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 dualists inclinations who, believing in the causal closure of the Physics, do not want to render consciousness epiphenomenal might find in Panpsychism (PP) and 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 expericences, is true.

Some philosophers, like David Chalmers, have either shown their sympathy for both principles or explicitely 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 offered by Chalmers to support the principle. Finally, in section 3, I argue that there is a tension between PP and OI; the same argument that back 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 about the intrinsic nature of the fundamental entitites (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 which exists throughout the universe: the most fundamental entities enjoy mentality.

In this paper I am interested in consciousness. If one believes that there is an important distinction between conscious and unconscious mental states, and 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.

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 don't 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 single out the circumstances in which conceivability is 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 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 S if and only if S 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 iff and ideal thinker who has no cognitive limitations can entertain S and is unable to rule it out through a priori reasoning.

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

There is a sense in which 'water is not H_2O' is not conceivable, call it

[1] Cf. Dennett (1991); Dretske (1995); Lewis (1990)

'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 but is not water, for 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, for what is secondary conceivable depends on empirical investigation. But there is another sense of conceivability, *primary conceivability*, in which we can say that 'water is not H_2O' is conceivable: in the sense that it cannot be ruled out a priori not even by an ideal thinker.

Parallel to these notions of conceivability Chalmers contructs two notions of possibility. A sentence S is 1-possible iff it is true in some world w considered as actual; we 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, we say that *w satisfies S*. (S's secondary intention is true at w).

With these tools in hand and considering primary ideal negative conceivability, we can present Chalmer's argument (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) $P\&\neg Q$ is (primary ideal negative) conceivable

(2) If $P\&\neg Q$ is conceivable, then $P\&\neg Q$ is 1-possible.

(3) If $P\&\neg Q$ is 1-possible, then $P\&\neg Q$ is metaphysically possible or PP is true.[2]

(4) If it is metaphysically possible that $P\&\neg Q$ then materialism is false.

$\therefore$ 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 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 $P\&\neg Q$

[2] In his argument Chalmers calls this alternative Russellian Monism or Type-F materialism. In this paper I focus on panprotopsychism 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 panpsychism shouldn't either for similar reasons.

is to hold that microphysical terms have different primary and secondary intentions and that their intrinsic nature is closely tied to consciousness; i.e., if PP is true. In this case, there are worlds that verify P and also verify $\neg Q$, 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 satisfy P also satisfy $\neg Q$.

Those who find the argument compelling are left with three 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 that 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 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 *neuron*† iff it satisfies the same pattern of causal interaction that a neuron. 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 experience.

Although principle has not gone without controversy, Chalmers (1996, ch. 7) provides two convincing arguments in its favor: the *fading qualia* and the *dancing qualia* arguments. Very roughly the arguments go as follows:

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

by replacing neurons by silicon chips with the same function. Two things might happen during the transformation: either the replacement of a single neuron switchs 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". The second one because it would require that a system, whose cognitive processes are perfectly functional and who 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 consider also a transformation process from a system with a *neuronal brain* to a system with a *silicon brain* and assume that, pace OI, they have different experience; for example, the later 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 later 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 circutitry 'C_2'. Suppose now that in a subject S we install a *backup circuit* with C_2 connected to a switch to change from C_1 to C_2. If OI were false, then we flip the switch from one position to the other then 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_1. What is 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 he still doesn't notice any change. This does not seem plausible.

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 I has been used to motivate PP is straightforward: if OI is true in every possible world, then $P\&\neg Q$ is not metaphysically possible, because microphysical duplicates are *fine-grained* functional duplicates and by OI enjoy the same qualitative experiences. The arguments presented by Chalmers do not support the truth of the antecedent of this conditional; in particular, they only support 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 hypotyheses. 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 notes "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 PP and OI are not compatible principles. I will present my argument in two steps. 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. One possible reply to this argument would be to give up on OI and accept something in the vicinity that keeps its spirit. So, in a second step I will argue that defenders of PP are committed to the existence of dancing and fading qualia in some worlds that verify P — a sentence of all the physical truths in the actual world — but do not satisfy it and that this is as implausible in these worlds as it is in ours. If one thinks that dancing and fading qualia arguments support the truth of OI in the actual world, then one should, *pace* Chalmers, give up PP.

3 The Argument against the conjunction of PP and OI

3.1 First Pass

On the one hand, OI maintains that what matters for consciousness is to satisfy a certain functional organization, 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 is that of neural networks. In this case, conscious states are made out 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 a Q_2 experience, where $Q_2 \neq Q_1$.
3. S has Q_1 experience.

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

If (2) is true, then OI is also false, for there is a change in qualitative character without a change in the required functional structure. We can, nonetheless, accept (2) while keeping in the spirit of OI by endorsing the following modified version of the principle:[3]

OI* Two systems with the same 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

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

it 'RED^*_{34}'. According to OI*, RED^*_{34} relates to other experiences in the very same way as RED_{34} does. 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 do not elicit different behavioral dispositions, it is unclear in what sense can they be said to be different kind of experiences.[4]

One might find support for supersimilar experiences in the research on change blindness, which shows that large changes in the experience might go unnoticed.[5] However, these changes are not *unnotizable* 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. But this is not the case for supersimilar experiences. If we ask S to concentrate in the color experience she has while looking at the apple while we flip back and forth the switch, changing her experience from RED_{34} to RED^*_{34}, she won't be able, *ex hypothesi*, to notice any difference.

Commitment to supersimilar experiences is not the worst problem for those willing to take this route as we will see in the next subsection. Let me first show some problems for option (3).

If (3) is true, then it *seems* that PP is false. The reason is that all that it takes to be a *neuron*† 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 intrisic features of strings; but *neurons*† might be made of different materials and have very different internal structure, thereby differing in the amount of strings and the relation 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 experience.

One possible reply would be to maintain that the intrinsic features of the fundamental particles of the actual world provide merely enabling conditions for the experience. Functional roles within a system determine the kind of experience (or sets of supersimilar experiences if one embrace option (2) and the particular kind of experience is determined by the structure of the *neurons*†). Along these lines, 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):

- Structure A (which satisfies function F in the system S) + stings realizing structure A = RED_{34}

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

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

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

Where RED^*_{34} and RED_{34} are either the same kind of experience (3) or supersimilar experiences (2). This alternative seems to make PP and OI (or OI*) compatible at the price of accepting that fundamental entities do not play any role in determining the kind of experience (or the existence of supersimilar experiences). In the next subsection I will argue that the intuitions that back up PP and OI are nonetheless incompatible and hence that one should give up on one of them.

3.2 Second Pass

In the actual world, assuming that strings are the fundamental entities of the actual world, tables, red apples, butterflies, chocolate, etc. supervene on properly organized strings. According to PP, so do conscious experiences.

Those who find in the conceivability argument the motivation for PP accept that there are worlds that verify P, where P is the conjunction of all the microphysical truths of the universe, but do not satisfy it because their fundamental entities differ in their intrinsic properties. Imagine one of these possible world 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 $P\&\neg Q$, being Q any positive phenomenal sentence like "there are headaches") but being made of strings- instead of strings does not satisfy P. 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 it. 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 exactly 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, as in the dancing qualia thought experiment, is instaled in Marta's brain and she is asked to concentrate in her pain experience. When C_{pain} is connected she has 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. The implausibility of cases like this is precisely what supports OI in the original argument. 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}: 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, thereby rejecting PP.

4 Conclusion

In this paper I have shown that the reasons that leads 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 so 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.[6]

References

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

Chalmers D (2010) The Character of Consciousness. Oxford University Press

Chalmers DJ (1996) The Conscious Mind: In Search of a Fundamental Theory, 1st edn. Oxford University Press, USA

Chalmers DJ (2002) Does conceivability entail possibility? In: Gendler TS, Hawthorne J (eds) Conceivability and Possibility, Oxford University Press, pp 145–200

Chalmers DJ (2009) The Two-Dimensional argument against materialism. In: McLaughlin BP, Walter S (eds) Oxford Handbook to the Philosophy of Mind, Oxford University Press

Dennett DC (1991) Consciousness Explained, 1st edn. Back Bay Books

Dretske F (1995) Naturalizing the Mind. MIT Press

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

Kripke SA (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: Lycan WG (ed) Mind and Cognition, Blackwell, pp 29–57

[6] Acknowledgements