What Panpsychists Should Reject: On the Incompatibility of Panpsychism and Organizational Invariantism

Keywords: Panpsychism, Russellian Monism, Organizational Invariantism, Conceivability, Consciousness, Chalmers.

On the one hand, materialists who find conceivability arguments compelling and those with dualist inclinations who, believing in the causal closure of Physics, do not want to render consciousness epiphenomenal might find in Panpsychism (PP)—roughly, the thesis that the mind is ubiquitous throughout the universe—an interesting route to explore.

On the other, there are good reasons for believing that Organizational Invariantism (OI)—the principle that holds that two systems with the same (sufficiently) fine-grained functional organization will have qualitatively identical experiences—is true.

Some philosophers, like David Chalmers, have either shown their sympathy for both principles or explicitly endorsed them. The purpose of this paper is to show the tension between the arguments that back up both principles. This tension should lead, or so I will argue, defenders of one of the principles to give up on the other.

The paper is structured in three sections. Section 1 is devoted to motivate PP. I will briefly sketch the conceivability argument as presented by David Chalmers and provide some reasons in favor of endorsing PP for those convinced by the argument. Section 2 deals with the principle of OI and outlines the dancing and fading qualia arguments also offered by Chalmers to support the principle. Finally, in section 3, I will argue that there is a tension between PP and OI; the same argument that backs up OI might be used, *mutatis mutandi* as I will show, to

argue against PP. I conclude that defenders of PP should give up on OI and those who believe that OI is true should reject PP.

1 Panpsychism (PP)

Physics only tells us about structures and functions; it remains neutral on the intrinsic nature of the fundamental entities (quarks, leptons, bosons, strings or whatever physics will ultimately determine) that give rise to macroscopic entities like chairs, tables, humans, etc. Panpsychism can be characterized as the doctrine that the mind is a fundamental feature of the world, it exists throughout the universe: the most fundamental entities enjoy mentality.

In this paper I am interested in a particular kind of mental states, (phenomenally) conscious ones. If one believes that there is an important distinction between conscious and unconscious mental states, leaving aside other mental properties, then one should endorse a less radical view that can be called 'Panprotopsychism'. Panprotopsychism can be roughly presented as the claim that the microphysical fundamental entities of the actual world, once properly related to each other, give rise to all sort of "physical entities" and due to its intrinsic properties also to consciousness. I will have this form of panpsychism in mind along the paper.

Motivation for PP might be found in anti-materialist arguments (Chalmers, 2009; Jackson, 1982; Kripke, 1980; Levine, 1983). In their general form these arguments are supported by the idea that structure and function do not suffice for explaining consciousness, what together with the claim that physical accounts explain at most structure and function entails the conclusion that physical accounts cannot explain consciousness. From this explanatory gap, some philosophers derive an ontological gap: consciousness is not physical, materialism is false.

The conceivability argument, for example, holds that (1) we can conceive that there is a possible world, w_z, which is a microphysical duplicate of the actual world, $w_@$, but such that some phenomenal truth in $w_@$ is not true in w_z and that (2) if we can conceive that there is w_z, then w_z is possible. But if w_z is possible then

(3) materialism is false, insofar as we take materialism to be committed to the claim that everything that is true in $w_@$ is true in any minimal duplicate of $w_@$—a world which satisfies all the physical truths in $w_@$ and "that's all".

Granting the first premise,[1] the entailment from conceivability to possibility involved in (2) has been rejected by many authors. Chalmers (2002, 2010) presents an analysis of conceivability that attempts to avoid clear counterexamples and singles out the circumstances in which conceivability is a good guide to metaphysical possibility. For this purpose, Chalmers distinguishes between a *positive* and a *negative* notion of conceivability. The notion of positive conceivability is a bit obscure and is characterized "in terms of what subjects can form a positive conception of" (2010, p. 144). However, the notion of negative conceivability is more clear, and it is what the argument, at least for the scope of this paper, requires. A sentence S is negatively conceivable for a subject if and only if she can entertain S and is unable to rule it out through a priori reasoning. Furthermore, to avoid the problems derived from cognitive limitations, Chalmers distinguishes *prima facie* from *ideal* conceivability. S is negatively ideally conceivable if and only if an ideal thinker who has no cognitive limitations can entertain S and is unable to rule it out through a priori reasoning.

A posteriori necessities have also been suggested as counterexamples to the entailment between conceivability and possibility, for some philosophers maintain that, to offer an original example, 'water is not H_2O' is conceivable while not metaphysically possible. In order to deal with these cases, Chalmers presents a two dimensional analysis of conceivability:

There is a sense in which 'water is not H_2O' is not conceivable, call it 'secondary conceivability'. In this sense, a situation in which it seems that water is not H_2O should better be understood as a situation in which there is watery stuff that is not H_2O and hence not water—because water is still H_2O. Secondary conceivability seems to be a good guide to metaphysical possibility but hardly one usable in a priori arguments like the conceivability one, because what is secondary conceivable often depends on empirical investigation. There is, nonetheless, another sense of conceivability, *primary conceivability*, in which we

[1] This premise is not uncontroversial, see Dennett (1991), Dretske (1995), Lewis (1990).

can say that 'water is not H_2O' is conceivable, precisely in the sense that it cannot be ruled out a priori.

Parallel to these notions of conceivability, Chalmers constructs two notions of possibility. A sentence S is 1-possible iff it is true in some world w considered as actual; we can say in this case that w verifies S (S's primary intension is true at w). On the other hand, a statement is 2-possible (metaphysically possible) iff it is true in some world considered as counterfactual; in this case, we can say that w satisfies S (S's secondary intension is true at w).

With these tools in hand and considering primary ideal negative conceivability, we can present Chalmers' argument (2010, p. 152). Let P be the conjunction of all the microphysical truths of the universe and Q a phenomenal truth like 'there is pain'.

(1) P&~Q is (primarily ideally negatively) conceivable.
(2) If P&~Q is conceivable, then P&~Q is 1-possible.
(3) If P&~Q is 1-possible, then P&~Q is metaphysically possible or PP is true.[2]
(4) If it is metaphysically possible that P&~Q then materialism is false.

Materialism is false or PP is true.

Premise 4 has been previously motivated and premise 1 is widely accepted. The entailment from primary negative conceivability to primary possibility seems to be free of counterexamples and this gives support to premise 2.

The interesting premise is 3. If Kripke is right and there is no distinction

[2] In his argument, Chalmers calls this second alternative 'Russellian Monism' or 'Type-F materialism.' In this paper I focus on panprotopsychism, a particular kind thereof, for two reasons. The first one is that most people accept that there is an interesting distinction to be drawn between conscious and non-conscious states, thereby ruling out the thesis that every entity in the actual world is conscious—as radical forms of panpsychism would hold. This makes, I think, panprotopsychism a more interesting option and one that more people will be willing to explore. The second one attends to expository purposes: panprotopsychism is a weaker thesis and if, as I argue in this paper, defenders of panprotopsychism should not endorse OI, defenders of stronger versions of panpsychism shouldn't either for similar reasons.

between appearances and reality in the case of consciousness, the entailment from 1-possibility to metaphysical possibility seems guaranteed in the case of phenomenal truths, and therefore, every world that verifies a phenomenal truth is a world that satisfies it. One way to reject the metaphysical possibility of P&~Q is to hold that microphysical terms have different primary and secondary intensions and that their intrinsic nature is closely tied to consciousness; i.e., that PP is true. In this case, there would be worlds that verify P and also verify ~Q—namely those worlds sufficiently close to ours in which the fundamental microphysical entities have a different intrinsic nature from ours, one that is not tied to consciousness—, while no worlds that satisfies P also satisfies ~Q.

Those who find the argument compelling are left with three (not mutually exclusive) theoretical frameworks to explore as Chalmers (2010, ch. 5) notes: dualism, epiphenomenism and PP. If one believes in the causal closure of physics but do not want to render consciousness epiphenomenal, then PP is definitely the way to go.

Let me now motivate the other main character in this story: the principle of Organizational Invariantism.

2 Organizational Invariantism (OI)

The principle of Organizational Invariantism (OI) holds that two systems with a sufficiently fine-grained functional organization—to fix the mechanisms responsible for the production of behavior, and to fix behavioral dispositions (Chalmers 2010)—will entertain experiences that are qualitatively identical. According to OI, what matters for the phenomenal character of experience is a certain—sufficiently fine-grained—functional organization, and once this functional organization is satisfied we can abstract from its particular realization. As Chalmers presents the idea:

> According to this principle, what matters for the emergence of experience is not the specific physical makeup of a system but the abstract pattern of causal interaction between its

components. (ibid, p.24)

Suppose that the required sufficiently fine-grained functional organization at which behavioral dispositions are fixed is that of neural networks. Neurons in our brain have a certain biochemical composition but, if OI is true, then—at least in the actual world—such a composition is irrelevant for our experiences. Conscious states are made out of *neurons‡*, where something is a *neurons‡* iff it satisfies the same pattern of causal interaction that a neuron does. If *neurons‡* can be made out of silicon, then it would be possible to replace our neurons by those silicon chips without a change in the required functional organization and therefore, according to OI, without a change in the phenomenal character of the experience.

Although this principle has not gone without controversy, Chalmers (1996, ch. 7) provides two convincing and complementary arguments in its favor: the fading/absent qualia and the dancing qualia arguments.

Before presenting the arguments, it will be useful to introduce a conceptual distinction between two components of phenomenal character: the qualitative character and the subjective character (Kriegel, 2009; Levine, 1983). The qualitative character is what distinguishes different kinds of experiences; for example, the kind of experience I have while looking at my red apple from the one I have while, say, looking at a golf course. On the other hand, a theory of subjective character abstracts from the particular ways having different experiences feel and concentrates on the problem of what makes it the case that having a conscious experience feels at all. Hence, the qualitative character is what makes a state the kind of phenomenally conscious state it is, and the subjective character what makes it a phenomenally conscious state at all (Kriegel 2009). The dancing qualia and the fading/absent qualia argument attempt to show respectively that the qualitative character and the subjective character are organizational invariants. Very roughly the arguments go as follows:

In the fading/absent qualia argument, we are asked to consider, for the sake of a *reductio*, the possibility that a functional duplicate of someone having, say, an experience as of red, but whose "brain" is made out of silicon neurons, had no experience at all—contrary to OI. As the two systems have the same functional organization, we can imagine gradually transforming one into the other by

replacing neurons by the corresponding silicon chips without changing the functional set-up. Two things might happen during the transformation: either the replacement of a single neuron switches off consciousness or the experience fades slowly along the process with every replacement. None of the alternatives is plausible, or so argues Chalmers. The first one because it requires that "there would be brute discontinuities in the laws of nature unlike those we find anywhere else" (ibid. p.238). The second one because it would require that a system, whose cognitive processes are not malfunctioning[3] and that is conscious, be systematically wrong about its own experience, complaining about its horrible pain while it is merely having a really mild one.

In the dancing qualia argument, we also consider a transformation process from a system with a *neuronal brain* to a system with a *silicon brain*. However, in this case, we assume that, *pace* OI, they have different experiences; for example, that after the replacement the subject has an experience as of blue while looking at a red apple. To ease my presentation of the argument let me distinguish the *total neural correlate* from the *core neural correlate* of a conscious state, where the former is the neural activity minimally sufficient for the experience and the latter is the part of the total neural correlate that distinguishes one conscious state from another—see for example Block (2007) for some details on this distinction. Let C_1 be the core neural correlate of an experience as of certain shade of red. Let's replace C_1 neurons with the corresponding silicon chips and call the resulting circuit 'C_2'. Suppose now that in a subject S we install a backup circuit with C_2 connected to a switch so that we can connect either C_1 or C_2 to the rest of the brain. If OI were false, when we flip the switch from one position to the other, S's experience would change from an experience as of red to an experience as of blue, but such a change in experience would go unnoticed for S.[4] What is

[3] As the functional description is satisfied all along the replacement (for the sake of the argument we can assume that the time required to replace one neuron can be as short as needed), if the cognitive system was not malfunctioning before, it will not be malfunctioning during the replacement process nor at the end of it.

[4] It can be confidently assumed, as Chalmers does, that “noticing” is one of the cognitive processes on which behavioral dispositions depend on, and hence, that there won't be a change in what S notices upon a change of the switch position.

more, we can imagine flipping the switch back and forth so that "the red and blue experiences "dance" before [S's] eyes" (Chalmers, 1996, p.253), but S doesn't yet notice any change. This does not seem plausible according to Chalmers.

The fading and the dancing qualia arguments provide good support for OI. One might think that the tension between OI and the conceivability argument that has been used to motivate PP is straightforward: if OI is true in every possible world, then P&~Q is not metaphysically possible, because microphysical duplicates are sufficiently fine-grained functional duplicates and by OI enjoy the same qualitative experiences. But the arguments presented by Chalmers, as he himself notes, do not support the truth of the antecedent of this conditional; in particular, they only support the claim that OI holds with nomological necessity. The reasons are, in the first place, that fading and dancing qualia, though implausible, seem to be coherent conceivable hypotheses; and second, that the arguments establish, at the very most, the logical necessity of the conditional: if a system with fine-grained functional organization F has a experiences E, then any system with organization F has experience E. But, as Chalmers remarks, "we cannot establish the logical necessity of the conclusion without establishing the logical necessity of the premise, and the premise is itself empirical." (1996, p.259)

Nevertheless, I will argue in the next section that the arguments that we have seen that back up PP and OI are not compatible. I will present two arguments in the next section. In the first one, I will argue that if PP is true, then there might be sufficiently fine-grained functional duplicates in the actual world that do not entertain the same qualitative experiences, against OI. In the second argument I will show that those who, convinced by the conceivability argument, endorse PP are left with no reason for endorsing OI, because if PP is true, there are worlds that verify P—the conjunction of all the physical truths in the actual world—in which fading and dancing qualia obtain and there is no principled reason to believe that ours is not one of them. If one thinks that dancing and fading qualia arguments support the truth of OI in the actual world, then one better, *pace* Chalmers, gives up PP.

3 Two arguments against the conjunction of PP and OI

3.1 First Argument

On the one hand, OI maintains that what matters for consciousness is to satisfy a certain functional organization—that we can abstract from the particular realization of such a functional organization. On the other hand, PP maintains that consciousness constitutively depends on the intrinsic features of our fundamental particles. There seems to be a tension between these two principles. I will explore this tension to show that the premises of the arguments that back them up are incompatible.

Following with the example above, let's assume that the sufficiently fine-grained functional organization that fixes behavioral dispositions is that of neural networks. In this case, conscious states are made of *neurons♯*, as we have seen. Imagine that S is looking at a red apple while having a horrible headache and that we decide to replace her neurons by other kind of *neurons♯*. If we call the phenomenal character of her experience before the replacement 'Q_1', we can consider the following three possibilities regarding S's experience after the replacement.

(1) S has no conscious experience.
(2) S has Q_1 experience.
(3) S has a Q_2 experience, where $Q_2 \neq Q_1$.

If (1) is true, then OI is false.

If (2) is true, then it seems that PP is false. The reason is that all that it takes to be a *neurons♯* is to satisfy a certain pattern of causal interaction. Hence, it seems reasonable to assume that *neurons♯* can be as different in their fundamental properties as we wish. Let's assume for the sake of simplicity that there is a unique kind of fundamental entity in the actual world; call this kind of entity 'string'. According to PP, consciousness depends on the intrinsic features of strings; but *neurons♯* might be made of very different materials and have very

different internal structure—just for illustration, consider the internal differences between a neuron and a circuit implementing the same functional role made of vacuum tubes or one made of transistors or maybe even a person realizing it (Block, 1978). They will thereby differ in the amount of strings and the relations among them required to realize different kind of *neurons‡*. If *neurons‡* can be so different at the microphysical level, then it seems that microphysical properties play no role in determining the particular kind of experience one undergoes. In reply, one might acknowledge this and step back maintaining that the intrinsic features of the fundamental particles of the actual world provide merely enabling conditions for the experience—they only determine the subjective character. However, this variation of PP does not offer a reply to the conceivability argument, which backs up PP in the first place, as we will see.

If (3) is true, then OI is also false, for there is a change in qualitative character without a change in the required functional structure. I see two routes one might try to explore in reply.

First, one can admit that OI is false but claim that something in the vicinity is true, endorsing a the following modified version of the principle:[5]

> **OI*** Two systems with the same sufficiently fine-grained functional organization will have the same *phenomenal structure*.

Imagine that S is having a RED_{34} experience while looking at a red apple before the replacement. We replace only the neurons of the core neural correlate of this experience and, as a result of this, S has a different kind of experience; call it *'RED^*_{34}'*. According to OI*, RED^*_{34} relates to other experiences in the very same way as RED_{34} does—and hence the *phenomenal structure* is maintained. We can say that RED_{34} and RED^*_{34} are *supersimilar experiences*, where two experiences of different kind are supersimilar iff there is no experiential way to tell the two experiences apart.

I think that postulating supersimilar experiences is problematic, to say the least. If RED^*_{34} and RED_{34} cannot be phenomenologically distinguished and they

[5] I am grateful to XX for suggesting me this possibility.

do not elicit different behavioral dispositions, it is unclear in what sense can they be said to be different kind of experiences.[6]

One might find support for supersimilar experiences in the research on change blindness, which shows that large changes in the experience might go unnoticed.[7] However, these changes are not *unnoticeable* and there is no reason to think that, if the subject is asked to attend to the particular feature that is changing, it would go unnoticed, contrary to what happens in the case of supersimilar experiences: if we ask S to concentrate on the color experience she has while looking at the apple while we change the position of the switch—or flip it back and forth—, changing her experience from RED_{34} to RED^*_{34}, she won't be able, *ex hypothesi*, to notice any difference. Be that as it may, commitment to supersimilar experiences is not the worst problem for those are willing to take this route as we are about to see.

The second option, already suggested in reply to (2) and compatible with the one above, is to maintain that only the subjective character depends on the intrinsic nature of the fundamental entities of the actual world. The following "variation" of OI would remain true in this case:

> **OI**** If two systems have the same sufficiently fine-grained functional organization, then if one of them has conscious experiences, so does the other.

On the contrary, the qualitative character, the particular kind of experience, would not be fixed by the such fundamental entities but rather by the internal structural properties of the *neuronal‡* network realizing the required functional role. So, a phenomenally conscious state is one that satisfies a certain functional role (OI) and is made out of the kind of entities that are fundamental in the actual world (strings):

---

[6] Note that the commitment to the existence of supersimilar experiences is what leads many philosophers to reject disjunctivism about phenomenal character.

[7] Impressed by this work, Chalmers (2010, p.24 fn.7) concedes that the dancing qualia argument is "something less than a reductio". He, nonetheless, endorses OI.

- Structure A (which satisfies function F in the system) + strings realizing structure A = RED_{34}
- Structure B (which satisfies function F in the system) + strings realizing structure B = RED^*_{34}

Where RED^*_{34} and RED_{34} are either the same kind of experience as in (2) or different experiences—supersimilar or not—as in (3). This alternative seems to make PP and OI (or a variation thereof) compatible at the price of accepting that fundamental entities do not play any role in determining the kind of experience—the kind of experience would rather be fixed somehow by the functional role, if they are the same kind of experience, or by the internal structure of the realizer, if they are not. This is not, however, a satisfactory option for those with materialist inclinations or those who willing to keep on the causal closure of physics look into PP for a solution to the conceivability argument that does not commit them to epiphenomenalism. The reason is that the conceivability argument can be reproduced just in terms of the qualitative character: the qualitative character would not be determined by the intrinsic nature of the fundamental entities of the actual world, and hence, not fixed in every world that satisfies P. Let me elaborate:

Defenders of PP reply to the conceivability argument by claiming that, although a zombie world—a microphysical duplicate of the actual one lacking consciousness—is conceivable, this would be a world in which the intrinsic nature of the fundamental entities would differ from that of ours. Therefore, although the zombie world verifies the microphysical description of the actual world it does not satisfy it. Materialism is saved, for no world that satisfies P is a world that differ from ours in regard to consciousness. Now, if as a result of my argument, defenders of PP maintain that the intrinsic nature of the fundamental entities of the actual world is constitutive of the subjective but not the qualitative character, then such intrinsic nature does not metaphysically determine the kind of experience that obtain in every world that satisfies P. Whereas the metaphysical possibility of zombie worlds would be ruled out, because every world that satisfies P is a world where there are experiences, the metaphysical possibility of microphysical duplicates of ours where different experiences obtain is not. We can

conceive worlds where P is the case and different experiences obtain. Some of those worlds will be made of the same fundamental entities that ours and hence, they do not only verify P but also satisfy it: the intrinsic nature of the fundamental entities only fixes the subjective character and therefore in some worlds the same kind of particles might give rise to different experiences. So, there would be microphysical duplicates of the actual world in which different experiences obtain (experiences with different character); i.e., materialism, as defined, would be false: there would be worlds that are microphysical duplicates of the actual world, which are not duplicates simpliciter.

To sum up, one might respond to my argument by endorsing a variation of PP according to which, the subjective character but not so the qualitative character is fixed by the intrinsic nature of our fundamental particles. But in doing so, this form of PP is left unmotivated because it is insufficient to reject the metaphysical possibility of a world that satisfies P and not Q, once its epistemic possibility is granted. So, this form of PP is committed to the very same kind of strong necessities that a posteriori materialists (like type-B materialists—see Chalmers, 2003) postulate and hence they are in no better position to offer a reply to the conceivability argument.

In the next subsection I will further argue that the intuitions that back up PP and OI are incompatible and hence that one should better give up on one of them.

3.2 Second Argument

In the actual world, properly organized strings, assuming that they are its fundamental entities, give rise to tables, red apples, butterflies, chocolate, and also consciousness. According to PP, consciousness depends in addition on the intrinsic nature of strings.

Those who find in the conceivability argument the motivation for PP accept that there are worlds that verify P but do not satisfy it because their fundamental entities differ in their intrinsic properties. Imagine one of these possible worlds, w_z, in which their fundamental entities, call them 'strings-', differ in their intrinsic nature from strings. Furthermore, strings- are such that they do not give rise to

conscious experiences: w_z is a zombie world. W_z is a world that verifies P,—and P&~Q; being Q any positive phenomenal sentence like "there are headaches"—but being made of strings- instead of strings it does not satisfy P—nor consequently P&~Q. Worlds like w_z are not problematic. But now, consider the semi-zombie world, w_{sz}. W_{sz} also verifies P, but has both strings and strings- as its fundamental entities and therefore does not satisfy P. In w_{sz} tables, butterflies and chocolate can be made of strings, of strings- or a combination of both kind of entities. In this case, we can run an argument against PP that mirrors the arguments in favor of OI:

Marta inhabits w_{sz}. Her brain is completely made out of strings and she enjoys conscious experiences. Imagine that she is having a terrible headache at time t and let C_{pain} be the core neural correlate of her painful experience. Let C_{nopain} be a physical duplicate of C_{pain} but made out of strings-. A commutator that allows to connect either C_{pain} or C_{nopain} to the rest of the brain is installed in Marta's brain and she is asked to concentrate in her pain experience. When C_{pain} is connected, she has a horrible headache, whereas when C_{nopain} is connected, she has no pain experience at all. However, she cannot notice any difference, the position of the switch makes no difference to her.[8] The implausibility of cases like this is precisely what supports OI in the original arguments. Now, recall that w_{sz} verifies P, and so there is no way for us to know whether we in fact inhabit a world like w_{sz}. So, if one is persuaded that OI is true of the actual world, then, for the very same reason, one should reject the claim that Marta's experience changes as we flip the switch—any good reason in favor of the claim that OI is true of the actual world will remain valid in any world that verifies the microphysical

[8] Chalmers has noted that he is not moved by the dancing qualia anymore and some panpsychist might be willing to follow him in this respect. During discussion in the CO5, he contrasted the dancing qualia and the fading/absent qualia arguments noting that, no matter how large, dancing qualia cases require momentary errors, but that there is a good explanation of those errors in terms of the massive switching processes that takes place at that moment, whereas the fading qualia case "requires huge ongoing errors (a subject believing that their consciousness is just like mine when instead it contains, just a few bits)." It should be clear that the example can be easily modified to accommodate this worry: just let C_{nopain} connected to the rest of the brain; Marta will believe that she has a horrible headache when she has not.

description of the actual world; i.e. P—thereby rejecting PP.

Defenders of PP can explore several objections to this argument, let me try to show that they are not felicitous.

The first one would be rejecting the metaphysical possibility of w_{sz} by holding that the intrinsic nature of the fundamental entities of the actual world is also constitutive of the physical set up of the actual world: that is, that any world that verifies P is a world that satisfies P and hence that neither the zombie world nor the semi-zombie world are metaphysically possible.[9] However, to do so, one would have to make much stronger commitments than those required by the conceivability argument itself, maintaining that causal relations as well as other physical dispositions are entered in virtue of the intrinsic properties of the relata. Some materialist might jump off the ship when they have also to accept these views on, say, causation. But it is important to note that something stronger than this is required: it should not just be the case that causation and other physical dispositions are grounded on the intrinsic properties of our fundamental entities, but rather that they can only have them as their categorical basis. Coherent as this position might be, and regardless of what its independent motivation might be, it is already several steps beyond the claim that consciousness depends on the intrinsic nature of the fundamental entities of the actual world, which is what the conceivability argument attempted to show.

Another possible worry,[10] might be that fading and dancing qualia (FQ and DQ respectively in what follows) arguments do not try to show that FQ and DQ are impossible but rather to show that it is not plausible that they obtain in the actual world; they just attempt to provide a justification for a belief about the actual world, so the metaphysical possibility of the semi-zombie world cuts no ice. This claim rests, or so I will argue, on a misunderstanding of the point that the argument makes. Let me be more specific:

If one is moved by the FQ and DQ arguments, then one thinks that one is justified in believing that FQ and DQ do not obtain in the actual world—for it is implausible that they obtain in the actual world. Now, call 'E' whatever evidence one takes to be in favor of the claim that FQ and DQ do not obtain in the actual

[9] I am indebted to YY for discussion here.

[10] I am indebted to XX and ZZ for pressing me at this point.

world. There are worlds containing only strings that verify the same truths than the actual world, call 'w_{s+}' one of them. Inhabitants of w_{sz} and w_{s+} who are microphysical duplicates, would, therefore, have the same evidence.[11] So, considering that these pairs of individuals will have the same evidence and have obtained it in the same way, then they would have to be equally justified in believing that FQ and DQ do not obtain in their respective worlds—if, for example, one thinks that the fact that FQ do not obtain because they would require discontinuities in the laws of nature that we do not find anywhere else (Chalmers 1996, p. 238), then both the inhabitant of w_{s+} and her counterpart in w_{sz} would think so. But the inhabitant of w_{sz} is wrong because FQ and DQ would obtain in w_{sz} if PP were true. Once PP enters into play, E is no longer evidence for the claim that FQ and DQ do not obtain in the actual world, because we don't know whether we inhabit w_{s+} or w_{sz}. In other words, the problem I raise, is that postulating the truth of PP acts as an *undercutting defeater* (Pollock and Cruz 1999, p. 196) of the evidence we might have in favor of the claim that FQ and DQ do not obtain in the actual world.

We can take for granted that if E is evidence for an hypothesis H, then if one believes that E is the case, then one is justified in believing that H.[12] Now, consider the following plausible principle:

> **EP** If the fact that E is compatible with a *relevant world* in which E obtains but H is not the case, then E is not evidence for H.[13]

---

[11] Note that we are, therefore, considering only inhabitants w_{sz} whose brain is made entirely out of stings, such as Marta before the surgery. This way we guarantee that they entertain the same experiences as their counterparts.

[12] One might object that one might believe that E is the case and not be justified in believing H, if E is not justified—if one has come to believe that E is the case through wishful thinking for example. For that purpose on might restrict the claim above accordingly. I have preferred not to do it for the sake of simplicity in the exposition. Moreover, in my argument we are considering two individuals who have come to believe E in the very same way and my opponent accepts that the inhabitant of w_{s+} would be justified; therefore, so would be the one in w_{sz}, at least insofar as the mechanism for acquiring the evidence is concerned.

[13] If one prefers a probabilistic account of evidence then the principle should demand, at the very

Whether E is evidence for H will depend on the situation (where the relevant parameters of the situation for our purposes might be, for example, pragmatic—see Stanley (2005)). Relevant worlds enter into play to model such a dependence. Providing an analysis of the notion of relevant worlds would require an entire book but fortunately this is not necessary for my purposes here and an intuitive grasp provided by examples would be sufficient. Whereas the fact that my clock indicates that it is ten o'clock might be evidence for it's being ten o'clock in a situation in which I have to meet a friend, it might not be if I have to attend a very important meeting at 10:30. In this latter case, I might want to rule out the possibility that the actual world is such that my clock has stopped to be justified in believing that it is ten o'clock. Similarly, whereas the fact that there are a lot of stars in the sky might be evidence for the hypothesis that the sun will shine tomorrow when I am considering to look for my umbrella for the trip next day, it does not count as evidence in a discussion about the laws of nature: in the latter case, but not in the former, relevant world include those in which suddenly the sun does not rise.

In our case, the relevance of the semi-zombie world is guaranteed by the very same nature of the discussion: once PP enters into play, worlds that are relevant are precisely those that verify the same physical truths than the actual world, as both w_{s+} and w_{sz} do. Consequently, the introduction of PP works as a defeater undercutting the evidential connection between E and the claim that FQ and DQ do not obtain in the actual world. The reason is that, being E whatever evidence we take to be in favor of the claim that FQ and DQ do not obtain in the actual world, once we introduce PP as a possibility there is a relevant world, w_{sz}, which we cannot rule out, where E obtains but so do FQ and DQ. So, with PP operating, E does not provide evidence in favor of the claim that FQ and DQ do not obtain in the actual world—for there is no way for us to decide whether the actual world is such that there are only strings (w_{s+}) or also string- (w_{sz})—; no reason for endorsing OI.

Summarizing, when presented with Chalmers' arguments in favor of OI

---

least, that there is no relevant world in which E is the case and the probability of H is not higher than the probability of not-H.

many might be persuaded that DQ and FQ are not possible of the actual world. The problem is that once PP enters into play whatever evidence one might have against such possibility is defeated. If PP is true and the actual world is w_{sz}—and within the debate we have no non-question-begging reason to rule out this possibility—, then it is theoretically predicted, *pace* Chalmers' reasoning in the FQ argument for example, either that there are brute discontinuities in the laws of nature unlike those we find anywhere else or that a system, whose cognitive processes are perfectly functional and who is conscious, be systematically wrong about her own experience. In reply, one cannot simply rule out the possibility that the actual world is w_{sz}. Just as one cannot simply reply to Hume that it is not plausible that the sun won't rise tomorrow because such a world is relevant in a debate about the laws of nature, any world that verifies P—and w_{sz} is one of them —is relevant in the debate about the truth of PP. Hence, if one endorses PP one is left with no justification for believing that FQ and DQ do not obtain in the actual world and hence with no justification for holding OI.

4 Conclusion

In this paper I have shown that the reasons that lead one to endorse PP as a solution to the conceivability argument and to believe that OI is true of the actual world are, *pace* Chalmers, not consistent. It might still be the case that PP is true and that OI is true of the actual world, but we are left with no reason to believe such a thing and there are good reasons to deny it.

If one finds the dancing and fading qualia arguments compelling one should reject PP and if one believes that PP is true, one should find a way to resist the dancing and fading qualia arguments.[14]

[14] Acknowledgments

References

Block, N.: 1978, Troubles with Functionalism, Minnesota Studies in the Philosophy of Science 9, 261-325.

Block, N.: 2007, Consciousness, accessibility, and the mesh between psychology and neuroscience, Behavioral and Brain Sciences 30, 481-548.

Chalmers, D. J.: 1996, The Conscious Mind: In Search of a Fundamental Theory, 1edn, Oxford University Press, USA.

Chalmers, D. J.: 2002, Does conceivability entail possibility?, in T. S. Gendler and J. Hawthorne (eds), Conceivability and Possibility, Oxford University Press, pp. 145-200.

Chalmers D. J.: 2003 Consciousness and its place in nature. In: Stich SP, Wareld TA (eds) Blackwell Guide to the Philosophy of Mind, Blackwell

Chalmers, D. J.: 2009, The Two-Dimensional argument against materialism, in B. P. McLaughlin and S. Walter (eds), Oxford Handbook to the Philosophy of Mind, Oxford University Press.

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

Dennett, D. C.: 1991, Consciousness Explained, 1 edn, Back Bay Books.

Dretske, F.: 1995, Naturalizing the Mind, MIT Press.

Jackson, F.: 1982, Epiphenomenal qualia, Philosophical Quarterly 32(April), 127-136.

Kriegel, U.: 2009, Subjective Consciousness: A Self-Representational Theory, Oxford University Press, USA.

Kripke, S. A.: 1980, Naming and Necessity, Harvard University Press.

Levine, J.: 1983, Materialism and qualia: The explanatory gap, Pacific Philosophical Quarterly 64(October), 354-61.

Lewis, D.: 1990, What experience teaches, in W. G. Lycan (ed.), Mind and Cognition, Blackwell, pp. 29-57.

Pollock, J. and Cruz, J.: 1999, Contemporary Theories of Knowledge, Towota, NJ: Rowman and Little eld Publishers.

Stanley, J.: 2005, Knowledge and Practical Interest, Oxford University Press.