His arguments for each of these claims will be explained in section 2.)

Chalmers then claims that his argument generalizes to support a version of functionalism:

> **Principle of Organizational Invariance:** Given any system that has conscious experiences, then any system that has [realizes] the same functional organization at a fine enough grain will have qualitatively identical conscious experiences. (Insertion mine, "Absent Qualia…" section 1)
>
> **Functional Organization:** 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. ("Absent Qualia…" section 1)

It is important to distinguish the generalization of Chalmers' argument from the argument itself. Section 2, which presents the General Argument, will explain Chalmers' argument in much

[8] No surrogate is perfect; after all, no matter how good a mechanical surrogate is in other respects, it will cause radiation passing through the brain to behave differently than a biological neuron would, simply because it is made of a different material. Really what I am saying is that the surrogate behaves the same way in all the kinds of behavior that matter.

more detail and expand it. That section will also discuss some of the implications of rejecting Chalmers' argument. Meanwhile, sections 3 and 4 will (among other things) attempt to figure out how far Chalmers' argument really does generalize. It is not at all clear that it generalizes in the way that he thinks it does. In fact, as subsection 1.3 is about to discuss, we have reason to think that it does not.

1.3 Historical Context: The Literature on Absurd Implementations

The Principle of Organizational Invariance (POI) is supposed to work like this:

> Systems (which are a kind of physical entity) map many-to-many onto functional organizations (which are a kind of mathematical abstraction). For example, every system maps to the trivial functional organization consisting of a single component with a single state that never changes. Every [finite, discrete] system also maps to a maximal functional organization, that has a component for every fundamental part of the system, with a state for every state of said part, with the exact same causal relations as the system. If two systems map to the same functional organization, then they also map to all the same "smaller" functional organizations.
>
> Some functional organizations are those that, when realized, give rise to consciousness. Thus, if a system S is conscious, then there must be some functional organization F that it maps to such that, should another system map to F, that other system would be conscious in exactly the same way as S. Thus, if a system S is conscious, all systems which have "the same functional organization at a fine enough grain" will be conscious as well.

Recall the definition of Functional Organization:

> **Functional Organization:** 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. (Italics mine, "Absent Qualia…" section 1)

The problems with the Principle of Organizational Invariance stem from the two italicized words.

First, the criteria for mapping systems to functional organizations are too liberal. All that is required is that the system *can* be divided in such-and-such a way. There are many ways to divide systems, and most of them are quite unnatural. As it turns out, if we make use of gerrymandered

divisions, we can make pretty much any system map to pretty much any functional organization. Combined with the POI, this would mean that pretty much every physical system has the same conscious experience as me right now!

A technical literature exists on this issue; the original proof can be credited to Putnam. (Putnam 1988, pp. 120-125) Chalmers himself admits that the above problem is serious and tries to solve it in his paper "Does A Rock Implement Every Finite State Automaton?" His proposed solution (as I understand it) is to fiddle with the italicized word "*divided*" in the above definition, so that the division must be into physical parts that *do not overlap*. This avoids Putnam's argument, though it may run into more sophisticated absurdities of its own. Indeed, some philosophers claim to have found such problems. (Schertz, "What it is…")

Interestingly, requiring the division of parts to avoid overlap would mean that a computer simulating a conscious brain in a virtual environment would not be conscious.[9] This explicitly contradicts one of the core claims of functionalism! Perhaps for this reason, Chalmers avoids saying that the POI is a necessary condition for consciousness; he presents it as merely a sufficient condition, so that simulated brains (with overlapping parts) could still be conscious. Chalmers himself admits that his solution is incomplete; he hopes to find a better version of POI that is both necessary and sufficient, that avoids absurdity while also endorsing the functionalist doctrine of simulated brain consciousness. ("Does a rock…" section 7)

Such attempts have been made, though perhaps not as successfully as Chalmers would like. Schertz, for example, has his own theory, as does Mallah. (Scheutz, "When Physical Systems…" and Mallah, "The Putnam-Searle-Chalmers Theorem") An overview of the different strategies can be found in the SEP. (Piccinini) As subsection 1.4 is about to discuss, sections 3 and 4 will contribute to this project by helping to focus it.

1.4 Methodology of this paper

This paper employs standard conceptual tools used in the literature on consciousness. For example, section 3 consists of science-fiction thought experiments, and section 4 is mostly intuition-

[9] In an ordinary serial-processing computer simulating a brain, the simulated neurons are not distinct physical parts of the computer, but rather overlapping parts. Even if they use different locations in memory—which they might not—they use the same processor to generate their causal relations with other neurons. This is especially clear for systems with virtual memory. (Chalmers, "Does a rock…" section 7)

based reflection on whether or not the systems imagined in them would be conscious. There are two things that this paper does which are mildly unusual and, hopefully, beneficial:

The first is to focus on transformations between hypothetical systems, rather than the systems themselves. The aim is to construct transformations that seem to preserve every important property save one; in other words, transformations that isolate a particular property for study. Then, when we apply our intuitions to whether or not this transformation preserves consciousness, we are applying our intuitions to whether or not that property is relevant to consciousness. This is important because if we try to apply our intuitions to the property in the abstract, we might pollute our intuitions by mistakenly imagining a scenario in which both that property and some other important property are lost.

The second is to focus on transformations that can be linked together to make a chain stretching from brains to ordinary rocks. By having a continuous series of transformations leading from common sense to absurdity, we both (a) narrow down the range of properties we must consider, to just those that the transformations in the chain fail to preserve, and (b) establish that at least one of the transformations in our chain fails to preserve consciousness. If nothing else, this is a helpful way to organize our thoughts and the debate as a whole.

Section 3 of this paper *tries very hard* to create a chain of transformations leading from brains to rocks such that no link in the chain is relevant to consciousness. If consciousness is a real property, then this task must be impossible; the hope is that we will learn more about consciousness by attempting it. Here is another way of putting it: There are so many differences between rocks and brains that we think at least one of them must be relevant to consciousness. Unfortunately, naive attempts to find such differences have failed, as Putnam's argument mentioned in 1.4 shows. (Putnam 120-125) This paper is a systematic attempt to find such differences, by scouring the conceptual space between rocks and brains. Section 3 can probably be improved upon: It is probably possible to construct a more detailed chain that isolates more properties more completely.

Section 4, which discusses each transformation in the chain, will attempt to find transformations that are relevant to consciousness. It will apply the General Argument, which Section 2 lays out, to those which are continua as a consideration against them being relevant to consciousness. In the process we will see more clearly how far, and in what ways, Chalmers' Argument truly generalizes. Section 4 has tremendous room for improvement: Most of the literature

on consciousness could be inserted to fill out the discussion about which transformations are relevant, and there is always a need for more thorough analyses of the transformations and the properties they purport to isolate.

Section 5 will discuss the implications of section 4. Though the results can only be described as preliminary so far, due to the aforementioned room for improvement, they do tentatively point us away from some theories and towards others.

2. THE GENERAL ARGUMENT

Many of the transformations I will discuss are what I call "Continua." This section will define what it is to be a continuum and present a general argument for the conclusion that consciousness is preserved across continua.

Recall that Chalmers' argument had essentially two components: (a) Consciousness doesn't disappear all at once as a result of one step in a transformation, and (b) consciousness doesn't disappear in portions over the course of many steps in a transformation. This General Argument extends Chalmers' argument in two ways. First, it adds a distinction having to do with time, and corresponding arguments, in order to more decisively address one objection (Searle's Hypothesis) that was raised against the original. Second, it goes into more detail about the reasons supporting (a) and (b). When doing so, it will explain Chalmers' justification for them as well.

2.1 The Definition of Continuum

A transformation is a Continuum iff it can be divided into steps, such that each and every step is both *physically isolated* and *indescribable* by our current best theories of consciousness.

> **Physically Isolated:** When described in the language of physics[10], the parts of the system which are not immediately internally affected by the change are never internally affected by the change. In other words, as far as physics is concerned, the change is entirely immediate. Perhaps higher-level properties like consciousness or aesthetic value will change gradually, but the low-level physical description of the system will not.

[10] Or, more accurately, the language of engineering or molecular biology.

> **Indescribable:** The step is small enough, and similar enough to its neighboring steps, that it is below the level of description of our current best theories of consciousness. (Note: this may not mean that it is *literally* indescribable) In other words, the kinds of entities and properties involved in the step operate on a much smaller scale than those that are involved in our current best theories. Talk of "levels of description" is analogical; the distinction being made here is not yet a precise one, but it should suffice.

Quite a lot hinges on whether or not a given transformation is a continuum. The definition was designed to be fairly uncontroversial when applied to the transformations in this paper. The controversy should instead arise in the next few subsections, where substantive conclusions are drawn about whether or not continua preserve consciousness.

2.2 The Chart

This paper is interested in measuring the change brought about by a given step in a given transformation. This change is measured relative to how things would have been without the step, rather than to how things were before it.[11] There are two important distinctions in how a step can result in change:

First, the step can either result in a Major Difference or a Minor Difference with respect to consciousness. For example, if a step results in a system changing from being conscious to not being conscious, that is a Major Difference. Similarly, if a step results in a system changing from experiencing a state of peace to experiencing a state of extreme agitation, that is a Major Difference. But if a step merely results in a system changing from being 100% conscious to being 99.99% conscious (if such notions of partial consciousness make sense), or if a step merely results in a system changing from experiencing bright red to experiencing slightly duller red, then that is a Minor Difference. The distinction between Major and Minor differences is not a sharp one, but it will suit my purposes.

Second, the change brought about by the step can happen entirely immediately, or some of it can happen over time, i.e. gradually. If a neuron is removed from the brain, there is an immediate effect (the brain is slightly smaller) but also a gradual effect, as the neighboring neurons react differently to the hole where their neighbor used to be. By contrast, if a neuron is replaced with a surrogate, there is an immediate effect (the brain is slightly more mechanical) but, since the

[11] Additionally, this change is not just in what actually happens, but in what might have happened as well. The same step might have different results due to indeterministic processes or something similar; what we are measuring is how the step affects the probability distribution over future events, rather than which events actually happen.

neighboring neurons will not notice, there is no gradual effect on behavior, or on the patterns of neuron-firings, or anything else that can be described on the level of physics. It remains to be seen whether this means there is no gradual effect on consciousness either; 2.5 will basically argue that there is not.

The following chart describes the resulting four ways that a step in a transformation can result in change. Helpful examples are given, with something besides consciousness as the subject matter, and with the major-minor distinction recast as applying to other properties besides consciousness.

The Chart	**(A) Immediate Change**	**(B) Gradual Change**
(1) Major Difference between systems develops over time; total change caused by step was major.	**1A:** e.g. what happens to a circle when a single point is deleted. (the circle is immediately destroyed)	**1B:** e.g. what happens when the proverbial last straw lands on the camel's back, causing an immediate but microscopic effect that gradually (over the course of many milliseconds) starts a chain reaction culminating in the collapse of the camel.
(2) Minor Difference between systems develops over time; total change caused by step was minor.	**2A:** e.g. what happens when a man with a full head of hair goes bald by losing hairs one by one[12]	**2B:** e.g. the effect of dropping sand grains on a trampoline on the depth of the well in the center.[13]

2.3 The General Argument

> **(P1)** If a continuum fails to preserve consciousness, it must do so in at least one of the four ways on The Chart: 1A, 1B, 2A, 2C.
>
> **(P2)** Every step in a continuum is *physically isolated* and *indescribable* by our current best theories of consciousness. (Directly from the definition of a continuum)
>
> **(P3)** History Doesn't Matter: Systems that are physically identical in every way save history are equivalent with respect to consciousness.

[12] Unless epistemicism about vagueness is true, in which case going bald by losing hairs is actually an example of 1A, and some other example must be used for 2A.

[13] Each grain of sand causes an immediate difference (the heap increases in size) and also a gradual difference (the equilibrium height of the trampoline shifts) but the total difference is minor.

(P4) If a continuum fails to preserve consciousness in manner 1B, then it also fails to preserve consciousness in manner 1A. (From P2, P3 & further argument)

(P5) It does not fail to preserve consciousness in manner 1A (From P2 & further argument)

(P6) It does not fail to preserve consciousness in[14] manner 2A or 2B (From P2 & further argument)

(C) Therefore, all continua preserve consciousness. (Directly from P1, P4, P5, and P6)

The argument for P1 has already been given in 2.2 and the argument for P2 has already been given in 2.1. In the following sections, the arguments for the other premises will be given.

2.4 P3: The History Doesn't Matter Assumption

This section presents reasons to believe **P3:** History Doesn't Matter: Systems that are physically identical in every way save history are equivalent with respect to consciousness.

For one thing, P3 is entailed by internalism about consciousness. We have good reasons to think that internalism about consciousness is true; see e.g. Ned Block (Block)

Yet P3 works with many forms of externalism as well. For example, if an organism must be "Currently in a normal environment" in order to be conscious, P2 will still hold. That being said, there are reasons to think that history does, in fact, matter. For example, perhaps some conscious mental states are representational, and perhaps what they represent is fixed by having had an appropriate causal interaction sometime in the past. Then there could indeed be two systems that physically differ only in their histories and yet are different mentally as well.

However, even this possibility is not fatal to the overall argument, since P2 could in fact be weakened further if necessary. Maybe history does matter—but as long as *the kind of history having to do with certain surgical part-replacements* does not matter, the argument will go through. Thus the aforementioned theory about representation might not go against the overall argument, since plausibly each part-replacement preserves the "Having once had an appropriate causal interaction" property.[15]

2.5 P4: 1B Implies 1A

[14] This could be weakened to "only in" and the argument would still go through.
[15] Our brain cells die and are replaced all the time. This suggests that replacement preserves consciousness.

This section argues that if a continuum fails to preserve consciousness in manner 1B, it also fails to preserve consciousness in manner 1A.

Suppose not. Then there is no step that causes an immediate major change, but there is a step (say, replacing neuron #345) that causes a gradual major change. Consider the latter. Imagine two identical systems (X and Y) that are undergoing this transformation and have reached the point right before this fateful step. Now perform the fateful step on X but not Y, and wait. As time passes, the gradual major change will materialize: X will be majorly different from Y with respect to consciousness.

Despite this, the two systems will still have all the same parts in all the same places behaving in all the same ways, except that X will have neuron #345 replaced. This is because of the *physically isolated* property of the continuum: The fateful step, whatever effects it may have had on consciousness, did not have a gradual effect describable on the level of physics.

Now, perform the same fateful step on Y, replacing Y's neuron #345. Since the only physically describable difference up until now was the fact that X had neuron #345 replaced and Y did not, there is no longer any physically describable difference—the two systems are physically identical. By the History Doesn't Matter assumption, they are equivalent with respect to consciousness as well. But that means that one of the two systems has undergone an Immediate Major Change, since the major difference between X and Y disappeared immediately. Since they both underwent the same transformation, one step on that transformation is an instance of 1A.[16]

2.6 P5 and the Implausibility of 1A (and 1B)

This section will attempt to justify P5 by appealing to the *indescribable* property of continua steps. In other words, this section will argue that it is implausible that a continuum would fail to preserve consciousness in a way that involved a single step causing an immediate major change in consciousness.

Chalmers actually has two arguments. One is that immediate major change implies that you could make consciousness "dance" in and out of existence by repeatedly replacing and then un-replacing a single surrogate neuron, which is absurd. I will not expand on this argument, since I do

[16] They are the same transformation because they involved the same steps in the same order. Transformations are not sensitive to how long you wait in between steps, since being independent of each other is how we individuate steps.

not think the "dancing" possibility adds any more absurdity than was already present in Immediate Major Change.

Chalmers' other argument is that immediate major change means that one neuron-replacement—or the replacement of an even smaller part; recall Note 1—is special; this seems to require an arbitrary and unparsimonious answer to the question "Which kinds of systems are conscious," which is implausible: ("Absent Qualia..." section 3)

> 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 quite arbitrary.

This argument works on me; I think it is a good articulation of a very strong intuition I have that instances of 1B are implausible for continua. I think I can expand on it slightly:

> (#1) If a step immediately results in a major change to consciousness, it must be because the ideal theory of consciousness says it would. (By contrast, if a step *gradually* results in a major change in consciousness, the ideal theory need not say so. After all, the ideal theory of whether-or-not-a-camel-is-standing-up cannot describe the addition of a single straw.)
>
> (#2) If a step is below the level of description of the ideal theory of consciousness, then the ideal theory does not say that the step would immediately result in a major change.
>
> (#3) Therefore, if a transformation consists of steps such that each and every step is at least that low-level, then the transformation cannot fail to preserve consciousness in manner 1A.
>
> (#4) Our current best theories are a good guide to the level of description on which the ideal theory will operate; that is, if a step is below the level of description of our current best theories, then it is probably below the level of the ideal theory as well.
>
> (#5) Thanks to the *indescribable* property of continua, every step in a continuum is below the level of description of our current best theories.
>
> (C) Therefore, continua cannot fail to preserve consciousness in manner 1A.

Premise #4 is the important one. Are our current best theories good guides to the ideal theory in that way? I think that intuition that they are is closely related to the intuitions about arbitrariness that underlie Chalmers' argument. Part of the reason that immediate major change in continua seems arbitrary is that our current best theories operate on a higher level of description, and vice versa.

2.7 P6 and the Implausibility of 2A & 2B

This section will present the reasons to think that P6 is correct. That is, it will explain why it is implausible that consciousness disappear in ways 2A or 2B.

The first thing to say is that the hypothesis we are considering—that consciousness disappears only in ways 2A and 2B—involves consciousness being a graded property. Since no step has a major effect, if your consciousness is to disappear it must happen in small pieces over many steps.

The second thing to say is you cannot notice this change. Chalmers' argument for this (in my words) is that if you noticed it, then you could behave differently as a result, but thanks to the *physically isolated* property, you cannot behave any differently than you otherwise would have. ("Absent Qualia…" section 3)

I think this is a good argument, since the first clause is supported by our notion of mental causation (if you notice something, you can decide to act as a result, and this will result in action) and the second clause follows from the definition of *physically isolated.* However, what Chalmers dubbed Searle's Hypothesis must be addressed: What if your decision to act after noticing the change cannot result in action? What if you find yourself powerless to communicate with the external world, listening to your own voice, operating without your permission, reassure people that you have not noticed anything?

Chalmers' response is that this sort of radical separation between the mental and behavioral realms is absurd. I agree, but I think I can go further: Not only is it absurd, it is an instance of 1B! To explain:

If whether or not you notice (and get disconnected from your behavior!) is a function of whether or not a certain step in the continuum has been reached, then clearly we have an instance of 1B, since getting disconnected from your behavior is clearly a major change, and it would all result from a single step. If it is a function of something else, such as your mental state at the time, then the picture is more complicated, but arguably still an instance of 1B. (See footnote.)[17]

[17] Imagine a multitude of copies of the same system (you), identical except that each one has undergone one more part-replacement than the previous one. There is one copy for every step on the continuum. Now, as time passes, some of them will have noticed and some will not. There will be an "earliest noticer" at any given time. After a very short time, each noticer will start to have radical thoughts like "Help, I'm trapped in my head!" So either we have instances of 1B, or no noticer is earliest for more than a very short time. The latter is absurd, because it would mean

Unnoticeable gradual disappearance of consciousness is very strange. It is not at all like other forms of gradual consciousness disappearance, such as those that accompany Alzheimer's or falling asleep.[18] One would like to think that we have good epistemic access to our own mental states; in particular, one would like to think that we know at least roughly how conscious we are. At least, so long as we are not inebriated, or very tired, etc. This is essentially Chalmers' argument against the possibility of 2A and 2B: That it would require a rational conscious agent to be radically mistaken about his or her own mental states, which goes against our notion of first-person authority. ("Absent Qualia…" sections 3, 4)

A further problem with gradual disappearance of consciousness is that it raises skeptical worries about whether or not we are fully conscious right now. After all, if I were only 10% conscious, I wouldn't know the difference! In particular, consider the possibility that *brains are not 100% conscious.* Maybe brains are only 40% conscious, and there is some other kind of system that is more conscious than brains!

Consciousness plays a role in our epistemology: If a system is not conscious right now, then I know that I'm not that system. If consciousness can come in unnoticeable degrees, then presumably this picture has to accommodate that, and the natural way to do it would be to say that the less conscious a system is right now, the more confident I can be that I am not it. Combining this with the above *brains are not 100% conscious* idea, we get an interesting argument for epiphenomenal dualism, or idealism, or some such radical view. (The details of the argument would depend on the details of our theory)

These ideas are interesting and worthy of further exploration, but arguably they are undesirable; accordingly, we should think twice before endorsing a graded notion of consciousness that allows for rational conscious agents to be radically mistaken about their own mental states, and that opens the door for brains to be less than 100% conscious.

2.8 Conclusion and Summary

---

that merely with the passage of time, more and more systems would notice each second, and after some finite amount of time has passed all of them would have noticed and become disconnected from their behavior, even the one that had zero parts replaced. I suppose one could bite the bullet and accept this conclusion, which would be: After some finite amount of time that is no more than, say, 100 billion seconds, (~3,000 years) and possibly much less, all humans suffer the horrible demise of Searle's Hypothesis! This would have implications for radical life extension technology at least.

[18] These are accompanied by changes in the rational capacities of the person involved, and they are noticeable.

This section has presented a general argument that transformations which are continua preserve consciousness. Immediate major change is implausible because it would require the ideal theory of consciousness to operate on a much lower level than our current best theories, drawing distinctions that seem to us to be arbitrary and completely irrelevant to consciousness. Meanwhile, gradual major change is implausible because it entails at least some immediate major change. Finally, if some of the steps contain minor change, then consciousness would diminish piece by piece in a way that could not be noticed, which goes against first-person authority and leads to interesting but unpleasant skeptical worries.

3. THE TRANSFORMATIONS

"The point of philosophy is to start with something so simple as not to seem worth stating, and to end with something so paradoxical that no one will believe it."

–Bertrand Russell

This section presents a series of thought experiments, dubbed Transformations. Each one describes a hypothetical procedure that transforms a system of one kind into a system of another kind. Strung together in a chain, these transformations constitute a way to transform an arbitrary brain into an arbitrary rock of similar size. This section will merely describe these transformations and whether or not they are continua; section 4 will discuss which transformations might be thought to preserve consciousness, and why.

This section will usually speak in the third person, about "the system" undergoing a transformation. However, it will also sometimes speak in the second person, talking about how you feel during the transformation, or about what happens to you. This is a stylistic choice; I would like you to imagine yourself undergoing these transformations, in the order that they are described.

I deliberately focus on the transformations towards the second half. This is because (a) I am mostly interested in functionalist/computationalist/POI theories, (b) the second half is the region on the chain that has been discussed least often in the literature, and (c) I am motivated by the specific question "Would uploads (i.e. conventional computers simulating brains in virtual environments) be conscious?" If they would, then at least one of the transformations in the second

half must fail to preserve consciousness, so the second half deserves close attention.[19] As a result, this section will get much more technical and detailed about halfway through.

3.1 The Robobrain Transformation

As described earlier in 1.2, the Robobrain transformation is what happens when all your neurons are replaced, one by one, with mechanical surrogates that behave the same way. Remember the two important notes: This transition could be done on a lower level, and it could be done on a wider selection of parts.

The Robobrain transformation is the archetypical example of a continuum. Each step in the transformation is *indescribable* because it is so small, and can be made even smaller if necessary. Meanwhile, each step in the transformation is *physically isolated* by hypothesis: Through dint of science fiction, each mechanical surrogate is as good as it needs to be at pretending to be a real neuron.

3.2 The EM Transformation

"EM" stands for Electronic Mind, which may be misleading. An EM is just like a Robobrain, except that instead of having many robotic neurons operating in tandem, it has some sort of conventional computer that simulates the neurons. That is, it has a few processors and memory disks that rapidly store and modify data, in a way that corresponds to what the robotic neurons would have done, but is structurally quite different: The processing is serial instead of parallel, for example.

The transformation to an EM is constructed as follows: Starting with a robobrain, take a *pair* of roboneurons and replace them with a computer chip that mimics both of them. The chip might have, for example, a memory where it stores the state of each neuron, and a processor which decides how to modify the states in response to input and the previous states, and which output to produce in response to the states. Once the first pair of roboneurons is replaced in this way, replace the pair formed by the resulting chip and one neighboring roboneuron, increasing the size of the chip so that it can serve as a surrogate for the other roboneuron as well. Repeat until all roboneurons have been "absorbed" into the computer chip—leaving us with a conventional serial-processing computer.

[19] This will become clear once the transformations are explained.

The EM Transformation is also a continuum. Each step in this transformation is *physically isolated* by hypothesis, just as before. The argument that each step is *indescribable* is this: The EM transformation can replace exactly the same (robo)neurons as the Robobrain transformation replaced neurons, in exactly the same order. So if the Robobrain steps were indescribable, then the EM steps will be indescribable as well.

3.3 Envatting

After having performed the EM transformation on you, we put you (or the computer that behaves like you, at any rate) in a vat, with sensors and stimulators hooked up in all the right ways to a big computer surrounding the vat. The big computer simulates a very convincing[20] virtual reality environment for you to interact with.

It is unclear whether or not Envatting is a continuum. Perhaps there is a way to divide it into smaller steps that are *physically isolated* and *indescribable*; perhaps not. This paper will conservatively assume that Envatting is not a continuum.

Helpful note: Up until this point we have been talking about systems which are in some sense directly comparable to brains. For example, a robobrain has the same boundaries, inputs and outputs as the brain it replaced. From here on out, we will talk about *larger* systems, systems like the combined brain+vat+supercomputer that we have now arrived at. Remember, what is being discussed is whether or not these larger systems contain consciousness, i.e. whether or not they have subsystems which are conscious.

3.4 Enviromerging

Having performed the Envatting step, we are left with a computer inside a skull inside a vat inside a big computer. Enviromerging is the process of replacing the skull and the vat with more computer chips, so that we are left with a computer (you?) inside a bigger computer, with no natural dividing line between them. The process is performed in the usual manner of part-replacement, where some tiny part is identified (say, a neuron-sized piece of eye tissue) and replaced with a tiny computer chip (equipped with the appropriate sensors and stimulators) that behaves the same way.

Since it is done in the usual manner of part-replacement, Enviromerging is a continuum for much the same reasons that the Robobrain transformation was. There are hundreds, if not hundreds of thousands, of distinct channels by which information flows to and from the brain.[21] Not only can

[20] Perhaps so convincing that you do not realize you have been Envatted.
[21] And that is where the information flow is narrowest.

each one of them be replaced individually, each one can be further divided into parts (e.g. particular neurons, particular photoreceptors, even particular parts of cells) all of which can theoretically be replaced by surrogates. This level of detail is below that of our current best theories of consciousness.

3.5 Isolation

In the Envatting transformation, we did not specify how the big computer interacts with the outside world. Isolation is simply the process of making sure that it does not interact in any meaningful way with the rest of the world. It has no user interface, no console, no sensors, and (if desired) no power plug.[22] Perhaps this step would be performed when the computer was constructed, in which case Isolation and Envatting would happen simultaneously.

As with the Envatting transformation, it is unclear whether or not Isolation can be made into a continuum. Perhaps there is a way to gradually limit the interaction between the big computer and the wider world, in a way that is *physically isolated* as well as *indescribable*, but this paper will not make that assertion.

3.6 GLUTification

GLUT stands for Giant Look-Up Table [Machine]. It can be thought of as the literal embodiment of a finite state machine. It is meant to be a more precise, more extreme version of the EM transformation. Note that this kind of GLUT machine is slightly different from the kind of Giant Look-Up Table that normally appears in *reductio* arguments against behaviorism: those GLUTs generally preserve the high-level behavior of the system but do not in any sense preserve the structure of the parts, whereas my GLUT does preserve the structure of the parts, at least under an unnatural interpretation. My GLUT is constructed as follows:

Divide up the system from 3.5 into tiny parts, such that each part is below the level of description of our current best theories of consciousness. Then take a *pair* of parts and replace them with a machine that mimics both of them. The machine is designed as follows:

The machine has four main components: A Look-Up Table, an Internal State Tracker, an Interaction Device, and a Messenger Web.

[22] It shouldn't matter whether it takes energy from the environment or generates its own; at any rate, we could construct it either way.

The Internal State Tracker (IST) is simply a receptacle that can display a single symbol at a time; it has N symbols on its repertoire, where N is the number of *pairs* of states of the two parts replaced. (Each symbol, therefore, corresponds to a *combined state* of the two parts replaced. In fact, while we are at it, let us stipulate that each symbol is a tiny picture of the two parts replaced, depicted in the combined state to which the symbol corresponds.)

The Look-Up Table consists of N entries. Each entry has two symbols, one in the "name" position and another in the "instruction" position. Each entry is "named" by the symbol in the "name" position; every symbol names exactly one entry.

The Interaction Device interacts with the outside world and changes the symbol on the Internal State Tracker accordingly. (The IST has a state for every combined state of the two parts, which is sensitive to what inputs they are in the process of receiving. So the Interaction Device makes sure that the IST is representing the right inputs at any given time.) The Interaction Device also reads the IST to figure out what output is being represented at any given time, and gives that output to the outside world.

The Messenger Web carries symbols back and forth between the IST and the Look-Up Table. It consists of N pathways, that each connect one entry to the IST.

The machine as a whole works as follows: Every time step, N copies are made of the symbol which is present on the IST. These copies travel down the N pathways (one each) to reach their respective entries on the Look-Up-Table. There, they are compared to the Name of their respective entries. Exactly one will match. That one will then be replaced by a copy of the symbol in the "Instruction" position in that entry. The new copy of the symbol in the "instruction" position will travel back to the IST, which will then change the symbol it displays to match the new copy. Then all copies will be erased, the Interaction Device will change the symbol so that it represents the right inputs, and the process will start over.

An appropriately designed GLUT could replace two parts as a surrogate, interacting with the neighboring parts in exactly the same way as the original pair of parts would have done.[23]

[23] The GLUT described above is deterministic. If we want an indeterministic GLUT we could make one. Though it would be more complicated, it would not be relevantly different. That being said, later transformations might require determinism to work. Might this be important? Perhaps consciousness depends on a fundamentally indeterministic

After the first pair of parts is replaced in this way, we replace the pair formed by the resulting machine and one neighboring part, and repeat until all parts have been "absorbed." This leaves us with one gigantic look-up-table, connected to the inputs and outputs in all the right ways, and behaving exactly as the previous system would have. As an aside, note that a GLUT made from an entire brain, or from something larger, would be truly gigantic; it would be larger than the known universe by far.

Note that with each replacement the inputs and outputs of the GLUT change, since some of the neighboring parts have been absorbed and since some new neighbors have been found. In the final step, no new neighbors are found due to Isolation; indeed there is no interaction with the outside world any more. Thus in the final step the Interaction Device disappears.

GLUTification is a continuum for much the same reasons that the EM transformation was a continuum. It is *physically isolated* because each time the replacement part is a surrogate; it is *indescribable* because the parts replaced are so tiny. Because the tiny parts replaced are bits of computers rather than bits of neurons, it is hard for us to establish the *indescribable* property with certainty, see the footnotes for a more careful argument to that effect.[24]

3.7 Lesser Symbol Relabeling

Before we continue, it will help to gather our thoughts by describing the system which has been reached thus far: a GLUTified Isolated Enviromerged Envatted EM. It is a gigantic, impenetrable black box, inside which there is a look-up-table, an internal state tracker, and a messenger web. The IST contains a symbol which is then relayed by the messenger web to the look-up-table which then relays back the appropriate instruction for how to update the symbol carried by the IST. This process repeats again and again very rapidly, which constitutes the operation of the system.

Lesser Symbol Relabeling is the process of rewriting all of the symbols in the machine so that they are no longer the tiny little pictures described earlier. Instead, we simply rename the entries

---

process, that cannot be captured by any deterministic surrogate. This would be very interesting and unexpected if true.

[24] Since the robobrain transformation was *indescribable,* the EM must simulate the brain on a level indescribable by our current best theories. Each simulated part corresponds to at least one part of the EM, so if we replace the parts of the EM one by one we are not replacing more than one simulated part at a time, which means that our replacements must be *indescribable.* All that being said, the fact that the EM transformation leads to overlap and serial processing might throw a monkey wrench into this.

on the look-up table with the numerals 1 through N, and then change all the remaining symbols accordingly so that everything behaves exactly as it did before. We do this one symbol at a time rather than all at once.

Since there are many more symbols on the GLUT than there are neurons in a brain, it is tempting to conclude that Lesser Symbol Relabeling is obviously a continuum. This temptation must be resisted, because the symbols on the look-up-table do not correspond at all to parts of the original brain. Each symbol is a representation of one state that the entire system can be in. So each step in this transformation is something that does, in a sense, affect the entire system. Nevertheless, each symbol is a representation *at a low level*; there are different symbols corresponding to each possible state of e.g. tiny pieces of cell wall in the original brain. Since the entries on the GLUT make distinctions that are below the level of description of our current best theories of consciousness, a step that changes one entry is arguably *indescribable*. Meanwhile, there is an obvious sense in which each step is *physically isolated*; after all, each step affects only one entry.

3.8 Symbol Diversification

In this transformation, the Internal State Tracker is modified so that instead of having one symbol for each entry on the Look-Up Table, it has up to NxR many symbols, which are grouped into N categories of size somewhere between one and N. (R is some large number) Whereas before each symbol the IST could display corresponded to a symbol in the name of an entry on the Look-Up Table, now each symbol the IST can display is part of a category, which in turn corresponds to a symbol on the name of an entry on the Look-Up Table.

All the symbols in a given category are related in a simple way to the symbol on the entry that their category corresponds to, so that the Messenger Web can easily figure out whether or not a given symbol from the IST is part of the category that corresponds to a given entry name.

Note that in order to keep the system behaving exactly as it did before, we need to give the IST a rule for updating when presented with instructions by the Web. After all, there are now multiple symbols that it could update to, that are all equally valid and would all lead to the same result. We give it any old rule for updating, since the behavior of the system as a whole will be the same regardless; we will discuss later whether this is relevant to consciousness.

Since this transformation can be done piecemeal, entry by entry, the argument for Symbol Diversification being a continuum is identical to the argument for Lesser Symbol Relabeling being a continuum.

3.9 Greater Symbol Relabeling

This is the most complicated and technical transformation in the paper. For each symbol on the Look-Up Table, make the following three changes:

(a) Change the symbols that the IST carries, so that they each consist of two parts: One part, dubbed the "Dial," can be in N different configurations, where N is the number of entries in the GLUT. The other part, dubbed the "Clock," can be in R configurations. The total number of symbols the IST can display is NxR, just as before. The configurations of the clock have a natural order to them; that is, there is a simple rule that well-orders them. For example, perhaps the configurations of the clock are simply printed Arabic numerals.

(b) Change the symbols that are written in the GLUT, so that they consist of a finite list of ordered pairs, where the first member of each pair matches a state of the Clock and the second member of each pair matches a state of the Dial. The ordered pairs in the list must be chosen carefully:

Define the function T(s)→s such that its domain and range are both entirely GLUT entries, and such that T(x) = the y such that the entry which x names has y as its instruction symbol.

Recall that there are exactly N states of the "dial." Find some bijective mapping between them and the entries on the look-up table.

Use the following recursive function F to map ordered pairs to entries. Let {X,D} be an ordered pair with the clock state which is ranked X by the well-ordering and the dial state which maps to entry D.

F({0,D}) = T(D)
F({C+1,D}) = T(F({C,D}))

Putting it all together: The set of ordered pairs that any given entry has as its symbol will be the set of ordered pairs that are mapped to it by F.

(c) Change the machinery of the IST and the Messenger Web appropriately, so that everything runs exactly as it did before. This isn't a dramatic change: it simply involves

(1) modifying the Messenger Web so that it searches the table not for an identical symbol, but for a symbol that includes an identical symbol as a part. (specifically, a symbol that has an ordered pair that matches the "Clock" and "Dial" the machine-head carries.)

(2) modifying the IST so that it updates the symbol it displays in the following way: It reads the relevant messenger symbol (the instruction), and updates the symbol it carries to match one of the ordered pairs listed the messenger symbol. There may be multiple such pairs in the output symbol. Since the next step will be the same regardless, arguably it doesn't matter which pair is picked. So we specify that the machine-head always picks the one that involves incrementing the "Clock" by one according to the well-ordering, and keeping the "Dial" constant. Thanks to the setup, there will always be such an option, until the Clock cannot be incremented any more.[25]

Since this can be done piecemeal, one entry at a time, the argument for Greater Symbol Relabeling being a continuum is identical to the argument for Lesser Symbol Relabeling being a continuum.

3.10 Lesser Process Simplification

The previous three transformations have changed the symbols used in various ways, but they have not altered the behavior of the system at all, in the sense that the GLUT continues to update from symbol to symbol in response to the entries that it communicates with via the Messenger Web exactly as it did before. However, we have now brought about this important feature:

> **Important Feature:** Every time the IST interacts with the messenger instruction symbol, it will end up updating the symbol it carries simply by incrementing the "Clock" part by one, and leaving the "Dial" the same.

Having established this, Lesser Process Simplification eliminates the redundancy by having the IST not even bother to interact with the messenger instruction symbol; instead, it just automatically updates the symbol that it carries when the messenger arrives, incrementing the "Clock" part by one and leaving the "Dial" the same no matter what the messenger says.

[25] If the finitude of the clock seems problematic, do not worry. Section 4.9 will discuss this.

Lesser Process Simplification is arguably a continuum for the same reasons discussed in the previous sections.

3.11 Greater Process Simplification

This final transformation stops the Messenger Web from interacting with the look-up table. Why bother, if the instructions gleaned will not be read? The Messenger Web continues to send signals back and forth between the look-up table and the IST, but they are blank. The IST updates just as before.

This too is arguably a continuum, for the same reasons discussed in the previous sections.

3.12 Table Encryption

This transformation encrypts the table so that the symbols on it no longer directly resemble the symbols in the IST. Instead, a translation (decoding) process must be followed to make the connections. There is of course no actual translation manual; it is just what we would have to do to make the connections. This encryption can be as light or as heavy as we want. It could be as light as swapping the position of the "clock" and "dial" components of all the ordered pairs on the table, or as heavy as encrypting the symbols on the table with a one-time pad, i.e. randomly scrambling them. We could even start by lightly encrypting it and then gradually increasing the level of encryption until the limit is reached.

The IST is unchanged by this transformation, which is also arguably a continuum for the usual reasons.

3.13 Table Deletion (CADification)

This transformation deletes the table and the messenger web leaving only the IST which now updates at the same speed anyway due to an internal clock. What we have now is called a Clock-And-Dial Machine, or CAD. This too is arguably a continuum for the usual reasons.

3.14 Rockification

By the time we get to a Clock-and-Dial machine, something clearly has gone wrong. You and your consciousness are surely long gone. However, the dial part of the clock-and-dial would need to have as many possible states as the original brain—i.e. astronomically many. Thus we are still dealing with outlandish hypothetical systems that are probably impossible for us to construct, and so accepting that such a clock-and-dial would be conscious will not have any practical implications.

This section attempts to bring our journey through science-fiction thought experiments back home, by building (and crossing) a bridge between the aforementioned clock-and-dial and an ordinary, real-life terrestrial rock.

The clock-and-dial system consists of a dial with astronomically many possible states, which does not change over time unless changed by an outside force, and a clock with some "large number" (R) of possible states, which it deterministically marches through unless changed by an outside force.

The internal structure of a large rock does not change much over time. The atoms inside may vibrate and exchange electrons, but the relative positions of, say, 10,000-atom clumps of rock do not change without outside intervention. Thus, a rock of any size can be conceptually divided up into tiny cubes of 100,000 atoms each. (much smaller than a neuron, or even a part of a neuron) Each cube can be thought of as being in one of two states: "Pure" or "Impure." These designations can be defined however we want, so long as whether a cube is classified as Pure or Impure depends only on the internal structure of the cube, and so long as the laws of nature ensure that the cube's classification will not change without outside intervention. What this conceptual division and classification gives us is a method for delineating possible states of a rock, such that each tiny cube within the rock has two possible states, and the rock as a whole has 2^C possible states, where C is the number of tiny cubes within it. Since each cube is so small, a rock of about the same size as a human brain would have many more cubes than it would have neurons, or even bits of cell wall. Thus it would have more than enough possible states to match the N states of the astronomical dial.

Can we imagine transforming an astronomically large dial into a brain-sized rock? Yes. Begin by removing half of the dial, cutting the number of states it has by two, and simultaneously graft on a tiny cube of rock. Then repeat until eventually the dial has only two states and the rock is missing only one cube; finally, replace the two-state dial with yet another tiny cube of rock. We now have a clock-and-a-rock.

Transforming the clock is trickier, but we do not need to worry about it, because the clock need not be astronomically large. Even if ordinary rocks are not by themselves conscious, if it turns out we can make them conscious by grafting a suitable clock onto their sides, they might as well be.

That being said, there are probably many natural processes that deterministically pass through a large number of states without outside intervention, such as the radioactive decay of a lump of uranium.[26]

The argument that Rockification arguably a continuum is the same as usual. Though it changes many entries at a time, each change is below the level of description of our current best theories in the same sense that each change in e.g. Symbol Diversification was.

4. THE PROPERTIES

In this section I will discuss whether or not the above transformations preserve consciousness. If a transformation preserves all properties that are relevant to consciousness, then it preserves consciousness. Each transformation is designed to isolate a single property; technically, infinitely many properties are lost in every transformation, so this is more an art than a science. I will go through each transformation and discuss properties that they fail to preserve that plausibly might be relevant to consciousness.

4.1 Does The Robobrain Transformation Preserve Consciousness?

One potentially relevant property that Robobrain does not preserve is *Biology*. The brain is biological; the robobrain is mechanical. While some (e.g. van Inwagen) have argued that this is relevant to the preservation of consciousness, there are reasons to believe it is not—for example, the General Argument applies to it, since it is a continuum. For reasons of space I will not discuss this issue further; I will simply say there are good arguments for both sides and direct you to the literature. (See footnote)[27]

4.2 Does The EM Transformation Preserve Consciousness?

When thinking about the plausibly relevant properties that this transformation fails to preserve, one immediately comes to mind: "The parts." That is, whereas the Robobrain transformation preserved the internal structure of the brain, in that it preserved each part of the brain (at a very low level of description), the EM transformation changes the internal structure by consolidating parts.

[26] This is not technically deterministic, but it might as well be.
[27] See e.g. the literature on Personal Identity surrounding Animalism. (Olson)

This is technically not true. There is a way to divide up the EM into parts such that each one maps in all the right ways to an old roboneuron. This division would be very complicated; the parts of the EM thus described would probably overlap with each other and would certainly be very unnatural. So technically, the properties that this transformation fails to preserve are *having non-overlapping parts that meet a certain description* (e.g. the description of the neural network of a rational adult brain) and *having naturally describable parts that meet a certain description.*

As mentioned in 1.3, Chalmers' solution to Putnam's problem is to say that *having non-overlapping parts that meet a certain description* is essential to consciousness. The problems with this approach are (a) it is not obvious why it should be relevant to consciousness, so an argument to that effect must be made, (b) this transformation is a continuum, so the General Argument must be defended against, and (c) the Lesser Symbol Relabeling transformation constitutes an intuition pump against this property being relevant to consciousness. (See section 4.7)

As for the property of *having naturally describable parts that meet a certain description*: Some philosophers (e.g. Lewis) have claimed that naturalness is an important property in various other philosophical fields; it is not too much of a stretch to claim that it is also relevant to consciousness in this way. (Weatherson) However, both problems (a) and (b) apply again here.

4.3 Does Envatting Preserve Consciousness?

Envatting need not change the internal properties of the purportedly conscious system, so only externalists can complain about it. Complain they will though; some philosophers have famously said that brains envatted in this manner would not be conscious, or at least not in the same way we are. (Brueckner) There are plausible arguments for this conclusion, having to do with the representational nature of consciousness. Since Envatting is arguably not a continuum, Externalism comes out looking pretty good; this paper could be read as an argument for a sort of externalism, though there are many well-known problems with externalism that should keep us looking for other solutions. (Block)

One of these problems is that Envatting seems to be no different in kind from the sort of virtual-reality devices that have already been tested, which we want to say preserve consciousness. Consider Oculus Rift, a recent virtual-reality headset. Imagine someone putting it on in a sensory deprivation chamber, and moving about in a detailed virtual environment. With current technology, they would still be able to tell the difference between that and real life, but as the technology

improves we will one day reach a point where only some people can tell the difference, and then after that a point where no one can. That point would be equivalent to envatting; indeed, it would literally be envatting applied to the whole body instead of just to the brain. (Recall that these transformations could be applied to the whole body if desired, instead of just the brain.)

If we decide that envatting does not preserve consciousness, then the implications are enormous. First, it calls for more research into the distinction between virtual reality devices and envatting. Second, this (probably) implies a view of consciousness under which skeptical hypotheses like "I'm a brain in a vat" are impossible, which is a plus. (Brueckner)

4.4 Does Enviromerging Preserve Consciousness?

One property not preserved by the Enviromerging transformation is the naturalness of the distinction between the purportedly conscious system and its immediate environment. Before the transformation, there were several relatively distinct layers of material between the system and the environment—neurons, blood, normal cells, skin, hair, air—but after, there is only more computer chip. The same considerations about naturalness discussed with the EM transformation apply here.

Another property not preserved by the Enviromerging transformation is the specification of the inputs and outputs of the system. When the eyeball is replaced by a mechanical surrogate, one might say that this merely changes the medium in which the input is received, having no effect on the consciousness of the system. However, if we instead say that the inputs themselves are changed, in a way that might affect the consciousness level of the system, then perhaps we ought to be suspicious of e.g. prosthetic limbs, especially those closely connected with the brain. In the same sort of way that virtual reality technology could be improved gradually until the point where it is equivalent to envatting, so too can prosthetic technology be improved tradually until it is equivalent to enviromerging, at least when applied to someone already envatted.

Finally, since Enviromerging is a continuum, saying that it fails to preserve consciousness involves defending against the General Argument.

4.5 Does Isolation Preserve Consciousness?

There are very strong reasons to think that Isolation preserves consciousness. For one thing, Isolation can be done to ordinary humans, and often is—e.g. when a mine collapses, trapping a miner or group of miners underground. It would be very radical to claim that those people lost consciousness!

For another, Isolation does not change the immediate environment of the purportedly conscious system, so objecting to it requires a particularly strong form of externalism—one that involves the non-immediate environment perhaps.

Finally, Isolation does not technically remove all interaction with the outside world—it merely removes all easy, natural interaction. After all, it is possible that the outside world will (using some science-fiction technology perhaps) overcome the barriers and make alterations to the system. For example, if a powerful outsider wanted to make the envatted, enviromerged, isolated EM have a blue experience, the outsider could surgically alter the relevant parts of the supercomputer so that the simulated environment appears blue.

4.6 Does GLUTification Preserve Consciousness?

Like the EM transformation, GLUTification results in a system which engages in serial processing rather than parallel processing. Like the EM transformation, GLUTification destroys (or at least reduces) the naturalness of the relevant parts of the system. Which properties are lost by GLUTification that are not also lost by the EM transformation?

The most plausibly relevant difference between the EM transformation and GLUTification seems to be that the latter uses a "brute-force" method to change from internal state to internal state (each entry in the table explicitly directs the machine-head to the next entry) whereas the former has a more nuanced, more efficient method. This may be an interesting avenue for exploration, but *prima facie* it does not seem promising: The difference between "brute force" methods and whatever it is that the EM is doing is probably a difference in degree, not kind, and it is unclear why efficiency should be relevant to consciousness anyway.

Another possibility is that GLUTification is problematic not because of any intrinsic difference between it and the EM transformation, but because it was performed *after* enviromerging, envatting, and isolation whereas the EM transformation was performed *before*. This is an interesting direction for further research, but this paper has not developed it any further thus far.

Finally, since GLUTification is just as much a continuum as the EM transformation, any claim that it fails to preserve consciousness must defend against the General Argument.

4.7 Does Lesser Symbol Relabeling preserve consciousness?

This transformation fails to preserve one property that, technically, was already lost in the EM transformation: *having non-overlapping parts that meet a certain description.* This transformation was created in part to illustrate the implausibility of this property:

First, imagine that we perform GLUTification directly on a robobrain, without it being envatted, isolated, etc. The resulting GLUT would still have an Interaction Device. Recall that the original GLUT had as symbols tiny pictures of the states of the parts being replaced. Thus, each replaced part can be mapped to all the tiny pictures of itself in all the symbols of the GLUTs. In other words, we can define objects (the collections of tiny pictures of an original part) that are very unnatural, and yet that do not overlap, and that retain all the relevant causal relations with each other to qualify as meeting *a certain description*. We could then perform Lesser Symbol Relabeling on it, and it would lose the property of *having non-overlapping parts that meet a certain description.* But it is very unintuitive that such a cosmetic change would be relevant to consciousness!

Stepping out of this imagined situation, and back into the Lesser Symbol Relabeling transformation as originally described, it seems that there is only one property it fails to preserve that might be relevant to consciousness. That being said, as mentioned in subsection 3.7, the argument that Lesser Symbol Relabeling is a continuum is actually weaker (because less clear) than the argument that previous transformations were continua. This point should be kept in mind as we discuss the next 6 transformations, since they all rise or fall as continua together. We will assume for convenience that the next 6 sections are continua, but remember that this assumption is shakier than the previous continuum-claims.

4.8 Does Symbol Diversification Preserve Consciousness?

Arguably Symbol Diversification is a redundant step, because of the everyday imprecision of physical objects. For example, the pre-diversification Internal State Tracker was probably capable of sporting scratched or otherwise slightly modified symbols that would still function normally. So already we had a machine that would treat many different, but similar symbols as if they were one. If this is so, then Symbol Diversification is nothing new.

Even if this is not so, it is hard to argue that Symbol Diversification fails to preserve consciousness. It is a continuum, and we have no intuitive reason to consciousness depends on whether or not there are multiple slightly different symbols being treated as just one symbol.

4.9 Does Greater Symbol Relabeling Preserve Consciousness?

Greater Symbol Relabeling, just like its lesser predecessor, is (with one exception, to be discussed) a purely cosmetic change to the system. All the symbols are changed, but each old symbol is changed to precisely one unique new symbol, and the causal dependencies and even spatial relationships between the symbols are preserved. Syntactically, nothing changes. It is true that now we have specified the rule that describes how the IST updates, (it always increments the Clock by one and holds the Dial constant) but since previously we said that any old rule would do, this need not be syntactically any different from what was already being done.

The one non-cosmetic change is this: Whereas before the system could operate indefinitely, or at least for some long period of time before running out of energy or breaking down, now the system has a definite lifespan: When the clock can no longer be incremented, when it has run through all R configurations, the whole system grinds to a halt. This may seem like a big deal, but actually it is not, for the following reason: First of all, R can be as large as we want it to be (it was defined as "some large number" in subsection 3.7) so we can set R to be so large that the system will break down for other reasons, like lack of fuel or the end of the universe, long before it runs out of clock states. Secondly, the Greater Symbol Relabeling transformation can actually be rewritten in a more complicated way, to allow for *indefinite* extension of the clock, and/or for *looping* of the clock with some large but finite R.[28]

Finally, it must be said that in this transformation naturalness takes another blow, in a new way. Already the naturalness of the description we would need to describe [a part of] this system as conscious was lowered in e.g. the EM and GLUTification transformations. Now, it is lowered again, since the symbols that are grouped into categories no longer *naturally* resemble each other. (Note, the entry symbols do naturally resemble the clock-and-dial symbols in the IST. But the clock-and-dial symbols in the IST are completely unrelated to their fellow category members, as far as natural relations are concerned.)

4.10 Does Lesser Process Simplification preserve consciousness?

The interesting thing about this transformation is that it destroys the causal dependency of the IST state on the entries in the table. Previously, the IST updated in response to what was written

[28] The former is how the transformation was originally written; the latter is less explored. The latter has to do with finding the repeat length for each entry (it is conveniently listed in each entry symbol) and then using that to tell the clock when to reset back to 0. Of course, this would then be lost in later transformations.

in the instructions on the table; now, it updates automatically. It is guaranteed to update in the same way as before, but plausibly something has been lost.

Since some kinds of causal dependencies are plausibly relevant to consciousness, this loss of causal dependence may prove to be relevant to consciousness. The case for this must explain why the look-up-table entries are important; after all, when we delete other "redundant" features of systems we generally do not think anything important is lost. The trick is to articulate the property that is lost, in a way that avoids obvious counterexamples. After all, the systems prior to GLUTs had no causal dependency between look-up-table entries and IST states either.

4.11 Does Greater Process Simplification preserve consciousness?

The interesting thing about this transformation is that it further destroys the naturalness of the way to characterize the system state. Previously, we could characterize the system state by the category that the IST symbol belonged to, or we could characterize it by the last instruction symbol carried by the messenger web. Now that the messenger web doesn't bother reading the entries, it treats them all the same, and the only way to characterize the system state is the first way: the category that the IST symbol belongs to. But these categories are extremely unnatural; they are generated by the function F in subsection 3.9. That being said, if we are allowed to make reference to the look-up-table when defining the categories, they are not so unnatural after all: An IST symbol belongs to the same category as another IST symbol iff the ordered pairs in the entry symbols that they correspond to are found together in the same entry.

If considerations about naturalness were relevant to consciousness, we probably would have decided that consciousness has been lost by now. Nevertheless this paper will continue to describe naturalness considerations.

4.12 Does Table Encryption preserve consciousness?

In the previous transformation, there were two ways to define the states of the system: Directly from the IST, using the function F, and indirectly from the IST, by making reference to the look-up-table. Encryption gradually increases the unnaturalness of the second option until it equals the unnaturalness of the first.

It would be strange if this was relevant to consciousness, however, because the table is not causally related to the IST in any meaningful sense.

4.13 Does Table Deletion (CADification) preserve consciousness?

Deleting the look-up-table entirely leads us to a situation that is clearly problematic for consciousness; the only serious question is whether consciousness was lost at this stage, or earlier.

Naturalness does not even decline in this stage, since a maximally encrypted look-up-table is no more helpful than empty space when it comes to defining the states of the system.

Many similar things can be said here as were said in section 4.9. Deleting the look-up-table entirely can be thought of as eliminating yet another redundancy, or it can be thought of as eliminating something important. But what, exactly, is being eliminated, and why is it important?

4.14 Does Rockification preserve consciousness?

Yet another drop in naturalness occurs in this transformation. Weird fusions of possible combinations of clock states and dial states are still more natural than weird fusions of possible combinations of clock states and rock states.

5. CONCLUSION

Perhaps, as you read through section 4, you developed or maintained a clear sense of which transformations failed to preserve consciousness, and why. I did not: Many of the transformations seem fishy to me, but I cannot articulate why any of them would be relevant to consciousness, and so each of them seems to me to preserve consciousness. Yet I think that rocks are not conscious (and that brains are) so that cannot be right. This is a problem.

This problem does not yet have an iron grip. The arguments I gave against their relevance to consciousness often relied heavily on intuition, and I certainly did not rule out the possibility of further arguments that could show one or more transformations to be relevant to consciousness after all. Further research in this area, I expect, will either solve the problem or make it stronger.

This section will briefly explain some of the implications of this problem and this paper more generally.

5.1 Externalism looks better

This paper makes externalism look good for a variety of reasons. For one thing, externalism allows us to reject the History Doesn't Matter assumption (P3), which in turn allows us to reject the General Argument. This gives us many more options on the menu of transformations to choose from, and it even allows us to get away with choosing none of them: Perhaps all the transformations do in fact preserve consciousness, but an ordinary rock would only be conscious if it had actually

been transformed from a brain, and that will probably never happen. Externalism also gives us an argument that Envatting is relevant to consciousness.

Of course, externalism is not one theory, but rather a term for a type of theory; thus not every externalist theory will grant all of the above advantages. The point I am trying to make is that these advantages seem appealing enough to motivate increased interest in externalist theories.

5.2 Further-fact and no-fact theories of consciousness look better

Theories of consciousness can be divided into three categories:

Further Fact: In addition to facts about the physical properties[29] of a system, there are facts about whether or not the system is conscious. There are of course correlations between the two; perhaps these correlations can be described by Laws of Nature.

Reducible Fact: (reductionist) There are facts about e.g. whether or not a system is conscious, but they are nothing over and beyond the facts about the physical properties of the system.

No Fact: (eliminativist) There are no facts about e.g. whether or not a system is conscious. When we think about consciousness we are simply confused, or else playing word games, or else making subjective value judgments.

Further Fact and No Fact theories each have an advantage over Reducible Fact theories, when it comes to dealing with the problem presented in section 4. This is because they each have an escape route:

No Fact theories can accept that none of the properties discussed seem relevant to consciousness; they say that there is no such thing as consciousness, and that in thinking about it we are either confused, or making some sort of subjective judgment. Thus we do not have to look for an explanation for why brains are conscious but rocks are not.

Further Fact theories can also accept that every transformation seems to preserve consciousness, because they think that consciousness really depends on something that was not mentioned in the description of the transformations. If consciousness depends on facts that are not physical, then a physical description of a system and a physical description of that system changing into another system will not tell us anything about consciousness. At least, not directly. Perhaps, with the help of Laws of Nature, we could deduce non-physical facts that would then tell us about

[29] As usual, this means "the properties describable in the language of physics."

consciousness. But if we do not yet know those Laws, it makes sense that we would not be able to find any physical properties that seem relevant to consciousness.

5.4 Explore Ways to Escape the General Argument

The General Argument seems to apply to most of the transformations in this paper; it says that all continua preserve consciousness. This gives us reason to focus on those transformations which are not continua, but it also gives us reason to reconsider the General Argument.

I envision this involving both (a) an exploration of the notion of partial, unnoticeable consciousness, particularly with the goal of deciding whether or not other systems like immaterial souls would be more conscious than brains, and (b) an exploration of the idea of an immediate major disappearance in consciousness, with the goal of deciding how likely it is that an ideal theory would be able to justify such a disappearance in a simple, non-arbitrary way.

ACKNOWLEDGEMENTS

This paper was made possible in part by discussions and correspondence with my friends, among them Ramana Kumar, Conor McMahon, Jeff Alstott, David Borg, Ryan Dominguez, Katja Grace, John Aspden, and Patrick Miller. Similarly, this research was funded through the summer of 2013 by a grant from the Glynn Family Honors Program. Finally and most importantly, I would like to thank my advisor, Jeff Speaks, for helping me through this entire process in multiple capacities.

BIBLIOGRAPHY

Bishop, Mark. "Counterfactuals cannot count: A rejoinder to David Chalmers." *Consciousness and Cognition*, 11:642-52, 2002. Web. 02 Apr. 2014 <http://www.doc.gold.ac.uk/~mas02mb/Selected%20Papers/2002%20Counterfactuals%20cant%20count.pdf>.

Bostrom, Nick. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." Mind & Machines (2006) 16:185–200 Web. 02 Apr. 2014. http://www.nickbostrom.com/papers/experience.pdf

Brueckner, Tony, "Skepticism and Content Externalism", *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2012/entries/skepticism-content-externalism/>.

Block, Ned. "The Real Trouble for Phenomenal Externalists." Forthcoming in *Phenomenology and the Neurophilosophy of consciousness.* Web. 03 Apr. 2014. <http://www.academia.edu/2744957/The_Real_Trouble_for_Phenomenal_Externalists>.

Chalmers, David. "Absent Qualia, Fading Qualia, Dancing Qualia." Published in *Conscious Experience*, edited by Thomas Metzinger. Imprint Academic, (1995). Web. 02 Apr. 2014. http://consc.net/papers/qualia.html

---. "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108: 309-33. (1996). Online. http://consc.net/papers/rock.html

---. "The Puzzle of Conscious Experience." *Metaphysics: The Big Questions,* Zimmerman & van Inwagen ed. (2008): 393-401. Print.

---. "A Computational Foundation for the Study of Cognition."*A Computational Foundation for the Study of Cognition*. Forthcoming in the Journal of Cognitive Science (2012), n.d. Web. 02 Apr. 2014. <http://consc.net/papers/computation.html>.

Chisholm, Roderick M. "Which Physical Thing am I?" *Metaphysics: The Big Questions,* Zimmerman & van Inwagen ed. (2008): 328-333. Print.

Chrisley, Ronald J. "Why Everything Doesn't Realize Every Computation." *Minds and Machines* November 1994, Volume 4, Issue 4, pp 403-420. Web. 02 Apr. 2014 <http://www.sussex.ac.uk/Users/ronc/papers/chrisley-computation.pdf>.

Godfrey-Smith, Peter. "Triviality Arguments Against Functionalism." *Philosophical Studies* (2008). Web. 02 Apr. 2014. <http://www.petergodfreysmith.com/TrivArgtsFnm-08-Zweb.pdf>

Van Gulick, Robert, "Consciousness", *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/spr2014/entries/consciousness/>.

Lycan, William G. *The Journal of Philosophy*, Vol. 78, No. 1 (Jan., 1981), pp. 24-50 Published by: Journal of Philosophy, Inc. Article DOI: 10.2307/2025395 Article Stable URL: http://www.jstor.org/stable/2025395

Mallah, Jacques. "The partial brain thought experiment: partial consciousness and its implications." (2009) Web. 02 Apr. 2014 <http://cogprints.org/6321/1/PBA-09.pdf>.

---. "The Putnam-Searle-Chalmers Theorem." (2011) Web. 02 Apr. 2014 <http://onqm.blogspot.co.uk/2011/10/putnam-searle-chalmers-theorem.html>.

Olson, Eric T., "Personal Identity", *The Stanford Encyclopedia of Philosophy* (Winter 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2010/entries/identity-personal/>.

Piccinini, Gualtiero, "Computation in Physical Systems", *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/computation-physicalsystems/>.

Putnam, H. *Representation and Reality*. Cambridge, MA: MIT, 1988. Print.

Russell, Bertrand. "The Philosophy of Logical Atomism" (1918), in *LK* pp. 177–281 and *CPBR8* pp. 157–244.

Scheutz, M. "What it is not to Implement a Computation: A Critical Analysis of Chalmers' Notion of Implementation." *Journal of Cognitive Science*. Vol 13 issue 1. pp. 75-106. Institute for Cognitive Science, Seoul National University (2012)

---. "When physical systems realize functions." *Minds and Machines, 9*, 161–196 (1999).

Searle, John R. "Is the Brain a Digital Computer?" *Is the Brain a Digital Computer?* University of Southampton, 20 Mar. 1996. Web. 02 Apr. 2014. <http://users.ecs.soton.ac.uk/harnad/Papers/Py104/searle.comp.html>.

Turner, Raymond, "The Philosophy of Computer Science", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/sum2014/entries/computer-science/>.

Weatherson, Brian, "David Lewis", *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/david-lewis/>.